Bindi: Affective Internet of Things to Combat Gender-Based Violence

Jose A. Miranda Calero, Esther Rituerto-González, Clara Luis-Mingueza, Manuel F. Canabal, Alberto Ramírez Bárcenas, Jose M. Lanza-Gutiérrez, Carmen Peláez-Moreno, *Member, IEEE*, and Celia López-Ongil, *Senior Member, IEEE*

***Abstract*—The main research motivation of this article is the fight against gender-based violence and achieving gender equality from a technological perspective. The solution proposed in this work goes beyond currently existing panic buttons, needing to be manually operated by the victims under difficult circumstances. Instead, Bindi, our end-to-end autonomous multimodal system, relies on artificial intelligence methods to automatically identify violent situations, based on detecting fear-related emotions, and trigger a protection protocol, if necessary. To this end, Bindi integrates modern state-of-the-art technologies, such as the Internet of Bodies, affective computing, and cyber–physical systems, leveraging: 1) affective Internet of Things (IoT) with auditory and physiological commercial off-the-shelf smart sensors embedded in wearable devices; 2) hierarchical multisensorial information fusion; and 3) the edge-fog-cloud IoT architecture. This solution is evaluated using our own data set named WEMAC, a very recently collected and freely available collection of data comprising the auditory and physiological responses of 47 women to several emotions elicited by using a virtual reality environment. On this basis, this work provides an analysis of multimodal late fusion strategies to combine the physiological and speech data processing pipelines to identify the best intelligence engine strategy for Bindi. In particular, the best data fusion strategy reports an overall fear classification accuracy of 63.61% for a subject-independent approach. Both a power consumption study and an audio data processing pipeline to detect violent acoustic events complement this analysis. This research is intended as an initial multimodal baseline that facilitates further work with real-life elicited fear in women.**

Manuscript received 4 July 2021; revised 4 March 2022 and 6 April 2022; accepted 13 May 2022. Date of publication 23 May 2022; date of current version 24 October 2022. This work was supported in part by the Department of Research and Innovation of Madrid Regional Authority, in the EMPATIA-CM Research Project (Reference Y2018/TCS-5046) funded by MCIN/AEI/10.13039/501100011033 under Grant PDC2021-121071-I00; in part by the European Union "NextGenerationEU/PRTR;" in part by the Spanish Ministry of Universities with the FPU under Grant FPU19/00448; and in part by the Madrid Government (Comunidad de Madrid-Spain) through the Multiannual Agreement with UC3M in the line of Excellence of University Professors (EPUC3M26), and in the context of the V PRICIT (Regional Programme of Research and Technological Innovation). *(Corresponding author: Jose A. Miranda Calero.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Research Ethics Committee of the Carlos III University of Madrid.

Jose A. Miranda Calero, Manuel F. Canabal, Alberto Ramírez Bárcenas, and Celia López-Ongil are with the Department of Electronics, Universidad Carlos III of Madrid, 28911 Leganés, Spain (e-mail: jmiranda@ing.uc3m.es; mcanabal@ing.uc3m.es; alramire@ing.uc3m.es; celia@ing.uc3m.es).

Esther Rituerto-González, Clara Luis-Mingueza, and Carmen Peláez-Moreno are with the Department of Signal Theory and Communications, Universidad Carlos III of Madrid, 28911 Leganés, Spain (e-mail: erituert@ing.uc3m.es; cluis@pa.uc3m.es; carmen@tsc.uc3m.es).

Jose M. Lanza-Gutiérrez is with the Department of Computer Science, Universidad de Alcalá, 28801 Alcalá de Henares, Spain (e-mail: jm.lanza@uah.es).

Digital Object Identifier 10.1109/JIOT.2022.3177256

***Index Terms*—Artificial intelligence of things, edge computing, fear recognition, microelectromechanical systems, multimodal data fusion, smart sensors.**

I. Introduction

GENDER-BASED Violence (GBV) is one of the most pervasive violations of human rights. According to the United Nations, International Children's Emergency Fund, and the World Health Organization [1], 30% of women worldwide have suffered or will suffer physical or sexual violence during their lives. This fact places GBV at a very critical level of social alarm, even surpassing armed terrorism in several countries. The United Nations defines violence against women as "any act of GBV that results in, or is likely to result in, physical, sexual, or mental harm or suffering to women, including threats of such acts, coercion, or arbitrary deprivation of liberty, whether occurring in public or in private life."[1] This type of violence includes GBV or domestic violence against women, men, or children living in the same domestic unit, causing severe harm to families and communities. As an example of the impact of GBV on women, who are the people suffering most from this problem, based on data 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 [2]. Another worrying statistic is that, from January 2003 to February 2022, there were 1130 GBV victims (GBVVs) were murdered in Spain by their male partners [3]. Moreover, the European Institute for Gender Equality (EIGE) has estimated that the cost of gender inequality across the European Union is 366 billion euros a year; GBV makes up 79% of this cost, amounting to 289 billion euros [4]. GBV and other long-term social problems should mainly be tackled through education, awareness, and sensitization programs. Although, technology may aid in preventing and combating their effects, which is one of the research motivations guiding this article.

In recent years, the growth of digital technology has benefited the development of novel Web and smartphone applications aimed at fighting against GBV. The applications range from mapping sexual violence exposure within a city, to offering a trusted and direct connection to law enforcement agencies (LEAs) [5], [6] and to providing with panic button

[1]https://www.un.org/en/genocideprevention/documents/atrocity-crimes/Doc.21_declaration%20elimination%20vaw.pdf

Fig. 1. Outline of Bindi.

devices. In some countries, such as India, a directive has been issued related on the mandatory inclusion of a panic button on every mobile phone sold from 2017. However, panic buttons present significant limitations regarding women's safety, such as the requirement of an active role in their self-protection, which is certainly not possible under some types of aggression, their lack of an inconspicuous design, which leads to GBV stigma, or even worse, the lack of infrastructure support [7]. Despite the technological efforts, this type of solutions presents different research gaps questioned by several GBV experts [8], who demand more advanced research and technology for these solutions regarded as outdated. Moreover, they agree that the technology should be complemented with better training of victim support professionals to avoid harmful revictimisation.

On this basis, the motivation of this work is to respond to the requirements discussed above to help fight and combat GBV by employing an autonomous system that guarantees the victims' protection. With this purpose, the UC3M4Safety[2] multidisciplinary research team was set up in 2017 with the objective of developing an innovative solution called Bindi,[3] whose outline is presented in Fig. 1. Bindi is an autonomous system powered by artificial intelligence and the Internet of Things (IoT) to automatically report when a woman is in a risky situation related to GBV. This identification will be performed by automatically detecting fear-related emotions in the victim. The IoT architecture in Bindi considers the usual three-layer division [9], i.e., edge, fog, and cloud. In this system, the edge-computing layer is conceived as a smart cyber–physical network composed of two devices (a pendant and a bracelet), measuring physiological and auditory data over time. In this first layer, a lightweight machine learning[4] system is running on the bracelet detecting possible risky situations. The fog computing layer in Bindi is conceived as a smartphone application, which implements a neural-based engine for auditory data providing information about risky situations and also feeds from the machine learning response in the bracelet. Thus, based on this information, if a risky GBV situation is predicted, an alarm will automatically be triggered to the corresponding protection services. Finally, the relevant information obtained throughout the whole process is securely stored in specific computing services in the cloud.

[2]https://portal.uc3m.es/portal/page/portal/inst_estudios_genero/proyectos/UC3M4Safety

[3]The Bindi system is an approved model of utility by the Spanish Office of Patents and Trademarks.

[4]Machine learning is the study of computer algorithms that allow computer programs to automatically learn about data and improve through experience [10].

This article presents the system hardware architecture of Bindi and the validation of its data processing pipelines. The goal is to analyze and gain a better understanding of women's responses to the fear emotion in risky situations. The main contributions of this article are as follows.

1) It introduces and uses a novel emotion recognition data set, i.e., the women and emotion multimodal affective computing (WEMAC) data set [11], targeting GBV-related fear elicitation. This data set, which is freely available, contains physiological and auditory information from nonacted emotions elicited in an immersive virtual reality environment.
2) Three multimodal data fusion strategies are evaluated and validated to make a final decision about risky situations in the fog layer. To the best of our knowledge, this is the first time that a multimodal fusion of physiological and speech data for fear recognition has been given in this context.
3) A novel audio data processing pipeline for the identification of acoustic events related to risky GBV situations is presented.
4) The experimental results show an average accuracy of the fear recognition rate of up to 63.61% with the leave-half-subject-out (LASO) method, which is an state-of-the-art subject-independent training classification strategy [12]. To the best of our knowledge, this is the first time a LASO model considering fear recognition, multisensorial signal fusion, and virtual reality stimuli has been presented. Note that the significance of the results is limited by the number of participants, i.e., 47 women. This fact is currently being addressed by increasing the database size.
5) A comprehensive power consumption analysis regarding the edge computing devices in Bindi is provided to compare the battery impact of each of the evaluated the data processing chains.

The remainder of this document is structured as follows. Section II discusses related work. The different elements of Bindi are detailed in Section III, followed by the data processing pipelines and their technical particularities in Section IV. The experimental methodology is explained in Section V, with the reported results detailed in Section VI. A comprehensive power consumption study for the different hardware elements closes the technical account in Section VII. Finally, a detailed discussion about the architecture of Bindi, the results, and their significance appears in Section VIII, followed by conclusions and future research directions in Section IX.

II. Related Work

GBV is already considered a pandemic by private and public organizations worldwide (such as the World Bank and EIGE) that can be enacted in different forms. Technological and scientific advancements can leverage new solutions to combat this social problem. First, this section addresses the use of technology concerning the GBV problem and introduces the potential of affective computing to deliver tools for preventing and combating GBV. Second, a detailed review concerning current trends in the Internet of Bodies (IoB) systems is

provided. Third, different fear and fear-related emotion recognition systems presented in the literature are analyzed. This analysis is done in both a unimodal and multimodal manner to state the research gaps related to the scarcity of emotion recognition systems based on multimodal data fusion. Finally, a review of the current multimodal data sets available in the scientific literature is provided, highlighting their deficiencies for the application in question.

A. Technology Against Gender-Based Violence

The Istanbul Convention [13] (a European Convention on combating violence against women) recognizes four main GBV manifestations: 1) physical; 2) sexual; 3) psychological; and 4) economic. Within this context, digital technologies expansion has a profound impact with two sides. On the one hand, the effects of technology-facilitated GBV (TFGBV) must be assessed and counteracted in any current and future technological advancements. Recently, Dunn [14] evaluated the expressions of GBV online, such as stalking, doxing, and impersonation. These new manifestations must be assessed and counteracted by current and future solutions (social or technology based) to combat the pervasiveness of TFGBV. On the other hand, technology is enabling the application and implementation of solutions toward preventing and combating GBV [15]–[17]. Specifically, in Spain, in the context of the comprehensive law against GBV (2004) [18], three technological tools have been implemented to support and protect GBVVs: 1) VioGen [19]; 2) ATENPRO [20]; and 3) Cometa [21].

First, VioGen can estimate the risk level faced by a GBVV and determine the adequate type and degree of protection for them. This risk level is updated according to the legal and social situation of the GBVV. VioGen is the result of intensive research by the Spanish Home Affairs Department with various Spanish university research groups, composed of experts in psychology, criminology, and sociology. The efficiency of this tool was validated in [22], achieving up to 85% and 54% average sensitivity and specificity risk prediction rates, respectively. Second, ATENPRO is a service that provides GBVVs with a direct and 24-7 hotline, triggered through a panic button. In this specialized telephonic assistance center, specifically trained attendants give an adequate response to handle GBV situations in real time, contacting Spanish LEAs if required. Finally, Cometa is a system conceived as a set of telematic control devices adopted when a restraining order is issued against an 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 center. The aggressor must also wear a lightweight bracelet-like radio-frequency device that connects to both geolocation devices. Although Cometa offers a technological solution for combating GBV, its limited battery life and outdated technology present a high false-positive rate [23], [24], apart from the risk of harassment for the victims.

Leaving aside public institutional resources to fight GBV, the private sector offers different technological solutions, such as Web and smartphone applications and wearable devices with panic button functionalities. For instance, several companies presented their technological solutions in the worldwide Anu and Naveen Jain Women's Safety challenge launched in 2017 by the XPrize Foundation [25]. The goal was to deliver an inconspicuous (and affordable) system capable of triggering an alarm in less than 90 s in the case of sexual assault detection. Most of the presented solutions[5] revolved around the panic button concept, although some of them proposed the use of artificial intelligence.

From this review, we can conclude that none of the public or private technological solutions to combat GBV benefit from key current state-of-the-art and consumer electronics progress, such as physiological and auditory analytics and affective computing. These advancements could be exploited for a better, autonomous, and more inconspicuous technological GBV prevention tool. The latter is the goal of the UC3M4Safety team and its technological solution Bindi.

B. Internet of Bodies

The growth of research on devices that monitor signals from the human body during the last years, as both edge devices in Bindi, supposes an imminent extension of the IoT domain. This trend emerges concerning interconnected devices (e.g., worn, implanted, embedded, and swallowed) located in-on-and-around the human body forming a network, which is currently being called the IoBs [26]. This novel field has many applications, such as human activity recognition [27], user authentication [28], and even emotion recognition [29]. This field also encompasses essential studies on the limitations of such sensors, such as time delay and energy consumption issues [30]. Thus, such in-body sensors can acquire different types of physiological information at the same time, which derives studies related to the use of multimodal data fusion techniques [31], [32].

This IoB proliferation is accompanied by advances in machine learning and deep learning technologies, resulting in an explosion of mobile intelligence and placing increasing demands on computing resources that mobile edge devices cannot meet. Consequently, edge computing capabilities are being boosted and explored to deliver better intelligence engine inference services to end users [33]. For instance, in [34], they worked on accelerating the training process of large machine learning models in IoT to meet the hardware limitations.

Within this IoB context, the presented work intends to provide and foster the generation of novel lightweight multimodal data fusion techniques fed by human body monitoring toward their applicability to current edge-computing devices, such as the ones in Bindi.

C. Emotion Recognition

Affective computing [35] is a multidisciplinary research field aimed at recognizing human emotions to provide better working conditions, entertainment, or services to people. It relies not only on smart sensors and digital signal processing but also on artificial intelligence techniques, such as machine

[5]https://www.xprize.org/prizes/womens-safety/teams

learning. This latter technology allows for an understanding of the connections between the emotional states and signals collected from a person being monitored or even the environment. For instance, the collaborative research among the psychology, computer science, smart sensors, and cognitive science fields [36] allows for the detection of different emotional states through physiological and physical signal monitoring. Some examples of physical signals include audio, voice, image, or video signals, tracking the background of the scene or the user. Some examples of physiological variables include heart rate (HR), galvanic skin response (GSR), skin temperature (SKT), electromyogram (EMG), and electroencephalogram (EEG).

Within this context, research on negative emotion detection in violent situations could help prevent and combat the GBV problem, since potentially risky situations for a user cause specific negative emotions, such as fear. In this regard, UC3M4Safety claims that the identification of the emotions felt when someone is a victim of violence is of paramount importance when trying to protect human lives [37]. This identification can help avoid violent assaults, including sexual assaults and violence toward vulnerable social groups. Although there is significant activity in the literature regarding emotion recognition through auditory and physiological signals for different purposes and considering different setups [38]–[40], none of the solutions found focuses on the GBV use case.

Regarding the design of an emotion recognition system, a crucial yet challenging learning process element is related to emotion labeling. Since the 19th century, different emotion theories and models have been proposed to understand the human response to external stimuli [41]–[43]. However, there are two main theories about emotion classification that are usually considered for labeling: 1) the categorical and 2) dimensional models. The former identifies various sets of discrete emotions common in different cultures and splits them into distinct categories [44]. The dimensional classification defines a continuous affective space with two or more dimensions, such as pleasure, arousal, dominance, and/or familiarity [45]. Note that a dimensional method based on three dimensions allows for a better differentiation of specific emotions, such as fear and anger [46].

From a physiological perspective, the distinction of fear among other emotions is not new [47]. However, to the best of our knowledge, there are only two fear recognition systems based solely on physiological information and self-reported labels. On the one hand, Bălan *et al.* [48] used all signals available from the database for emotion analysis using physiological signals (DEAP) [49] to provide a specialized fear recognition system. They achieved a fear accuracy detection rate below 90%, although they also considered EEG, which is not currently feasible as an inconspicuous wearable device. On the other hand, in our own previous research [50], only three physiological variables available from the multimodal analysis of human nonverbal behavior in real-world settings data set (MAHNOB) [51] were used, obtaining a fear recognition accuracy rate of up to 76.67% for a subject-independent approach using data from 12 women volunteers. Other works presented in the literature are based on valence and arousal quadrant classification rather than binary fear classification. For instance, Zao *et al.* [52] developed a valence and arousal classification system and obtained an accuracy of 75.56% for a subject-dependent approach. Hassan *et al.* [53] proposed a deep learning emotion recognition-based method and obtained an accuracy up to 89.53% for a subject-independent model considering five discrete emotions (happy, relaxed, disgusted, sad, and neutral). Although these latter works also employed a reduced set of physiological signals, they were focused on a different use case than the one pursued in this research.

Regarding the use of speech signals, emotion detection has been widely reported in [54] and [55]. The lack of existing speech corpora with strong elicited fear in real situations is a problem in speech emotion research. However, a few studies have managed to achieve results in this regard. For instance, Clavel *et al.* [56] developed an audio-based abnormal situations detection system for movie clips. Their results achieved up to 70.3% accuracy for fear detection via a leave one trial out (LOTO) strategy for 30 movies. In [57], they performed emotion detection with paralinguistic cues in a dialog corpus containing real agent-client recordings obtained from a medical emergency call center. As a result, they achieved a recognition rate with up to 64% accuracy for fear recognition.

When dealing with emotion recognition combining different data modalities, some comprehensive reviews can be found presenting current state-of-the-art data fusion techniques [58], [59]. These works state the need for: 1) novel approaches to advance the community knowledge on the multimodal casuistry and 2) subject-independent emotion recognition models to ease the further deployments under real-life conditions. They also agree on the potential performance improvements with multimodal approaches compared to unimodal ones. In fact, recently research in multimodal experimentation has been on the rise. For instance, Cimtay *et al.* [60] proposed a hybrid multimodal fusion emotion recognition system, including facial expressions, GSR, and EEG. Their results yielded a maximum subject accuracy of 91.50% and a mean accuracy of 53.80% using a leave-one-subject-out (LOSO) strategy and a publicly available database (DEAP) for different emotion detection use cases, such as angry, disgust, afraid, happy, neutral, sad, and surprised. Moreover, they created their own data set with which they achieved a maximum subject accuracy of 81.2% and a mean accuracy of 74.2% using a LOSO strategy for three emotion classes, i.e., sad, neutral, and happy. In [61], a weighted-based fusion strategy accompanied by transfer learning techniques was applied for multimodal emotion recognition using EEG and spontaneous spatial expression detection. The work employed a LOTO subject-dependent configuration and reported an average accuracy up to 69.75% and 70.00% for the valence and arousal classification, respectively. In addition to these works, more research can be found regarding multimodal data fusion for stress-related use cases [62], [63].

Analyzing these related works, most emotion recognition systems do not target the fusion of physiological and auditory modalities nor consider vulnerable groups, such as GBVVs. Specifically regarding such bimodal fusion of physiological and vocal information, the only work found is in [64], to

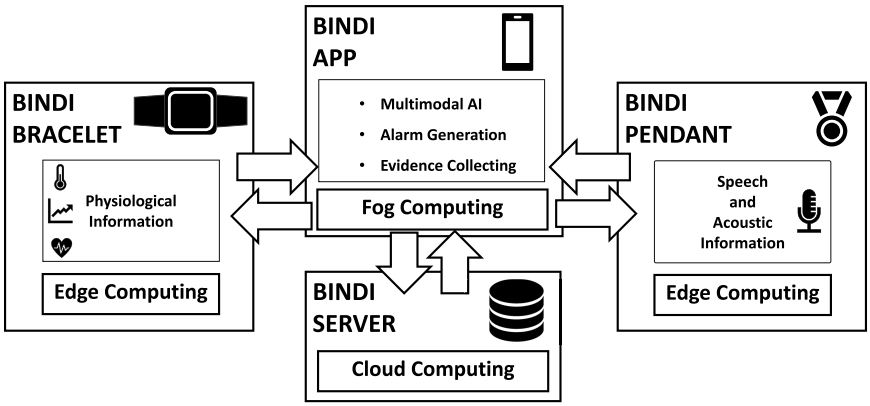

Fig. 2. Simplified Bindi architecture.

the best of our knowledge. This work considered different data fusion schemes and achieved an average accuracy of up to 55.00% for a subject-independent strategy using a feature fusion when targeting a valence and arousal binary classification. Consequently, there is a current need for research on these topics, which this work aims to deepen.

D. Open Public Multimodal Data Sets

There are various useful available databases in academia providing emotional labels together with auditory or physiological variables, such as DEAP [49], MAHNOB [51], WESAD [65], AMIGOS [66], FAU, Reg, and Ulm TSST Corpora [67], and BioSpeech [68]. However, they are not specifically intended for fear emotion detection, as most of them are based on a general framework in which the target is a set of emotions [69]. That means that fear-related samples are scarce for dealing with robust fear recognition. Moreover, no difference between men and women for their proposed intelligent emotional recognition systems is included. This latter fact is essential given that stimuli interpretation is strongly affected by gender [42]. Therefore, as concluded in [50], one of the main shortcomings of generating fear detection systems is the lack of adequate data sets, which should provide well-balanced labels (fear/not-fear) with a sufficient number of volunteers, real emotions, a gender perspective, and considering the target group of people.

As a result of the previously discussed limitations found in the literature regarding public multimodal data sets, UC3M4Safety is generating a database specially designed for fear detection in the GBV use case—the so-called UC3M4Safety database. In this database, women volunteers are exposed to a set of audio-visual stimuli that elicit specific emotions. Physiological and auditory variables are recorded during the viewing, and annotations and self-reports from the users are also registered. A more detailed description of the UC3M4Safety database is provided in Section V-A.

III. System Hardware Architecture

A simplified system architecture of Bindi is presented in Fig. 2. The following sections provide a technical overview regarding each system component.

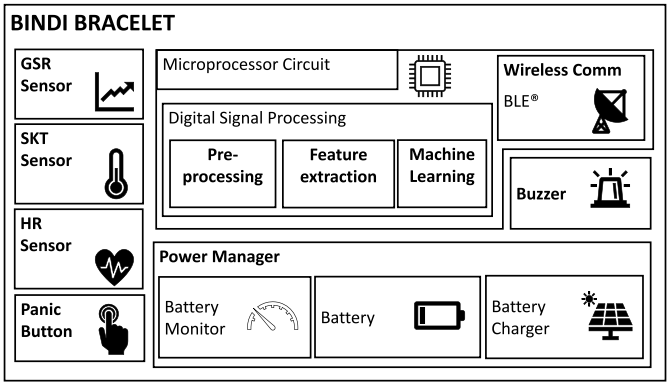

Fig. 3. Simplified Bracelet architecture.

A. Edge Computing

The edge devices in the Bindi architecture are the Bracelet and the Pendant. These two elements are described as follows.

1) Bracelet: This device runs an embedded intelligence engine for fear detection based on physiological information. Fig. 3 shows the hardware components integrated into this device, which can be classified into four groups: 1) physiological sensors; 2) actuators; 3) power manager elements; and 4) the microprocessor unit. The latter is the nRF52840 system on chip ARM Cortex-M4, an ultralow-power-consumption microcontroller unit with 1-MB memory flash and 256-kB RAM, a single-precision floating-point unit, a Thumb-2 instruction set, a 64-MHz clock, and some integrated peripherals (USB, UART, SPI, I2C, I2S, ADC, PDM, and AES). Note that the radio-frequency module through Bluetooth low energy communication is also integrated within this host unit. Regarding the power manager elements, the BQ2407xT and MAX17055 components by Texas Instruments and Maxim Integrated, respectively, are used. These two integrated circuits are responsible for charging and monitoring the battery. The Bracelet is equipped with a conventional electro-mechanical button for manual user activation, acting as a panic button. The details of the physiological sensors included are as follows.

1) *HR:* This is based on a photoplethysmography sensor that detects blood volume pulse (BVP) changes by measuring the absorption of light emitted through

the skin. This sensor is the MAX30101 high-sensitivity reflective pulse oximeter, with different integrated LEDs (red, green, and infrared), 18-bit ADC, I2C communication, and digital noise cancellation. More particularities about this sensor are described in [70].

2) *GSR:* This sensor utilizes two electrodes to measure the skin conductivity through a dc exosomatic measurement. More particularities about this sensor and its analog front end are described in [71]. Note that data provided by this sensor are analog, and then, the acquisition is performed using the ADC of the microcontroller unit.
3) *SKT:* The MAX30208 component is proposed to acquire a reliable SKT measurement. This integrated circuit is defined as a clinical-grade sensor for wearable applications, providing an accuracy of $\pm 0.1\,^{\circ}\mathrm{C}$ over a $30\,^{\circ}\mathrm{C}$ to $50\,^{\circ}\mathrm{C}$ temperature range. It integrates 16-bit ADC and I2C communication.

The previously discussed physiological variables were chosen due to their proven strong relationship with emotion recognition [38] and their ease of implementation in wearable devices. The latter point is particularly relevant and led us to discard other typical physiological sensors used in the field (such as EEG), which do not meet the inconspicuousness requirement. The digital sensors considered provide integrated analog front-end circuitry but also include digital processing to relieve the host processing unit of some preprocessing tasks. The digital signal processing pipeline within the Bracelet entails both, the acquisition and filtering of the physiological signals and the feature extraction and inference stages. These steps and their particularities are detailed in Section IV-A.

2) Pendant: This device captures audio and speech information, which is fed to an intelligent engine for fear detection. Note that such as engine is currently running in the Bindi app, in the future, it will be executed in the Pendant. The Pendant has the same hardware architecture as the Bracelet but integrates a microphone instead of physiological sensors. Its architecture is shown in Fig. 4. The microphone is based on a microelectromechanical system with an omnidirectional audio sensor. This part includes a capacitive sensing element and an integrated circuit interface, allowing a digital signal to be obtained directly. The digital signal processing pipeline within the Pendant entails both, the reception and filtering of the auditory signals (audio and speech) and the wireless transmission to the Bindi app. Note that due to the limited bandwidth of the wireless communication, the audio is compressed prior to being transmitted.

B. Fog Computing

The fog computing within Bindi is represented by the Bindi app running on a smartphone. It provides an end-user graphical interface and performs the following technical functionalities.

1) It requests physiological and auditory data from the Bracelet and the Pendant, respectively, according to the data processing pipelines implemented, as discussed in Section IV.
2) It handles the alarm triggers (SMS/protection unit or emergency services alerts) and logs them into the server

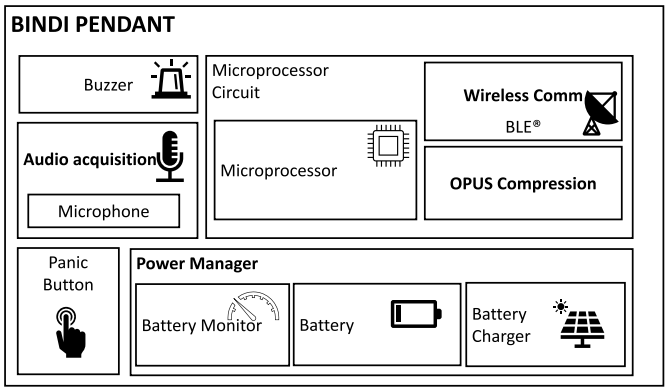

Fig. 4. Simplified Pendant architecture.

based on the intelligent engine response or the manual panic button.
3) It keeps track of each user's location using GPS.
4) It manages secure communications with the server adapted by the current smartphone battery status.
5) It collects and uploads auditory and physiological ciphered data to the cloud as evidence of an alleged crime if the alarm is triggered.
6) It performs the feature extraction and inference processes for the auditory monomodal system. Moreover, it handles different data fusion strategies, which are discussed in Section IV.

C. Cloud Computing

The cloud computing part is where the Bindi server comes into operation. The Bindi server implementation consists of a MongoDB[6] database and a NodeJS[7] Web application server. This Bindi server stores the information captured in the edge with three main goals. First, it serves as an activity monitor, indicating potential problematic situations regarding victims' long-term affective evolution for people supervising the well being of the users. Second, it stores encrypted data, serving as digital evidence in an eventual trial. Third, it makes decisions after the alarms are triggered by the following predetermined safety procedures.

IV. BINDI Data Processing Pipeline

As stated in Section I, one of the key objectives of this work is to validate the data processing chain within Bindi, from data acquisition to alarm generation. Different arrangements of the system components have been applied and compared to achieve this goal. This fact has led to a design space exploration of different multimodal (physiological and auditory information) system architectures. Specifically, three arrangements have been evaluated.

1) The first version is Bindi 1.0 [72], which is based on a hierarchical strategy. In this version, physiological information is continuously collected by the Bracelet, which runs a lightweight monomodal physiological

[6]https://www.mongodb.com
[7]https://nodejs.org/es/

intelligence engine. When it detects that the user is experiencing fear, it triggers a prealarm 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 auditory signal is then sent to the Bindi app to perform fear detection using a speech-based monomodal 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 Bindi server.

2) The subsequent version, Bindi 2.0a, is based on the same two monomodal data processing pipelines in Bindi 1.0 but at the final decision stage applies a late fusion technique rather than a hierarchical agreement or confirmatory strategy [73]. It inherits the prealarm functionality from Bindi 1.0 for low-energy consumption for the microphone.
3) As a variation of Bindi 2.0a, Bindi 2.0b follows the late fusion scheme introduced in Bindi 2.0a but bases it on continuous physiological and auditory data acquisition, meaning that the prealarm functionality is not enabled.

The following sections detail the physiological and auditory data processing pipelines and the different data fusion strategies considered in the three arrangements evaluated. The particular nature of the data types (physiological, speech, and audio) entails different challenges. Thus, the data processing schemes, methods, and feature extraction techniques are tailored to each signal. Note that the feature extraction process refers to a typical machine learning step that transforms filtered signals' data into numerical features that can be processed more quickly by the classification/regression algorithms while preserving and highlighting the information from the original data. The experimental results in this article are an account of the validation process performed offline, to evaluate the functionality of the data processing pipelines and later embed such modules in the architecture, balancing the tradeoffs observed. A preliminary acoustic event detector data processing pipeline is also described in this section but has not been integrated into the arrangements evaluated. This acoustic detector is intended as a proof of concept from which some interesting experimental results are presented in Section VI-B, paving the way for Bindi 3.0.

A. Physiological Data Subsystem

The first physiological data processing stage is signal acquisition and windowing. In this first stage, the sampling frequency and signal segmentation are critical parameters because they represent a tradeoff between the amount of information that can later be extracted and the resource usage. In our case, the selected sampling frequencies are 100, 10, and 5 Hz for the BVP, GSR, and SKT, respectively. These frequencies are adequate to capture signal dynamics with the appropriate temporal resolution. For signal segmentation, an overlapping fixed-length strategy is used. This segmentation approach is the most common method to process physiological signals in emotion recognition systems [38]. Note that the effect of the selected window length directly impacts two key factors. First, this parameter determines the frequency resolution available for the feature extraction stage. Second, it is related to the ability to extract emotion-related events. There is no agreement in the literature about the optimal window length for emotion recognition analysis, as it also depends on the subject and the emotion felt. The proposed physiological data processing chain uses 20-s windows with a 10-s overlap. This configuration provides a frequency resolution of 0.05 Hz, which results in a good tradeoff between the data storage and physiological information available to be extracted. However, some limitations appear when dealing with this specified length. For instance, some GSR phasic events cannot be completely separated from the tonic component, as the maximum event duration is 30 s [74].

Once the signals are captured and segmented, the filtering stage removes the out-of-band noise. The BVP filter architecture has been selected through design space exploration considering the specifications required and resource usage [70]. Specifically, the BVP filter is a two-stage FIR filter, with high-pass and low-pass stages with 0.5- and 4-Hz cut-off frequencies, respectively. Afterward, the signal is scaled by employing an automatic gain control to limit the amplitude and improve the peak detection. The GSR is filtered to preserve information below 1.5 Hz, which is the maximum frequency for phasic activity. Moreover, it is downsampled to 5 Hz, reducing memory and computation requirements while increasing resolution. The GSR filter is also applied to the SKT signal to store only one set of filter coefficients.

Feature extraction is the next stage in the processing pipeline. This block extracts the information contained in the physiological signals. Specifically, there are 25 features for BVP (two in the time domain, nine in the frequency domain, and 14 nonlinear ones), 17 features for GSR (six in the time domain, three in the frequency domain, and eight nonlinear ones), and six features for SKT (four in the time domain and two in the frequency domain). An extensive description of the features is provided in [50]. For classification, a lightweight K-nearest neighbors (KNNs) binary supervised machine learning algorithm is used. During the training stage, cost-sensitive learning is applied by modifying the misclassification cost of KNN, which increases the sensitivity, i.e., the system will be less likely to omit a dangerous situation for the use case [75]. Moreover, the different hyperparameters are optimized using a sequential model-based optimization technique [76]. Some of the nonlinear features include recurrence analysis computation [77]. This could lead to unaffordable computational complexity for a constrained wearable device. Accordingly, a sequential forward feature selection algorithm is used during the training stage. This process removes redundant information, lowering the computational load and resource usage for the inference [78]. Finally, the physiological data subsystem output is a binary label every 10 s. This physiological pipeline has been tested in previous work using a public data set [50].

B. Speech Data Subsystem

The speech data processing includes the following fundamental modules: voice activity detection (VAD),

frequency domain filtering, feature extraction, normalization, and a neural-network-based classifier.

A basic lightweight VAD module [79] based on spectral energy is employed to detect and remove silent parts of speech signals where the posterior feature extractor would not extract any relevant speech information due to the absence of speech. Silence detection is crucial for correct functionality of the device, as women in dangerous situations frequently react with shock and remain silent.

In combination with the VAD module, to ease the handling of the signals while keeping all significant information from the speech data, it is necessary to downsample the signals at 16 kHz. Next, a low-pass filter is applied at 100 Hz to remove low-frequency noise captured by the microphone and possibly caused by air conditioning and electrical network buzzing, among other factors, during laboratory experiments. Afterward, the signals are processed, starting with a low-pass filter at 8 kHz. Then, the speech feature extractor computes 38 speech features dedicated to emotion detection using a 20-ms window with 10-ms overlapping, both of which are standard values from the literature. Among the features considered are pitch, Mel frequency cepstral coefficients, formants, energy, and additional spectral features, all of which are calculated through the librosa Python toolkit [80]. The features are aggregated per second by computing their mean statistics to be later normalized. Preliminary ablation experiments are performed before fixing this 1-s aggregation, varying the temporal context of the aggregated speech features for $1, 5$, and 10 s.

Feature normalization is done by applying the z-score mean and standard deviation values from the baseline features extracted when the user is in a resting or neutral state. Other normalization schemes (e.g., per video, per user, and traditional z-score) are informally tested before considering the basal state normalization described.

The normalized aggregated features are fed into a user-adapted neural network classifier trained for fear detection. This subsystem generates a binary label every 1 s. The labels predicted by the monomodal speech subsystem every second are smoothed in time using a 7-s long window to maintain consistent and stable detection.

C. Data Fusion Strategy

Data fusion is a powerful way to improve the robustness of the multimodal intelligence engine in Bindi. The late fusion strategy in Bindi 2.0a and Bindi 2.0b is fed from the binary labels provided by the physiological and speech monomodal intelligence engines, as shown in Fig. 5.

As discussed previously, the physiological and speech monomodal subsystems estimate a binary label, $y_k^m \in \{0, 1\}$, for every time window k and modality $m \in \{\text{phy}, \text{sp}\}$, with phy and sp referring to the physiological and speech subsystems, respectively. Note that each of the modalities uses a different time window length T_m in seconds, due to their specific peculiarities. Bindi is intended to output a response per time period n (each one of same length L), with $n \in 1, 2, \ldots$, in seconds. Thus, an estimation of fear probability p_n^m for the nth time period and the mth modality is computed as

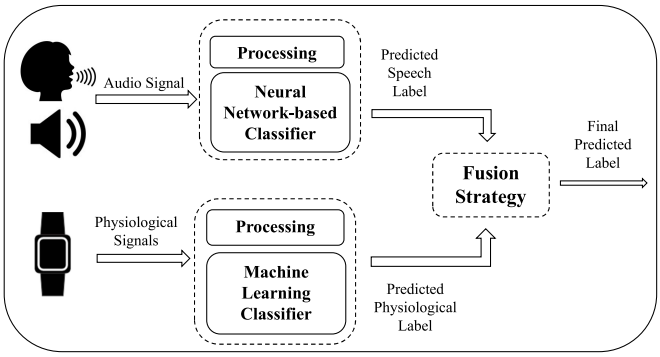

Fig. 5. Data fusion block diagram.

$$p_n^m = \frac{\sum_{k=1}^{K_m} y^m_{[K_m \cdot (n-1)+k]}}{K_m} \tag{1}$$

where $K_m = \lfloor (L/T_m) \rfloor$, 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 based on p_n^m can be calculated as

$$Y_n^m = \begin{cases} 0, & \text{for } p_n^m < \text{th}_m \\ 1, & \text{otherwise} \end{cases} \tag{2}$$

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 th_{phy} and th_{sp} values are discussed in Section VI-A.

As a metric to represent how confident each monomodal system is for the class label predicted in a given period, entropy h_n^m for the nth time period and mth modality is calculated as

$$h_n^m = -\left[p_n^m \cdot \log\left(p_n^m\right) + \left(1 - p_n^m\right) \cdot \log\left(1 - p_n^m\right)\right]. \tag{3}$$

On this basis, three late fusion strategies are studied to produce fused system response Y_n^f for the nth time period.

1) *Case 1 (Lowest Entropy):* The system's response corresponds to the binary label produced by the monomodal system with the smallest entropy, i.e., the most confident one. To this end, fused fear probability p_n^f for the nth time period is calculated as

$$p_n^f = \begin{cases} p_n^{\text{phy}}, & \text{if } h_n^{\text{phy}} < h_n^{\text{sp}} \\ p_n^{\text{sp}}, & \text{otherwise.} \end{cases} \tag{4}$$

Next, applying the same rationale as in (2), a fused binary label is obtained as

$$Y_n^f = \begin{cases} 0, & \text{if } p_n^f < \text{th}_f \\ 1, & \text{otherwise} \end{cases} \tag{5}$$

where for now, th_f is the conventional 0.5.

2) *Case 2 (Inverse Entropy Weighted Combination):* Fused fear probability p_n^f for the nth time period is computed as a weighted sum of probabilities, as given by

$$p_n^f = \sum_m w_n^m \cdot p_n^m \tag{6}$$

where

$$w_n^m = \frac{1/h_n^m}{\sum_m 1/h_n^m}. \tag{7}$$

Next, a fused binary label is obtained according to (5).

3) *Case 3 (Logical OR):* The system response corresponds to the logical OR computation over the binary labels for each monomodal system. That is

$$Y_n^f = Y_n^{\text{phy}} \vee \ Y_n^{\text{sp}}. \tag{8}$$

The three fusion strategies are based on the literature (e.g., [61]) and are proposed as a tradeoff 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 tradeoff between the probabilities and entropies for each monomodal subsystem. Thus, the confidence of this last strategy might be higher than that of the others.

D. Acoustic Information Subsystem

This section proposes an audio processing pipeline for acoustic scene threat detection. As stated before, this component has not yet been included in the arrangements studied in this article for Bindi and is here conveyed as a proof of concept for further versions of Bindi. This subsystem is based on the architecture presented in [81]. Its main task is to detect whether the sounds recorded from the microphone represent a threat to the user's safety according to the use case.

The acoustic event detection system proceeds as follows. First, the audio signal is normalized, just as for the speech pipeline. Second, a log-Mel spectrogram is computed to obtain a time–frequency representation of the signal in an image form to later feed it to the network. Thus, an initial spectrogram is computed through a short-time Fourier transform (STFT) with the following parameters: a window size of 25 ms, window hop of 10 ms, and Hanning window. The frequency dimension of the spectrogram is mapped to 64 Mel bins to cover frequencies ranging from 125 to 7500 Hz and the amplitude is transformed into a log scale with an offset of 0.001. These features are framed into examples of 0.96 s with an overlapping of 50%. Each example covers 96 frames of 10 ms each and 64 Mel frequency bands. Therefore, the dimensions of these features are 96×64. The resulting features are fed into a pretrained convolutional neural network (CNN) to detect the audio events in a scene.

The selected model for this task is YAMNet. Specifically, the MobileNet_v1 [82] depthwise separable convolution architecture is considered. This model has been pretrained on 521 classes of the AudioSet YouTube corpus [83], a general-purpose multilabel sound event classification database, and is ready to perform inference for the detection of acoustic events. The performance of these types of networks has been widely studied in the field of sound event detection [84].

The procedure to feed the network is as follows. First, the 96×64 patches from the feature extraction stage are transformed into a 3×2 array for the 1024 kernels of the top convolutional layer. After being processed through the feature extraction layers, these examples are averaged to obtain a 1024-dimension embedding. Then, a logistic layer performs the classification in 521 target classes.

V. Experimental Methodology

A detailed description of the UC3M4Safety database and WEMAC data set is provided in this section, with a discussion of the technical aspects that affect the generated models. This description is followed by a detailed analysis of the specific training and testing methodologies applied for both modalities to provide a satisfactory fit for the experimental data.

A. UC3M4Safety Database Technical Aspects

As introduced in Section II-D, the UC3M4Safety team is currently developing the UC3M4Safety Database, a novel multimodal database considering women's emotional responses to audio-visual stimuli [85]. The process began with the selection and validation of appropriate stimuli [37], which have been publicly released [86], [87]. The UC3M4Safety database comprises the generation of different multimodal data sets, some of which contain the physiological and auditory variables of the subjects being monitored. Note that one of the main goals is to understand and obtain models of the relationship between the physiological and auditory activation mechanisms in GBVVs.

The first data sets contain a list of 79 audio-visual stimuli that were labeled and selected from among 160 clips with criteria of quality, balancing different emotions, and agreement among viewers [37], [87]. The 160 audio-visual stimuli were selected by different expert judges [85] to achieve a good balance between fear and the rest of the emotions elicited. This stimuli selection was evaluated by 1332 independent people, obtaining a percentage of 44.44% for fear and 55.55% for the rest of emotions, as shown in Table I. The other data sets contain the results from the experiments performed in a laboratory environment with only women volunteers who had never experienced GBV. These are employed in this article for the space design exploration of the different system architectures of Bindi. The other data sets of the UC3M4Safety database consist of capturing data from GBVVs in laboratory conditions and from GBVVs and non-GBVVs in real-life conditions. These two latter data sets are currently being recorded.

The details of the employed WEMAC data set are as follows: 47 volunteers were exposed to 14 validated audio-visual stimuli through a virtual reality environment with the Oculus Rift-S Headset[8] to maximize the immersive experience and, consequently, achieve better emotion elicitation. These video clips were selected from a pool of 28 audio-visual stimuli, resulting in two batches of videos. Note that these videos were extracted from a pool of 79 audio-visual stimuli in [87]. During the experimentation, the volunteers self-reported information

[8]https://www.oculus.com/rift-s/

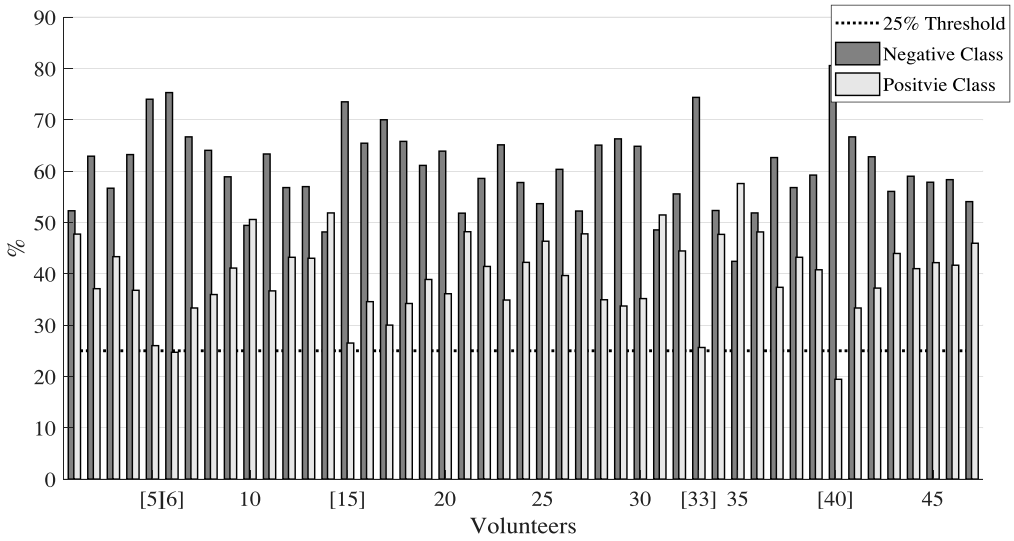

Fig. 6. Positive and negative class distributions for the binarized self-reports in WEMAC. Volunteers in brackets are those excluded.

based on well-known emotion-related labeling methodologies, as described in Section II-C. Also, their physiological information was collected using BioSignalsPlux,[9] one of the most common sensing research toolkit systems in the physiological monitoring field, used for the sake of comparison with the proposed sensors. The speech data were captured using the Oculus(R) Rift-S Headset microphone.

The experimental protocol followed in the laboratory was as follows: Every volunteer was exposed to the collection of audio-visual stimuli, and their physiological information was captured. Every stimulus visualization was preceded by a neutral clip to reset the user to a neutral state. This step was followed by different labeling stages, which consisted of descriptive speech recording and interactive self-reports. In the descriptive speech stage, each volunteer was requested to provide a voice recording by answering questions that were intended to make them relive the emotion felt so that it was reflected in the recorded voice signal. In the interactive self-reports, each user was presented with different discrete possible emotion labels from which they were to select the primary emotion felt. As mentioned before, every video clip was expected to elicit only one targeted emotion. Thus, the information obtained by the labeling stages matched the target emotions with more than 90% agreement with respect to the labels reported by the volunteers. For further information on WEMAC, we refer the readers to [11].

To adapt WEMAC to our purposes of validating and evaluating the different system architectures of Bindi, we first binarized the reported discrete emotions to transform the modeling problem into a binary classification, where "1" (positive class) represented *fear* and "0" (negative class) any other emotion. After the experiment was conducted, it was observed that some particular volunteers presented with a considerably unbalanced distribution in their self-reported labels, as shown in Fig. 6. Therefore, it was decided to exclude volunteers 5, 6, 15, 33, and 40 from the evaluation since they had only around 25% of the positive class distribution. Consequently, the evaluation was to be performed with only 42 of the 47 initial volunteers. The class distribution for these 42 volunteers was around 60% and 40% for the negative and positive classes, respectively. This distribution fits the information presented in Table I for the different emotions. As stated in Section I, the UC3M4Safety team is working toward enlarging the size of this data set by increasing the number of volunteers.

[9]https://biosignalsplux.com/products/kits/researcher.html

TABLE I
EMOTIONS ELICITED BY THE UC3M4SAFETY DATABASE AUDIO-VISUAL STIMULI VALIDATED BY AN INDEPENDENT GROUP OF 1332 PEOPLE, IN TOTAL PERCENTAGE

Fear	44.44%	**Tedium**	2.22%
Joy	8.89%	**Tenderness**	6.67%
Hope	2.22%	**Calm**	11.11%
Surprise	4.44%	**Disgust**	8.89%
Anger	4.44%	**Sadness**	6.67%

B. Training and Testing Considerations for the Monomodal Subsystems

Some points had to be considered to design the training and testing strategies of the two monomodal subsystems. First, according to the database design, it should be noted that physiological data were gathered during the stimulus elicitation, whereas speech recording was registered during the subsequent speech annotation. That means that the physiological and speech data were not aligned in time in WEMAC. However, both data types had to be fused inBindi 2.0b for every emotional reaction per user or experiment, unlike for Bindi 1.0 and Bindi 2.0a where the fusion was conditioned to the physiological prealarm. Therefore, we obtained a single p_n^m per experiment and modality, according to (1); note that L is the length of the audio-visual stimuli for the physiological modality and the total length of the audio recording for the speech modality. During the labeling, the volunteers were requested to relive the emotions felt during the stimulus elicitation, so it was assumed that the correspondence was solid enough

between both time instants. However, this assumption will need further validation when the rest of the subsets in the UC3M4Safety Database become available.

Second, for the train-test split, a LASO strategy was applied. This was an adapted subject-semiindependent approach procedure for training the 42 models required, i.e., one per user. This approach was chosen due to the fact that the subject personalization provided by LASO is crucial for an emotion detection model such as ours [12]. Thus, each model was trained with all available data from the rest of the users plus half the instances of the subject to be tested, particularly, the data acquired from the first seven audio-visual stimuli. The rest of the utterances of the last seven videos of the session were to be used as test samples. Thus, the test data were not seen during the training stage but some information about the subject was obtained, as intended.

Third, regarding specific training particularities, for the physiological monomodal subsystem, the same misclassification cost of 1.6 to the positive class to deal with the commented class imbalance was considered for all physiological models generated. This cost was fixed by an experimental parameter sweep. Moreover, the training was validated by a stratified k-fold cross-validation strategy, with $k = 5$. Finally, the normalization applied for the data set was based on the z-score technique applied to the features extracted from all volunteers.

For the speech monomodal subsystem, the classifier consisted of a shallow lightweight neural network with input, fully connected hidden, and fully connected output layers. The network had 38 units in its input layer, i.e., one per feature. The number of hidden units in the dense layer was fixed to 250 to avoid largely increasing the computational cost but achieve fairly good prediction rates. The output layer yielded one predicted label as an output. All samples, except the ones from the user of interest, were used to train the model during 300 epochs, with early stopping after a 30-epoch plateau in the model validation loss, a binary cross-entropy loss function, using Adam optimizer, and a learning rate of 0.001. Then, samples from the user of interest (half of the ones available according to the LASO strategy) were used to fine-tune the model for a maximum of 100 epochs, with an early stopping approach (i.e., stopping after a 10-epoch plateau in the model loss). Regarding the z-score normalization used, the features extracted from the speech recordings of the sixth audio-visual stimuli were used as the baseline. This video was expected to elicit a calm emotion and was assumed to evoke a neutral state in the user.

Finally, regarding the testing procedure, as discussed in Section IV-C, the monomodal subsystem's outputs were arrays of binary labels. Specifically, for the UC3M4Safety database, the length of the arrays was equal to dividing the duration of each stimulus by the monomodal sampling periods, i.e., 10 and 1 s for the physiological and speech subsystems, respectively. Afterward, those collected arrays were processed by calculating the probabilities and their corresponding binary labels by applying the physiological ($formulaeth_{\text{phy}}$) and speech ($formulaeth_{\text{sp}}$) thresholds. The data fusion strategies proposed also generated their corresponding binary labels, as described in Section IV-C. The evaluation metrics selected, i.e., the accuracy and $F1$-score, fed on the hard labels obtained. The accuracy could fairly represent the prediction rates since the class imbalance was low. The $F1$-score was considered to deal with the slight imbalance observed. Although the $F1$-score should be a good metric for a detection problem such as the one addressed—in which the number of positives should have been relatively low in comparison with the negatives—the experimental setting considered here was almost balanced, and therefore, this metric was not as significant as it was expected to be when testing with data captured in real-life conditions.

VI. Experimental Results

This section presents the experimental results regarding the prediction of fear using WEMAC for the different configurations of the system, as discussed in Section IV. Note that this is the first time this database has been used; therefore, these results represent the first step toward real (nonacted) fear emotion detection from physiological and auditory variables for the problem of GBV and are meant as a baseline for future developments. Additionally, an analysis of the acoustic events contained in the audio-visual stimuli is introduced.

A. Fear Detection Analysis

The first analysis concerns the performance of the physiological and speech subsystems working independently in a continuous setting, i.e., taking into account all samples. This experiment was essential to determine the thresholds, th_{phy} and th_{sp}, which convert the set of binary labels predicted during a video visualization, into a single binary label for such period [see (2)]. This step was relevant to determine whether the architecture was more or less prone to false alarms, regardless of the version of Bindi being considered. Thus, each parameter was swept in the range [0.3, 0.6] with steps of 0.1 while generating the corresponding 42 monomodal subsystems following the LASO approach. In this regard, Fig. 7(a) and (b) shows the th_{phy} and th_{sp} values versus the accuracy and $F1$-score average metrics for the 42 testing groups in the physiological and speech subsystems, respectively.

Analyzing Fig. 7(a), it can be observed how the $F1$-score decreases as th_{phy} grows, whereas the accuracy remains rather stable. Note that the F1-score depends to a great extent on the number of true positives (TPs) predicted but mostly disregards the true negatives (TNs). Thus, if TPs increase and the sum of false positive (FP) and false negative (FN) rates decrease, then the $F1$-score increases. This tradeoff caused the behavior observed, where the lower the th_{phy} gets, the higher the $F1$-score becomes. According to this analysis, th_{phy} was fixed to 0.40, obtaining 66.66% and 64.60% for $F1$-score and accuracy, respectively. The reason behind choosing this value was the good compromise observed between both metrics and the fact that missing a TP could be dramatic for the GBVV. The combined multimodal system should also refrain from triggering false alarms to avoid overwhelming the institutions in charge of protecting the users, and this is why the speech subsystem was chosen to be more conservative in this regard. Fig. 7(b) shows how the $F1$-score and accuracy

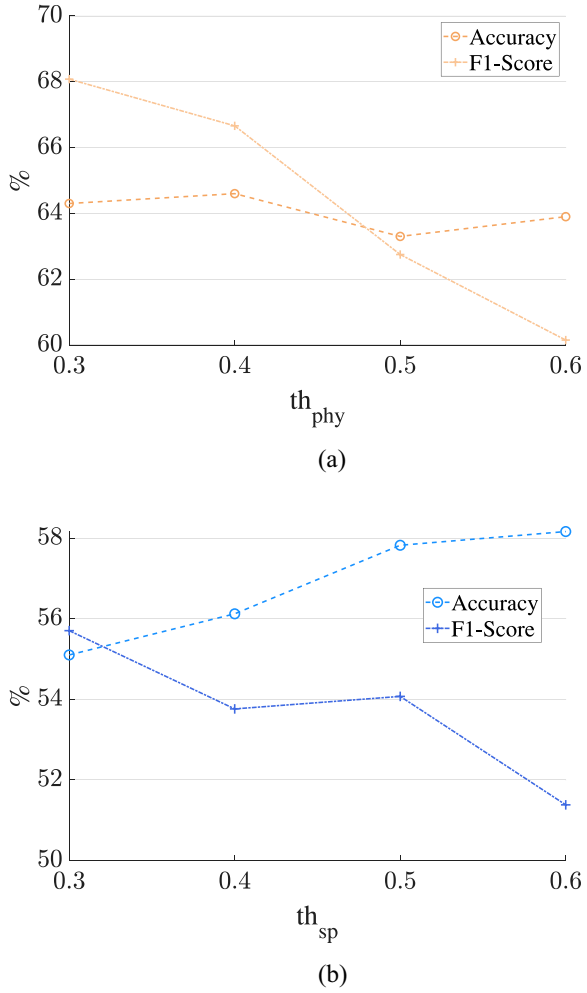

Fig. 7. Parameter sweep for (a) th_{phy} and (b) th_{sp} in the physiological and speech monomodal subsystems, respectively.

began to diverge from 0.50 onward for the speech subsystem. Therefore, th_{sp} was fixed to this value, obtaining 54.07% and 57.82% for the $F1$-score and accuracy, respectively. Note that the accuracy could be increased by choosing a higher th_{sp}.

Once th_{phy} and th_{sp} were fixed, we studied the average performance prediction over the 42 testing groups for the different architecture configurations, as shown in Fig. 8. The physiological monomodal subsystem achieved the highest accuracy of 64.63%, surpassing even the fusion schemes. For the $F1$-score metric, this subsystem also provided the second highest rate of 66.67%. This behavior could be related first, to the bias introduced toward detecting the positive class with the misclassification cost of the classifier and second, with the parameter sweep of th_{phy}. The speech monomodal subsystem provided significantly lower metrics than the physiological subsystem. This fact could be related to the limited number of samples available to train the neural network and, possibly, some fading of the emotion felt when the samples were taken. This situation caused Bindi 1.0 to provide the lowest metrics since the final system response relies on the speech subsystem. Bindi 2.0a and Bindi 2.0b both provided similar accuracies close to those of the physiological subsystem in most cases. However, Bindi 2.0b achieved the highest $F1$-score in all cases, especially with the logical OR data fusion. This latter strategy provided the highest $F1$-score of 67.59%, although the accuracy was limited. This performance of the $F1$-score could be related to the positive bias contributed by the physiological subsystem due to the lower th_{phy} chosen, which introduced a conservative bias toward not missing TPs at the cost of increasing FPs. However, as for the other architectures with fusion strategies, the speech subsystem may have been slightly deteriorating the system performance in terms of the $F1$-score and accuracy but preventing Bindi 2.0a andBindi 2.0b from producing too many FPs. Moreover, auditory information was expected to play an important role in detecting silences, which could mean that the user is in a state of shock caused by a GBV situation, and provide acoustic information about the environment. The meaning and consequences of these indicators over the real-life system performance should be thoroughly analyzed in the light of more robust metrics, such as in [88]. A short preview of this analysis and discussion of the confusion matrices obtained for each configuration can be found in the Appendix.

To elaborate on the results shown in Fig. 8, Table II 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 presented the lowest standard deviation, which could be seen as a good generalization, its scores were surpassed by most of the configurations, as previously stated. Moreover, it can be observed that the standard deviation values obtained are relatively high, especially for the $F1$-score. The cause is shown in Fig. 9, where the $F1$-score and accuracy are provided for each of the 42 tests and monomodal subsystems. It can be noted that some volunteers had an $F1$-score of zero for the speech subsystem. This situation occurs because the $F1$-score depends on the TPs detected and there were no positive predictions for some users.

B. Acoustic Information Analysis

This analysis aims to characterize the problem of GBV detection from an acoustics perspective since the development of an empirical description of the problem is important for its automatic detection. Thus, the acoustic information subsystem was applied to the audio signal of the audio-visual stimuli in WEMAC to observe the acoustic scene of a violent situation in the context of GBV. The results obtained appear in Fig. 10, where the occurrences of the YAMNet labels in the audio-visual stimuli of WEMAC are analyzed. This figure also shows the YAMNet labels found in the fear audio-visual stimuli of WEMAC. Thus, some labels were exclusively found in fear audio-visual stimuli, such as *heartbeats*, *explosions*, and *breathing*, whereas other labels never appeared for fear, such as *tender music*, *lullabies*, and *crowds*. There were also intermediate cases in which labels appear for both types of stimuli, such as spatial-contextualization labels (indoors or outdoors related), *animals*, *silence*, and *laughter*. Therefore, automatic classification of acoustic events seems to be promising as certain patterns can be deduced from extreme cases in

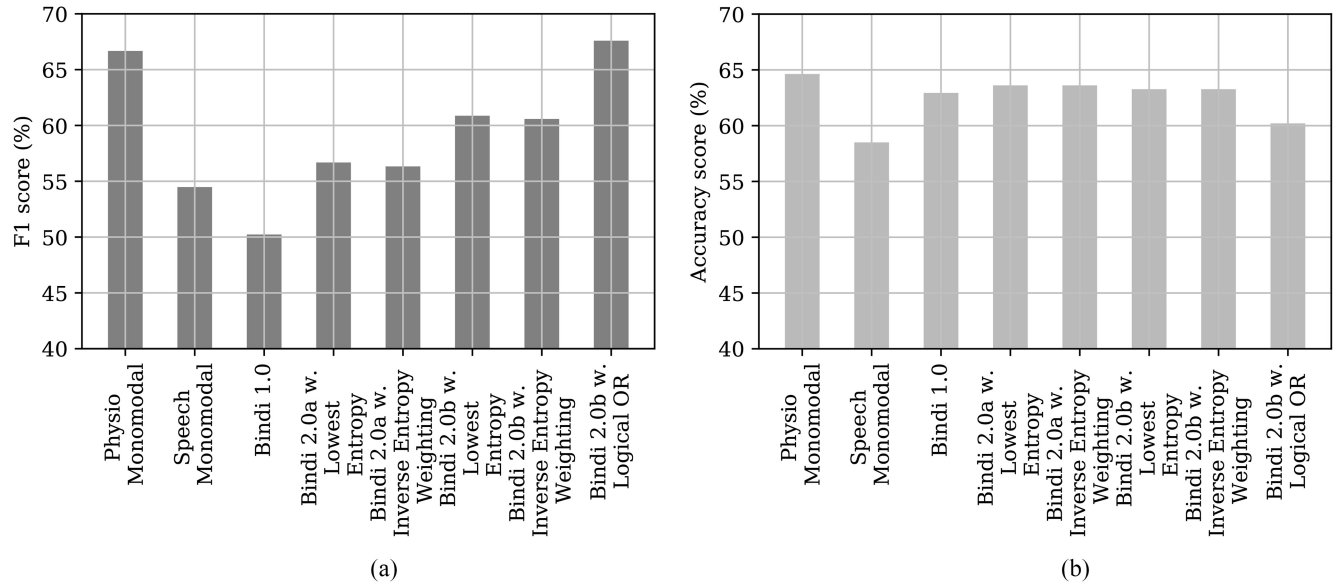

Fig. 8. Average performance analysis predicting over the 42 testing groups for the different architecture configurations. (a) $F1$ score. (b) Accuracy score. From left to right, the configurations are: physiological monomodal subsystem, the speech monomodal subsystem, Bindi 1.0, Bindi 2.0a with lowest 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.

TABLE II
AVERAGE PERFORMANCE ANALYSIS PREDICTING OVER THE 42 TESTING GROUPS. MEAN AND STANDARD DEVIATIONS (STD)

		Physiological Monomodal	Speech Monomodal	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	58.50	62.93	63.61	63.61	63.27	63.27	60.20
	Std	16.56	16.73	14.30	14.35	14.35	17.94	18.21	15.75

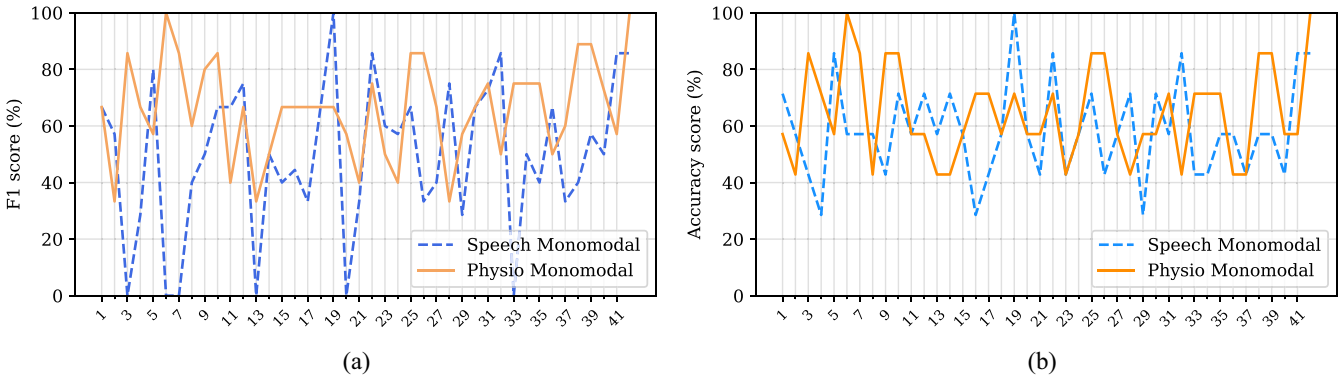

Fig. 9. Individual performance analysis for the two monomodal subsystems. (a) $F1$ score. (b) Accuracy.

which labels exclusively appear for one of the two types of audio-visual stimuli. It must be noted that YAMNet labels are very general themselves, i.e., they can appear to be related to many circumstances and scenes. Thus, they must be analyzed as a set, which is a feasible way to infer some qualities of the context of a particular scene, e.g., violence.

From this analysis, we can conclude that the information extracted from acoustic events can be very beneficial to disambiguate potential GBV situations detected automatically in Bindi with the rest of the sensors. The surrounding sound events of a scene can help infer its context, which is critical to determine whether the scene is violent or not. Thus, we expect the acoustic information subsystem to play an important role in the evaluation of data sets in the UC3M4Safety database, where volunteers are performing everyday activities out of the lab setting currently being explored.

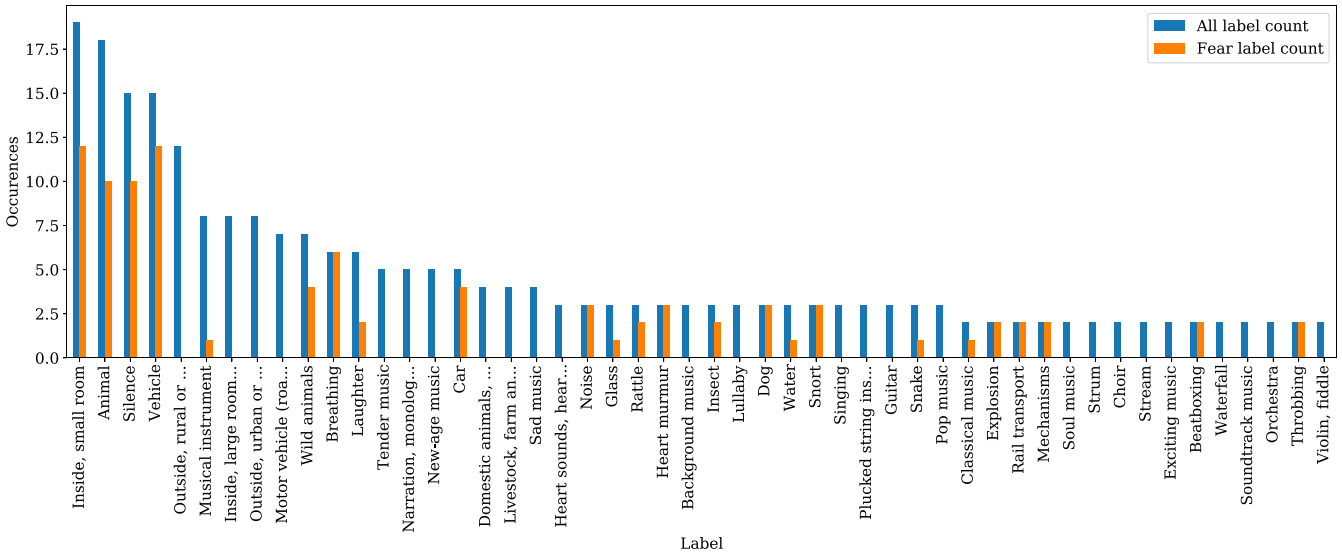

Fig. 10. Occurrences in absolute numbers of YAMNet labels in fear versus all audio-visual stimulus in WEMAC.

VII. BINDI Power Consumption

Power consumption management is a requirement for the design of a wearable system. In Bindi, an accurate measure of the state of the battery charge and autonomy of the two wearable devices is essential to ensure that the system works when needed. As commented on Section III-A, the two wearable edge devices integrate a battery charge monitor that provides information about power consumption during operation. Thus, the end user can be informed about the battery state to allow for the planification of the charge. This monitoring is performed by the Maxim Integrated MAX17055, which implements the Maxim Model Gauge m5 EZm algorithm, combining a coulomb counter with a voltage-based method.

This section provides a quantitative current consumption analysis for the Pendant and Bracelet, which will later be linked to the architectures discussed along with the work. This analysis was performed by measuring the most energy-demanding actions through the monitoring part described previously. Thus, in the Bracelet, the electric current consumed by acquiring data through each physiological sensor was measured separately. In the Pendant, the electric current consumed by acquiring data using the microphone was measured. Moreover, the power consumption incurred by making use of the buzzer at soft, medium, and strong intensities was also measured for both devices. 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 consumption analysis appear in Figs. 11 and 12 for the Pendant and the Bracelet, respectively. Analyzing the results obtained, the vibration modes in both devices were the most current-consuming actions. The higher the vibration produced, the higher the current required, as expected. However, the buzzer's impact on the autonomy was reduced because it was activated for a short

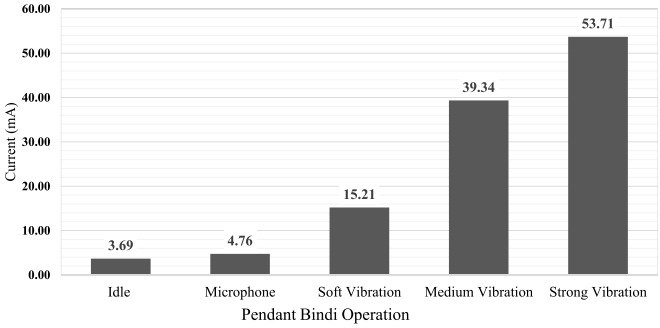

Fig. 11. Average current consumption in the Pendant.

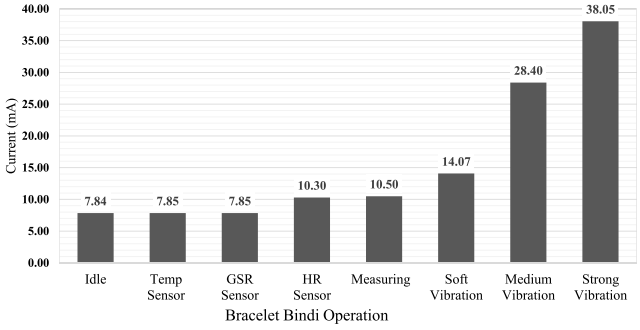

Fig. 12. Average current consumption in the Bracelet.

time in situations identified as GBV, meaning that its activation will usually be sporadic. In the Pendant, the current required by the microphone produced a small increment compared to the idle state. In the Bracelet, the temperature and GSR sensors also produced a small increment from the idle state, even less than the microphone required. However, the HR sensor had a higher impact than the other sensors.

Due to the low power consumption in the idle state, the Pendant's battery life is approximately 38 h using a 140-mAh battery, while the Bracelet needs a 400 mAh battery to last

the same amount of time. Note that these calculations are based on no-alarm situations and are independent of the system architecture. Related to the power consumption comparison between the architectures discussed in this article, the main difference is the existence, or not, of a prealarm. This fact affects the microphone directly since it is working all the time inBindi 2.0b, while in Bindi 1.0 and Bindi 2.0a, it is only activated when a prealarm occurs. Therefore, the Pendant battery life has been reduced to approximately 30 h in theBindi 2.0b architecture. Despite that, the Bracelet is still the most power-demanding device due to the current consumption of its sensors.

VIII. Discussion

Regarding the usual IoT layer architecture (edge, fog, and cloud) considered in Bindi, a relevant system design question concerns which part of the system should be implemented in each of the layers.

First, the cloud computing layer is intended to collect and process great amounts of data without limitations regarding computing resources, energy demand, or response times [89]. This definition fits the needs of the centralized computing services of Bindi, which are therefore placed in the cloud layer to manage potential criminal evidence and historical information for victims' long-term monitoring.

Second, edge computing takes place in the IoT nodes that capture data in the edge of the network. These devices are constrained by their computing and energy capabilities because, in most cases, they are powered by batteries or situated in hazardous environments [90]. This definition fits with the devices by which physiological and auditory data are captured over time in Bindi, i.e., a bracelet and a pendant.

Finally, the fog computing layer follows a concept similar to that of the edge computing layer. However, fog devices are less constrained in computing and energy capabilities while still remaining close to the data origin [91]. According to this description, Bindi's smartphone can be considered a fog device because it does not capture data but is close to the data origin, and both the computing and energy capabilities are less constrained than the ones in the edge devices (the bracelet and the pendant). Some authors assert that the fog does not exist, and then implement the fog layer functionalities described before, inside the edge layer [92]. Under this focus, it is still possible to structure devices in different layers inside the edge. From this point of view, the smartphone would be in an upper layer inside the edge, whereas the bracelet and the pendant would constitute the bottom layer. For further discussion about and review of the edge, fog, and cloud layers, the readers are referred to [9] and [93].

The proposed data fusion techniques in this work achieved a maximum of up to 63.61% average accuracy for a subject-independent fear recognition use case. This result was obtained using multimodal speech and physiological signals and the lowest entropy fusion strategy approach. The obtained average accuracy fell within the range of accuracy rates achieved by similar works presented in Section II and outperformed the system proposed in [64], which considered the same multimodal sources of information. It should be noted that as a differentiating feature of our system, we make use of noninvasive signal monitoring, rather than EEG headsets or face detection sensors [60], [61]. Additionally, the number of users considered (i.e., 42) provides more variability in the data and, therefore, the model more robust.

It is worth highlighting that the configurations described here for fear detection through physiological and speech data and the identification of threatening acoustic events are just possible ways to characterize the situations and contexts in which Bindi users could be involved. These are meant as initial baselines for further developments and have allowed for the identification of important challenges. To start, finding a suitable tradeoff between TPs and TNs and FPs and FNs 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 GBVVs and saturating the protection services with false alarms.

Thus, in this work, we tried to reduce FNs as much as possible, while FPs were maintained at an adequate rate. To this end, we considered strategies based on misclassification costs and threshold parameter setting. Specifically, we fixed th_{phy} in the physiological subsystem to obtain 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 andBindi 2.0b) would help in correcting the bias while trying to maintain the TP prediction. During this experimentation, the current speech monomodal system provided lower performance rates than expected. A possible explanation for this behavior could be the temporal misalignment of the physiological and speech data in WEMAC. The vanishing of the emotion elicited by the time the voice sample is collected could be behind this decrease in performance. Moreover, only classical processing and classification techniques have been used as a baseline for future exploration with this novel data set. A similar situation applies to the fusion strategies, conceived to check the reliability of the prealarms triggered by the physiological model and acting as modulators to lower the FP class prediction rate.

Several problems 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 are acted or synthetic. This situation leads to the need to generate databases with real elicited emotions, which is highly challenging and time consuming. Above all, working with strong negative emotion elicitation, such as that 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 comprises women who have suffered GBV. This is because the failures of the system have critical consequences for them. For this reason, the following data set currently being collected within the UC3M4Safety database comprises only GBVV volunteers. Although the investment of resources to provide

safety and comfortability during the recording of the database is considerable, we are totally committed to the volunteers' well being, providing constant medical assistance as the probability of triggering their posttraumatic stress disorder is very high.

Regarding future work, this study opens the door for further research in many directions. For example, the use of recurrent neural networks to exploit the temporal context of signals, the analysis of other fusion alternatives, or the evaluation of alternative score metrics, such as mutual information or area under the curve, could be used to continue finding a proper balance between false alarms and miss probability. Additionally, adding data acquired from more volunteers in laboratory conditions would add robustness to the models. Likewise, including GBVV data would help to better understand the GBVV activation mechanisms under fear-related situations. Finally, it should be noted that the development of subject-adaptation techniques is critical for our GBV use case.

IX. Conclusion

This article presented Bindi, an end-to-end autonomous multimodal system that leverages affective IoT throughout auditory and physiological commercial off-the-shelf smart sensors, hierarchical multisensorial signal fusion, and secure server architecture, with the final objective of providing safety for and ensuring the well being of GBVVs. Specifically, this article proposed three system architectures for Bindi, consisting of specific arrangements of the data processing subsystems developed, i.e., physiological, speech, and data fusion subsystems, plus a novel acoustic information subsystem to extract acoustic information from the acoustic scene in the near future of Bindi. These architectures were validated and evaluated using the WEMAC data set belonging to the UC3M4Safety database. Note that the data set was specifically built to detect fear in women in a laboratory environment. The reported results achieved an overall fear classification accuracy of 63.61% for a subject-independent approach. The obtained metrics are in line with similar multimodal-based state-of-the-art systems, such as the ones reviewed in Section II [60]–[63]. Moreover, our system outperforms the only system in the literature dealing with the same bimodal combination as in this work [64]. Their results reported an overall accuracy of up to 55.00% for a subject-independent strategy using a feature fusion when targeting a valence and arousal binary classification.

This experimentation serves as an initial multimodal approach toward working with real elicited fear in women and its proper processing. Finally, a power consumption analysis was also presented for the sensors in the Bindi edge wearable devices since its critical application scope. Bindi is a very complex system that requires a thorough balance of many aspects, such as battery consumption, computational power, resource usage, and algorithm performance. We aimed to point out that the ultimate goal of this work is to ignite the community's interest in developing solutions to the very challenging problem of GBV.

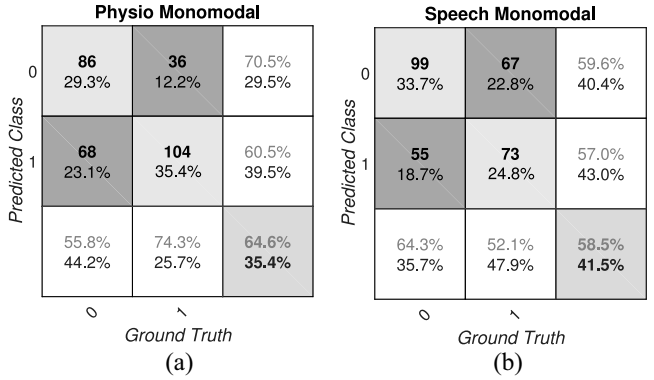

Fig. 13. Confusion matrices for (a) physiological and (b) speech monomodal systems.

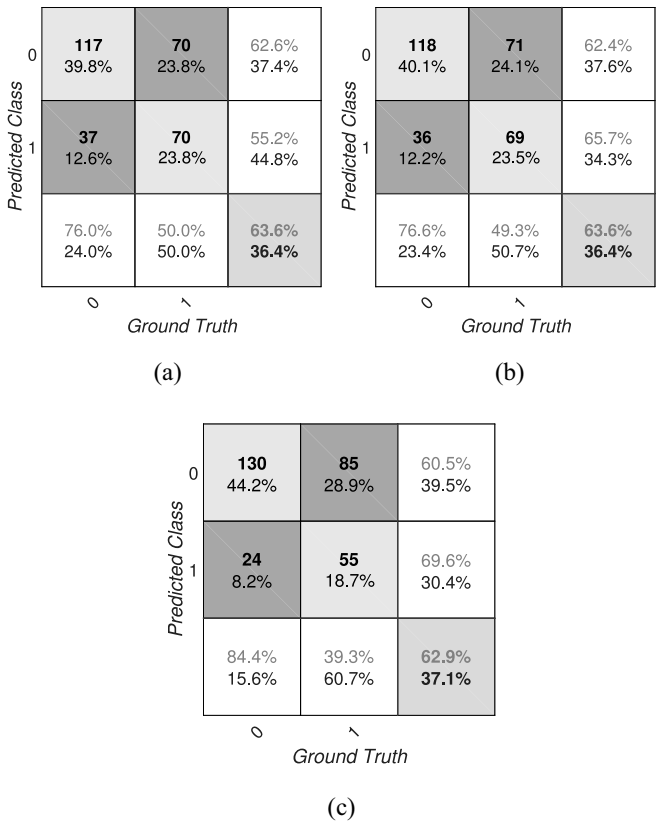

Fig. 14. Data fusion confusion matrices for Bindi 2.0a and Bindi 1.0. (a) Bindi 2.0a. Lowest entropy fusion confusion matrix. (b) Bindi 2.0a. Inverse entropy weighting fusion confusion matrix. (c) Bindi 1.0 confusion matrix.

Appendix
Confusion Matrices for the Monomodal and Fusion Systems

Figs. 13–15 show the confusion matrices for the arrangements evaluated. In these figures, 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 TN, FP, and false omission rates. The next row shows the FN, TP, and precision rates. The last row shows the FN rate, specificity, and overall accuracy.

The physiological subsystem confusion matrix reflects its tendency to predict the positive class at the cost of missing

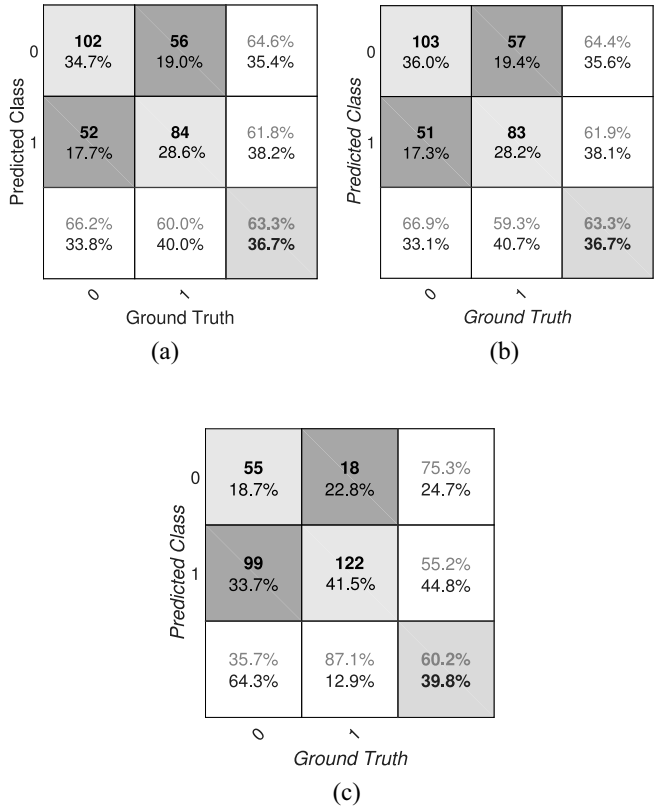

Fig. 15. Data fusion confusion matrices forBindi 2.0b. (a) Lowest entropy fusion confusion matrix. (b) Inverse entropy weighting fusion confusion matrix. (c) Logical OR function fusion confusion matrix.

TNs. Meanwhile, the speech monomodal subsystem achieves lower overall rates than the others but achieves a higher TN rate. Finding a balance between these two behaviors is very important in our application, where missing alerts can be dramatic for the users, but triggering too many false alerts could overwhelm the institutions in charge of protection. Thus, the fact that the speech subsystem can hold back the FPs triggered by the physiological monomodal system looks very promising. In this line of work, the fusion strategies whose confusion matrices are shown in Figs. 14(a) and (b) and 15(a) and (b) differ only in a couple of instances but are more balanced between TNs and TPs. However, the strategy shown in Fig. 15(c) reflects much higher FP and TP rates than the others but misses more TNs than any other, and Fig. 14(c) shows how the hierarchical decision making of Bindi 1.0 performs poorly, proving that fusion is indeed essential.

ACKNOWLEDGMENT

The authors thank the whole UC3M4Safety team for their contribution and support, and specially David Larrabeiti Lopez (Universidad Carlos III of Madrid, Spain, e-mail: dlarra@it.uc3m.es) for his help toward this manuscript.

REFERENCES

[1] (UNICEF, New York, NY, USA). *Gender-Based Violence in Emergencies*. (May 2019). Accessed: Apr. 4, 2021. [Online]. Available: https://www.unicef.org/protection/gender-based-violence-in-emergencies

[2] L. Sardinha, M. Maheu-Giroux, H. Stöckl, S. R. Meyer, and C. García-Moreno, "Global, regional, and national prevalence estimates of physical or sexual, or both, intimate partner violence against women in 2018," *Lancet*, vol. 399, no. 10327, pp. 803–813, 2022. [Online]. Available: https://doi.org/10.1016/S0140-6736(21)02664-7

[3] "Gender-Based Violence Victims Killed in Spain by Their Partners or Former Partners." S. D. Government Against Gender Violence. [Online]. Available: https://violenciagenero.igualdad.gob.es/violenciaEnCifras/victimasMortales (Accessed: Apr. 4, 2021).

[4] V. Collins, *The Costs of Gender-Based Violence in the European Union, European Institute for Gender Equality (EIGE)*, Eur. Inst. Gender Equal., Oct. 2021. Accessed: Mar. 31, 2022.

[5] R. Jewkes and E. Dartnall, "More research is needed on digital technologies in violence against women," *Lancet Public Health*, vol. 4, no. 6, pp. e270–e271, 2019.

[6] "Alertcops Smart-phone Application." Spanish Ministry of the Interior and Public Security. [Online]. Available: https://alertcops.ses.mir.es/mialertcops/en/index.html (Accessed: Apr. 4, 2021).

[7] N. Karusala and N. Kumar, "Women's safety in public spaces: Examining the efficacy of panic buttons in New Delhi," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2017, pp. 3340–3351.

[8] T. Martínez, "A travel of the institutional system in the field of gender violence," *Revista de Estudios Socioeducativos*, no. 7, pp. 256–257, 2019.

[9] J. Portilla, G. Mujica, J.-S. Lee, and T. Riesgo, "The extreme edge at the bottom of the Internet of Things: A review," *IEEE Sensors J.*, vol. 19, no. 9, pp. 3179–3190, May 2019.

[10] T. Mitchell, *Machine Learning*. Cambridge, MA, USA: McGraw Hill, 1997.

[11] J. A. Miranda *et al.*, "WEMAC: Women and emotion multi-modal affective computing dataset," Mar. 2022, *arXiv:2203.00456*.

[12] A. Ferrari, D. Micucci, M. Mobilio, and P. Napoletano, "On the personalization of classification models for human activity recognition," *IEEE Access*, vol. 8, pp. 32066–32079, 2020.

[13] J. Ulla and S. Rosamund, *The Istanbul Convention: A Tool to Tackle Violence Against Women and Girls*, Eur. Parliamentary Res. Service, Brussels, Brussels, 2020.

[14] S. Dunn, *Technology-Facilitated Gender-Based Violence: An Overview, Supporting a Safer Internet Paper No. 1*, Centre Int. Governance Innov., Waterloo, ON, Canada, 2020.

[15] C. Hayes. "STackling Gender-Based Violence with Technology: Case Studies of Mobile and Internet Technology Interventions in Developing Contexts." STATT. Sep. 2014. Accessed: May 31, 2022. [Online]. Available: http://www.gendermatters.co.uk/pdfs/STATT%20Tackling%20GBV%20with%20Technology.pdf

[16] T. W. Bank. "Hackathon Explores Innovative Solutions to Overcome Violence Against Women in Nepal." [Online]. Available: https://www.worldbank.org/en/news/press-release/2013/06/16/violence-against-women-hackathon-nepal (Accessed: Apr. 4, 2021).

[17] Á. González-Prieto, A. Brú, J. C. Nu no, and J. L. González-Álvarez, "Machine learning for risk assessment in gender-based crime," Jun. 2021, *arXiv:2106.11847*.

[18] (Spanish Justice Ministry, Madrid, Spain). *Integrated Protection Measures Against Gender Violence*. [Online]. Available: https://violenciagenero.igualdad.gob.es/definicion/pdf/Ley_integral_ingles.pdf (Accessed: Apr. 15, 2021).

[19] J. J. López-Ossorio *et al.*, "Predictive effectiveness of the police risk assessment in intimate partner violence," *Psychosocial Intervention*, vol. 25, pp. 1–7, Apr. 2016.

[20] "ATENPRO: Servicio Télefónico de Atención y Protección a Las Víctimas de la Violencia de Género." Delegation of the Spanish Government against Gender Violence. [Online]. Available: https://violenciagenero.igualdad.gob.es/informacionUtil/recursos/servicioTecnico/home.htm (Accessed: Apr. 4, 2021).

[21] "Dispositivos de Control Telemático de Medidas y Penas de Alejamiento." Spanish Government Against Gender Violence. [Online]. Available: https://violenciagenero.igualdad.gob.es/informacionUtil/recursos/dispositivosControlTelematico/home.htm (Accessed: Apr. 4, 2021).

[22] J. J. López-Ossorio, J. L. González-Álvarez, and A. Andrés-Pueyo, "Eficacia predictiva de la valoración policial del riesgo de la violencia de género," *Psychosocial Intervention*, vol. 25, no. 1, pp. 1–7, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1132055915000496

[23] L. A. García, "The efficacy of electronic monitoring in gender violence: Criminological analysis," *Int. E-J. Criminal Sci.*, vol. 10, 2016.

[24] R. S. Recio *et al.*, "Prevention of violence against women: Policies and actions on gender violence," *Informació Psicològica*, no. 111, pp. 35–50, 2016.

[25] V. Woollaston, "The XPrize: What has it done for us lately?" *Eng. Technol.*, vol. 15, no. 2, pp. 50–54, 2020.

[26] A. Celik, K. N. Salama, and A. M. Eltawil, "The Internet of Bodies: A systematic survey on propagation characterization and channel modeling," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 321–345, Jan. 2022.

[27] Y. Zhang, Y. Chen, Y. Wang, Q. Liu, and A. Cheng, "CSI-based human activity recognition with graph few-shot learning," *IEEE Internet Things J.*, vol. 9, no. 6, pp. 4139–4151, Mar. 2022.

[28] T. Zhao, Y. Wang, J. Liu, J. Cheng, Y. Chen, and J. Yu, "Robust continuous authentication using cardiac biometrics from wrist-worn wearables," *IEEE Internet Things J.*, early access, Nov. 16, 2021, doi: 10.1109/JIOT.2021.3128290.

[29] T. Zhang, M. Liu, T. Yuan, and N. Al-Nabhan, "Emotion-aware and intelligent Internet of Medical Things toward emotion recognition during COVID-19 pandemic," *IEEE Internet Things J.*, vol. 8, no. 21, pp. 16002–16013, Nov. 2021.

[30] T. Wang, Y. Shen, L. Gao, Y. Jiang, X. Zhu, and F.-C. Zheng, "Long-term energy consumption and transmission delay tradeoff in wireless-powered body area networks," *IEEE Internet Things J.*, vol. 9, no. 6, pp. 4051–4064, Mar. 2022.

[31] A. John, S. J. Redmond, B. Cardiff, and D. John, "A multimodal data fusion technique for heartbeat detection in wearable IoT sensors," *IEEE Internet Things J.*, vol. 9, no. 3, pp. 2071–2082, Feb. 2022.

[32] S. Qiu *et al.*, "Sensor combination selection strategy for kayak cycle phase segmentation based on body sensor networks," *IEEE Internet Things J.*, vol. 9, no. 6, pp. 4190–4201, Mar. 2022.

[33] Y. Bai, L. Chen, M. Abdel-Mottaleb, and J. Xu, "Automated ensemble for deep learning inference on edge computing platforms," *IEEE Internet Things J.*, vol. 9, no. 6, pp. 4202–4213, Mar. 2022.

[34] H. Wang *et al.*, "A comprehensive survey on training acceleration for large machine learning models in IoT," *IEEE Internet Things J.*, vol. 9, no. 2, pp. 939–963, Jan. 2022.

[35] R. W. Picard, "Affective computing for HCI," in *Proc. HCI*, 1999, pp. 829–833.

[36] J. Tao and T. Tan, "Affective computing: A review," in *Affective Computing and Intelligent Interaction*, J. Tao, T. Tan, and R. W. Picard, Eds. Heidelberg, Germany: Springer, 2005, pp. 981–995.

[37] 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?" *Int. J. Environ. Res. Public Health*, vol. 17, no. 22, p. 8534, 2020. [Online]. Available: https://www.mdpi.com/1660-4601/17/22/8534

[38] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," *Biol. Psychol.*, vol. 84, no. 3, pp. 394–421, 2010.

[39] P. Schmidt, A. Reiss, R. Dürichen, and K. V. Laerhoven, "Wearable-based affect recognition—A review," *Sensors*, vol. 19, no. 19, p. 4079, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/19/4079

[40] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: A review," *Int. J. Speech Technol.*, vol. 15, no. 2, pp. 99–117, 2012.

[41] R. Plutchik, "A psychoevolutionary theory of emotions," *Social Sci. Inf.*, vol. 21, nos. 4–5, pp. 529–553, 1982.

[42] D. L. Robinson, "Brain function, emotional experience and personality," *Netherlands J. Psychol.*, vol. 64, no. 4, pp. 152–168, Dec. 2008.

[43] W. Wundt, *Vorselung Über die Menschen—und Tierseele*. Leipzig, Germany: Voss Verlag, 1863, pp. 145–172.

[44] P. Ekman, "Are there basic emotions?" *Psycol. Rev.*, vol. 99, no. 3, pp. 550–553, 1992.

[45] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychol. Sci.*, vol. 18, no. 12, pp. 1050–1057, 2007.

[46] 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,"' *Behav. Cogn. Neurosci. Rev.*, vol. 4, no. 1, pp. 3–20, 2005.

[47] A. F. Ax, "The physiological differentiation between fear and anger in humans," *Psychosomatic Med.*, vol. 15, no. 5, pp. 433–442, 1953.

[48] 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, p. 1738, 2019.

[49] S. Koelstra *et al.*, "Deap: A database for emotion analysis ;using physiological signals," *IEEE Trans. Affective Comput.*, vol. 3, pp. 18–31, Jan.–Mar. 2012.

[50] 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. 1581, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/5/1587

[51] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 42–55, Jan.–Mar. 2012.

[52] B. Zhao, Z. Wang, Z. Yu, and B. Guo, "Emotionsense: Emotion recognition based on wearable wristband," in *Proc. IEEE SmartWorld Ubiquitous Intell. Comput. Adv. Trusted Comput. Scalable Comput. Commun. Cloud Big Data Comput. Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, 2018, pp. 346–355.

[53] 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," *Inf. Fusion*, vol. 51, pp. 10–18, Nov. 2019.

[54] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320310004619

[55] C. Busso *et al.*, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. 6th Int. Conf. Multimodal Interfaces*, 2004, pp. 205–211. [Online]. Available: https://doi.org/10.1145/1027933.1027968

[56] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette, "Fear-type emotion recognition for future audio-based surveillance systems," *Speech Commun.*, vol. 50, no. 6, pp. 487–503, 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016763930800037X

[57] L. Devillers and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs," in *Proc. INTERSPEECH*, 2006, pp. 801–804.

[58] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, Sep. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253517300738

[59] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Inf. Fusion*, vol. 59, pp. 103–126, Jul. 2020.

[60] Y. Cimtay, E. Ekmekcioglu, and S. Caglar-Ozhan, "Cross-subject multimodal emotion recognition based on hybrid fusion," *IEEE Access*, vol. 8, pp. 168865–168878, 2020.

[61] 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.

[62] A. Muaremi, A. Bexheti, F. Gravenhorst, B. Arnrich, and G. Tröster, "Monitoring the impact of stress on the sleep patterns of pilgrims using wearable sensors," in *Proc. IEEE-EMBS Int. Conf. Biomed. Health Inform. (BHI)*, 2014, pp. 185–188.

[63] E. Kanjo, E. M. Younis, and N. Sherkat, "Towards unravelling the relationship between on-body, environmental and emotion data using sensor information fusion approach," *Inf. Fusion*, vol. 40, pp. 18–31, Mar. 2018.

[64] J. Kim and E. Andre, "Emotion recognition using physiological and speech signal in short-term observation," in *Proc. Int. Tutorial Res. Workshop Percept. Interactive Technol. Speech Based Syst.*, 2006, pp. 53–64.

[65] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in *Proc. 20th ACM Int. Conf. Multimodal Interaction*, 2018, pp. 400–408.

[66] J. A. M. Correa, M. K. Abadi, N. Sebe, and I. Patras, "AMIGOS: A dataset for affect, personality and mood research on individuals and groups," *IEEE Trans. Affective Comput.*, vol. 12, no. 2, pp. 479–493, Apr.–Jun. 2021.

[67] A. Baird *et al.*, "An evaluation of speech-based recognition of emotional and physiological markers of stress," *Front. Comput. Sci.*, vol. 3, Dec. 2021, Art. no. 750284. [Online]. Available: https://www.frontiersin.org/article/10.3389/fcomp.2021.750284

[68] A. Baird, S. Amiriparian, M. Berschneider, M. Schmitt, and B. Schuller, "Predicting biological signals from speech: Introducing a novel multimodal dataset and results," in *Proc. IEEE 21st Int. Workshop Multimedia Signal Process. (MMSP)*, 2019, pp. 1–5.

[69] J. A. Miranda, M. F. Canabal, J. M. Lanza-Gutiérrez, M. P. García, and C. López-Ongil, "Toward fear detection using affect recognition," in *Proc. Conf. Design Circuits Integr. Syst. (DCIS)*, 2019, pp. 1–4.

[70] J. A. Miranda, M. F. 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 *Proc. XXXV Conf. Design Circuits Integr. Syst. (DCIS)*, 2020, pp. 1–6.

[71] 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 *Proc. XXXV Conf. Design Circuits Integr. Syst. (DCIS)*, 2020, pp. 1–6.

[72] J. A. Miranda, M. F. Canabal, M. Portela García, and C. Lopez-Ongil, "Embedded emotion recognition: Autonomous multimodal affective Internet of Things," in *Proc. Cyber Phys. Syst. Workshop*, vol. 2208, 2018, pp. 22–29.

[73] 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, Switzerland: Springer Int., 2020, pp. 223–237.

[74] W. Boucsein *et al.*, "Publication recommendations for electrodermal measurements," *Psychophysiology*, vol. 49, no. 8, pp. 1017–1034, 2012.

[75] J. A. M. 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 *Proc. Conf. Design Circuits Integr. Syst. (DCIS)*, 2018, pp. 1–6.

[76] F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Sequential model-based optimization for general algorithm configuration," in *Proc. Int. Conf. Learn. Intell. Optim.*, 2011, pp. 507–523.

[77] N. Marwan, N. Wessel, U. Meyerfeldt, A. Schirdewan, and J. Kurths, "Recurrence-plot-based measures of complexity and their application to heart-rate-variability data," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 66, no. 2, 2002, Art. no. 26702.

[78] D. W. Aha and R. L. Bankert, "A comparative evaluation of sequential feature selection algorithms," in *Learning From Data*. New York, NY, USA: Springer, 1996, pp. 199–206.

[79] J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Commun.*, vol. 42, no. 3, pp. 271–287, 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639303001201

[80] B. McFee, "librosa: Audio and music signal analysis in Python," in *Proc. 14th Python Sci. Conf.*, 2015, pp. 18–24.

[81] E. Rituerto-González, C. Luis-Mingueza, and C. Peález-Moreno, "Using audio events to extend a multi-modal public speaking database with reinterpreted emotional annotations," in *Proc. IberSPEECH*, 2021, pp. 61–65. [Online]. Available: http://dx.doi.org/10.21437/IberSPEECH.2021-13

[82] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[83] J. F. Gemmeke *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2017, pp. 776–780.

[84] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2017, pp. 131–135.

[85] M. Á. B. Ruiz *et al.* "UC3M4Safety Database Description." 2021. [Online]. Available: http://hdl.handle.net/10016/32481

[86] M. Á. B. Ruiz *et al.* "UC3M4Safety Database—List of Audiovisual Stimuli." 2021. [Online]. Available: https://doi.org/10.21950/CXAAHR

[87] M. Á. B. Ruiz *et al.* "UC3M4Safety Database—List of Audiovisual Stimuli (Video)." 2021. [Online]. Available: https://doi.org/10.21950/LUO1IZ

[88] F. J. Valverde-Albacete and C. Peláez-Moreno, "10% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox," *PLOS ONE*, vol. 9, no. 1, pp. 1–10, 2014. [Online]. Available: https://doi.org/10.1371/journal.pone.0084217

[89] P. Mell and T. Grance, "The NIST definition of cloud computing," Inf. Technol. Lab., Nat. Inst. Stand. Technol., Gaithersburg, MD, USA, Rep. NIST SP 800-145, 2011.

[90] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 637–646, Oct. 2016.

[91] M. Iorga, L. Feldman, R. Barton, M. J. Martin, N. S. Goren, and C. Mahmoudi, "Fog computing conceptual model," Inf. Technol. Lab., Nat. Inst. Stand. Technol., Gaithersburg, MD, USA, Rep. 500-325, 2018.

[92] G. Premsankar, M. D. Francesco, and T. Taleb, "Edge computing for the Internet of Things: A case study," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 1275–1284, Apr. 2018.

[93] F. Firouzi, B. Farahani, and A. Marinšek, "The convergence and interplay of edge, fog, and cloud in the AI-driven Internet of Things (IoT)," *Inf. Syst.*, vol. 107, Jul. 2022, Art. no. 101840.

Jose A. Miranda Calero received the B.Sc. degree in industrial electronics and automation engineering and the M.Sc. degree (Hons.) in electronic systems and applications engineering from the Universidad Carlos III of Madrid, Leganés, Spain, in 2012 and 2015, respectively, where he is currently pursuing the Ph.D. degree with the Microelectronic Design and Applications Research Group.

From 2012 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 gender violence 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.

Esther Rituerto-González received the B.Eng. degree in audio-visual systems and the M.Eng. degree in multimedia communications with specialization in signal and data processing from the Universidad Carlos III of Madrid, Leganés, Spain, in 2017 and 2018, respectively, where she is currently pursuing the Ph.D. degree in speech technologies with the Group of Multimedia Processing.

She was a Visiting Researcher with Augsburg Universität, Augsburg, Germany. She has been associated with the University Signal Theory and Communications Department, Universidad Carlos III of Madrid since 2016. She is currently working in the EMPATIA-CM Project with the UC3M4Safety Group, where they are developing a wearable solution that will detect a user's panic, fear, and stress through physiological sensor data, speech and audio analysis, machine-learning algorithms, and multimodal data fusion, for gender-violence detection and prevention. Her interests rely in very different research communities focusing on computer science categories, such as affective computing, speech communications, speaker recognition, emotions in speech, biosignals processing, artificial intelligence, and deep learning.

Clara Luis-Mingueza received the sound and image engineering degree from the Universidad Carlos III of Madrid, Leganés, Spain, in 2020, where she is currently pursuing the M.Sc. degree in information health engineering, where she has studied topics, such as computer vision, medical image, and speech processing.

In her bachelor thesis, she explored music generation through artificial neural networks. She is currently a Research Assistant at EMPATIA: a multidisciplinary research project focused on comprehensive protection for victims of gender-based violence through multimodal affective computing and wearable solutions. As a part of UC3M4Safety Group, she is researching sound event detection in stress, panic, and fear situations to detect violent scenarios with deep learning given an acoustic scene.

Manuel F. Canabal received the B.Eng. degree in industrial technologies and the M.Sc. degree in electronic systems and applications engineering from the Universidad Carlos III of Madrid, Leganés, Spain, in 2015 and 2017, respectively, where he is currently pursuing the Ph.D. degree in industrial electronics program.

Since 2015, he has been working in different research fields, such as mixed-signal microelectronic, space instruments, affective computing, and the integration of this technology in edge devices. He is part of the UC3M4Safety team that aims to provide a technological solution to combat gender violence from a multidisciplinary perspective. Bindi is the wearable system that he has developed with the rest of the team to achieve this goal integrating Universidad Carlos III of Madrid and physiological smart sensors in a constraint platform along with machine learning algorithms. Moreover, he is researching the use of new biosensors to improve Bindi capabilities and functionality.

Alberto Ramírez Bárcenas received the B.Eng. degree in industrial electronics and automation engineering and the M.Sc. degree in electronic systems and applications engineering from the Universidad Carlos III of Madrid, Leganés, Spain, in 2018 and 2019, respectively, where he is currently pursuing the Ph.D. degree in industrial electronics program.

His research field comprises fault-tolerant design, online testing, and hardware/software co-design. He is part of UC3M4Safety team, participating in the development 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 gender violence situation employing emotion recognition using the physiological and auditory signals of the user. Moreover, he participates in the space exploration project DS-EXOMARS20 with the Universidad Carlos III of Madrid.

Jose M. Lanza-Gutiérrez received the B.S and M.S degrees in computer science, the master's degree in grid computing and parallelism, and the Ph.D. degree in computer science (under the guidance of Prof. Dr. Juan A. Gomez-Pulido) from the University of Extremadura, Caceres, Spain, in 2008, 2009, 2010, and 2015 respectively.

He is currently an Assistant Professor with the University of Alcalá, Alcalá de Henares, Spain. He has authored or coauthored more than 50 publications, including Journal Citation Report papers in journals, such as *Applied Soft Computing*, *Expert Systems with Application*, *BMC Bioinformatics*, *Soft Computing*, *Reliability Engineering and System Safety*, IEEE ACCESS, and IEEE INTERNET OF THINGS. His main research interests include metaheuristics, Internet of Things, digital embedded systems, and machine learning.

Carmen Peláez-Moreno (Member, IEEE) received the engineering degree in telecommunications from the Public University of Navarre, Pamplona, Spain, in 1997, and the Ph.D. degree from the Universidad Carlos III of Madrid, Leganés, Spain, in 2002.

She was a Visiting Researcher with the University of Westminster, London, U.K., in 1996; the University of Strathclyde, Glasgow, U.K., in 2003; the University of Trento, Trento, Italy, in 2013; and the International Computer Science Institute, Berkeley, CA, USA, in 2004 and 2006. She is currently an Associate Professor with the Signal Theory and Communications Department, Universidad Carlos III of Madrid, a Collaborator of the Spanish National Research Agency, and has earned the Full Professorship National Habilitation. Her research interests include speech recognition and perception, affective computing, machine learning and data analysis, and information theory and education.

Celia López-Ongil (Senior Member, IEEE) received the industrial engineering degree from the Polytechnic University of Madrid, Madrid, Spain, in 1995, and the Ph.D. degree in industrial engineering from Polytechnic University of Madrid in 2000.

In 1998, she joined the Universidad Carlos III of Madrid, Leganés, Spain, as an Assistant Professor, where she has been an Associate Professor with the Department of Electronic Technology since 2010. Her research was framed in the line of robust circuits for space applications in the first 20 years, focused on the generation of HW/SW tools to ensure the quality and reliability of microelectronic circuits, and produced two Ph.D. theses, 19 articles in scientific journals, and 90 contributions at international conferences. Some designs in space missions (ASCAT-METOP, SADE-ROSETTA, and DS-EXOMARS20) can be highlighted. Since 2016, her research has been focused on cyber–physical systems, dedicated to detecting, preventing, and fighting violence against women. She is the Leader of the UC3M4Safety multidisciplinary team that proposes to use technology to protect women at risk of sexual or gender-based violence.

Dr. López-Ongil's team has been a semifinalist in the International XPRIZE Women's Safety Competition and has won the Vodafone Foundation Award "Connecting for Good" in 2019.