Research on Biomedical Engineering (2023) 39:613–638
https://doi.org/10.1007/s42600-023-00293-9

ORIGINAL ARTICLE

Emotion Recognition from Multimodal Data: a machine learning approach combining classical and hybrid deep architectures

Maíra Araújo de Santana[1] · Flávio Secco Fonseca[1] · Arianne Sarmento Torcate[1] · Wellington Pinheiro dos Santos[1,2]

Received: 15 March 2023 / Accepted: 15 July 2023 / Published online: 2 August 2023

Abstract

Purpose The expression of emotions is essential in human relationships. However, the aging process associated with some pathologies such as Alzheimer's Disease and other dementias can affect our ability to express emotions.
Methods In this context, we propose a method for automatic recognition of emotions from multimodal data. We based this approach on Artificial Intelligence algorithms, as part of the development of a human–machine interface to support the personalization of therapy for elderly people with dementia. From this tool, emotional feedback can modulate the therapy. By doing this we hope to improve the therapeutic results. In this work, the performance of the proposed architectures was evaluated regarding to their ability to recognize emotions in physiological and speech signals and in images of facial expressions.
Results In the context of physiological and speech signals, we achieved promising results with the use of Random Forest. We found an accuracy of up to 99% in classifying emotions from physiological signals and almost 80% with speech signals. In the images assessment, we found more than 82% of accuracy when adopting a hybrid architecture.
Conclusion The good results in the test stage are encouraging and point to the possibility of adopting the method in the analysis of emotions in multimodal data. These findings are even more interesting due to the large amount and variety of emotions.

Keywords Emotion Recognition · Multimodal Data · Therapy · Dementia · Elderly · Biofeedback

Introduction

Motivation and problem characterization

Emotions are present on many situations in our daily lives. They shape our choices, desires, tastes, memories, and other human aspects. Throughout human history, the range of emotions that can be felt and expressed has always been a topic that has attracted the attention of behavioral scientists. Even in the 1900s, some studies were conducted in order to find patterns to map the different human emotions (Russell 1980; Izard 1977, 1991).

Nowadays we know that emotions are distinguishable from each other and are built from the subjective experiences of each individual. Furthermore, these feelings can be interpreted as involuntary physiological responses. However, emotions are not isolated and easily identified variables, since they are manifested from combined elements such as sensations, changes in voice and facial expressions (Oliveira and Jaques 2013). The use of Artificial Intelligence (AI) techniques has contributed to this field of study (Santana et al. 2021; Saxena et al. 2020). AI algorithms are already successfully applied in the analysis of many complex and non-linear problems of everyday life (Gupta et al. 2018; Andrade et al. 2020; Oliveira et al. 2020; Santana et al. 2020a, 2020b, 2018; Cruz et al. 2018; Barbosa et al. 2020; Silva and Santana 2020; Gomes et al. 2020; Freitas Barbosa et al. 2021). Furthermore, this tool is commonly successful in analyzing large volumes of data (Deshpande and Kumar 2018).

Facial expression analysis currently is the most common way to perform automatic emotion recognition (Santana et al. 2021). Despite being a well-explored field of study, there are still several gaps associated with this task. The development of solutions in this context requires a large amount of data, due to the huge human variability, especially regarding demographic aspects (Lawrence et al. 2015;

✉ Wellington Pinheiro dos Santos
wellington.santos@ufpe.br

[1] Núcleo de Engenharia da Computação, Escola Politécnica da Universidade de Pernambuco, Recife, PE, Brazil

[2] Departamento de Engenharia Biomédica, Universidade Federal de Pernambuco, Recife, PE, Brazil

Reyes et al. 2018). This factor directly interferes with the facial expressions and, therefore, needs to be considered. The need for many data leads to an increased computational costs associated with facial expression data analysis. In addition, pathologies, injuries and human aging itself commonly affect the face and the individual's ability to express emotions (Lawrence et al. 2015; Harms et al. 2010; Kohler et al. 2003). These challenges can be overcome by associating facial expression data with data from other sources (Silva et al. 1997; Abdullah et al. 2021). Therefore, the face is not the only source of information for decision making regarding the classification of emotions.

From a neurological point of view, human emotions activate a series of affectivecognitive brain structures. We can assess the neuronal activity generated by emotions from an electroencephalogram (EEG) (Izard 1977, 1991). EEG is one of the main techniques for acquiring human neurophysiological activity. Mainly due to its reliability, effectiveness, simplicity, portability and accessibility (Gupta et al. 2018; Andrade et al. 2020; Oliveira et al. 2020; Santana et al. 2020a; Alarcao and Fonseca 2017). In addition to neurophysiological activations, emotions have an effect on peripheral physiological signals. We noticed common changes through galvanic skin response (GSR), heart rate, temperature, and respiratory rate. The association of these peripheral and central physiological signals favors the recognition of emotions (Santana et al. 2020a; Doma and Pirouz 2020; Vijayakumar et al. 2020; Khalili and Moradi 2009; Shu and Wang 2017).

Emotions can also be perceived and differentiated from patterns of human voice recordings. Changes in the time and frequency domains of these signals often appear during the expression of different emotions. Several studies have been dedicated to the recognition of emotions in speech, especially with the aim of incorporating this analysis into human–computer interfaces (Santana et al. 2021; Schuller et al. 2003; Livingstone and Russo 2018; Issa et al. 2020). However, developing models that understand the nuances in natural language and speech is still a complex task. Therefore, there is a tendency to combine this analysis with other types of data related to the manifestation of emotions (Santana et al. 2021).

One of the main factors that make it difficult to recognize emotions is the existence of some pathology. Neurodegenerative pathologies such as Alzheimer's disease and other dementias commonly lead to neurological impairments that affect both the identification of emotions and their expression (García-Casal et al. 2017; Behere et al. 2011; McIntosh et al. 2015). In addition, with the current and growing process of population aging around the world, we are also experiencing an increase in cases of this type of pathologies (Mundial and da Saúde 2018; Saúde 2021). According to Ferreira and Torro-Alves (2016), emotions are fundamental in the regulation of social interactions, as they guide our preferences, motivations and decision making (Ferreira and Torro-Alves 2016). They are also indispensable to provide good verbal and non-verbal communication (Chaturvedi et al. 2021; Dorneles et al. 2020). Thus, it is essential to develop tools that help in the identification of emotions for a dignified and pleasant quality of life.

In the therapeutic context, automatic emotion recognition tools are important to improve interventions in the most diverse audiences. Some studies demonstrate that emotional response can be used to improve patient engagement in the therapy process (Marinoiu et al. 2018; Schipor et al. 2011; Sourina et al. 2012; Delmastro et al. 2018; Aranha et al. 2017; Arroyo-Palacios and Slater 2016). It is important to highlight that greater engagement tends to increase the effectiveness of these therapeutic interventions (Lenze et al. 2011).

Therefore, this study proposes a method for recognizing emotions from multimodal data. This method will be incorporated as the core of a human–machine interface to support the therapy of elderly people with dementia. The aim is to contribute to the personalization and consequent optimization of the therapeutic process. We base the proposed method on artificial intelligence algorithms to deal with data from physiological parameters, facial expressions and speech signals.

We organize the article as follows. In the next section we present some recent and relevant studies of emotion recoginition from these type of data. After this section, we detail our approach in the Materials and Methods topic, followed by the results and discussion sections. Finally, we draw some conclusions, highlighting the main findings and future possibilities.

Related Works

Nowadays, automatic recognition of emotions has strong relevance in the therapeutic scenario. The emotional response of patients has already been used to shape the therapeutic experience, so that interventions become more appropriate to achieve the particular goals of each individual. Different therapy modalities can benefit from emotion recognition tools. Some studies are already carried out in the context of physical therapy for motor rehabilitation (Aranha et al. 2017), in speech therapy (Schipor et al. 2011), in music therapy (Sourina et al. 2012), in addition to cognitive behavioral therapies (Marinoiu et al. 2018; Arroyo-Palacios and Slater 2016).

Aranha et al. (2017) proposed a serious game adaptation approach for motor rehabilitation. From the implemented framework, physical therapists can use the affective response of patients to adapt the commands of a game. The recognition of emotions was performed from the analysis of the user's facial expression. With this, it was possible to achieve the goals of rehabilitation more effectively. An increase in the effectiveness of the therapeutic process was also identified by (Schipor et al. 2011), but now in the context of speech therapy. Since speech quality is also influenced by the individual's emotional condition, the authors implemented an emotion recognition module in a Computer Based Speech Therapy System (CBST) to assess the quality of word pronunciation in the context of speech therapy in children. The results were promising and point to the close relationship between the human voice and emotional states.

Music therapy sessions can also be favored by assessing the affective state of the patient. (Sourina et al. 2012) built a tool to identify the emotional state of the user in real time and use it to adjust the songs used during music therapy. This approach classifies emotions into fear, frustrated, sad, happy, pleasant, and satisfied from EEG signals. Thus, using perceived emotion, the system automatically selects the most appropriate music to meet the patient's needs.

Marinoiu et al. 2018) investigated the expression of emotions in the context of robot-assisted therapy of children with Autism Spectrum Disorder (ASD). The authors performed emotion recognition in 3D videos collected with a Kinect system. After analyzing the data, they realized that emotional state identification has great potential to improve human–machine interaction and, consequently, improve therapeutic intervention in these individuals.

In order to modulate the cognitive-behavioral state of the participants, (Arroyo-Palacios and Slater 2016) proposed a virtual reality scenario to identify and modulate the affective state of the user. In the proposed interface, participants were represented by virtual dancers and had to control the rhythm of the dance by modulating their own mood. Thus, people who were agitated should make the avatar move more calmly. In the opposite way, people who were more relaxed should make the character dance more frantically. Participants' mood were identified from the physiological signals of skin conductance, heart rate and respiratory rate. Only by modulating these parameters it was possible to control the avatar's activity. The authors concluded that by using this game, participants were able to emotionally awaken in when in the activation condition and relax in the relaxation condition.

In order to contribute to the development of studies related to emotional response, (Soleymani et al. 2011) gave rise to the MAHNOB-HCI database. The base was built to acquire information about different manifestations of affective responses to audiovisual stimuli. This acquisition was performed from multimodal data that, among other information, includes records of physiological signals from the central nervous system (EEG) and peripheral nervous system (electrocardiogram (ECG), GSR, respiratory amplitude and skin temperature). Given the vast amount of information, this database has been used in several studies for emotion recognition. The authors of the database themselves conducted a promising preliminary study of automatic recognition of emotions from this data. For this analysis, the authors extracted spectral and statistical features from the physiological signals. In total, 318 attributes were extracted, 20 from the GSR signal, 64 from the ECG, 14 from the breathing pattern, 4 from the skin temperature and 216 from the EEG signal. After feature extraction, the authors evaluated the performance of a Support Vector Machine (SVM) model with a radial basis function (RBF) kernel to classify data in terms of valence and arousal. The proposed model was able to classify peripheral physiological signals with an accuracy of 46.2% for the arousal class and 45% for valence. In the classification of EEG signals, the method obtained slightly better accuracies, 52.4% for arousal and 57% for valence.

A few years after the development of this database, (Wiem and Lachiri 2017) used peripheral physiological signals from the MAHNOB-HCI database to propose a method for classifying emotions. Initially, the authors removed artifacts and noise using Butterworth filters. Then, 169 statistical attributes were extracted from each signal. Finally, the authors evaluated the performance of 4 SVM configurations for signal classification. SVM algorithms with linear, polynomial, sigmoid and gaussian kernels were evaluated. These different configurations showed similar results to each other. The use of ECG signals resulted in the best classification performances, with accuracies around 65% for arousal and 60% for valence using SVM with linear kernel. This same SVM configuration reached accuracies between 53 and 63% in the classification of affective states using the other peripheral physiological signals. However, when the physiological signals were combined, SVM algorithm with polynomial kernel showed a better performance in classifying arousal levels, with an accuracy of 64.23%. In the case of valence, the SVM with Gaussian kernel presented the best performance, with an accuracy of 68.75%.

Another relevant work that uses the physiological signals of the MAHNOB-HCI base is that of (Wei et al. 2018). In this study, the authors sought to perform emotion recognition from EEG, ECG, respiration amplitude, and GSR signals. For the feature extraction, the authors used a combination of attributes from the time and frequency domains. After extracting attributes, the authors submitted the data for classification with an SVM algorithm with RBF kernel. The hyperparameters of this algorithm (C and γ) were optimized by the grid search method. The authors evaluated the algorithm's performance in rating 5 emotions: Sadness, Happiness, Disgust, Neutral, and Fear. The classification was made separately for each of the physiological signals. Thus, an accuracy of 74.52% was obtained using EEG signals, 68.75% with ECG signals, 54.33% using respiration amplitude, and 57.69% with GSR signals. Subsequently, the authors also evaluated the performance from the combination of the four physiological data, reaching an accuracy of up to 84.62%.

Still in the efforts to find of strategies to perform automatic emotion recognition, Livingstone and Russo (2018) proposed the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone and Russo 2018). The database has voice and video recordings of professional actors expressing 8 emotions: calm, happy, sad, angry, fearful, surprise, disgust, and neutral. In this study, the authors also validated the database based on the analysis of 72 human evaluators. This evaluation showed that emotions are better identified using audio associated with video or simply video than using audio alone. Overall, the authors reported accuracies between 58 and 67% in classifying emotions through speech. These results are associated with Kappa between 0.41 and 0.52. The "neutral" and "angry" states were the most easily identified. The highest amount of misclassification was associated with "sad" emotion.

In 2020, (Issa et al. 2020) proposed a method for classifying emotions from voice signals. Part of the method evaluation was done using the RAVDESS database. The proposed architecture consists of extracting honeyfrequency cepstral coefficients, chroma-gram, honey-scale spectrogram, Tonnetz representation, and spectral contrast features from sound files. After the feature extraction, the data were classified by a convolutional neural network (CNN) with a rectified linear activation function (ReLU). This model correctly classified 71.61% of the data from the RAVDESS database. Better rating performances were obtained for stronger emotions like "angry". There was greater confusion in the classification of emotions closer to each other such as "calm" and "sad" or "happy" and "surprised".

The following year, (Luna-Jiménez et al. 2021) proposed the use of a CNN architecture for emotion recognition with the RAVDESS database. Pre-trained CNN architectures with AlexNet were used for feature extraction. The authors obtained better results with RBF kernel SVM as the classifier. The best evaluated model resulted in an accuracy of 76.58% in the identification of the 8 emotions in the database. Emotions "angry" and "disgust" were ranked higher. Higher error rates were associated with the "sad" class, commonly confused with "calm" and "fearful".

The FER facial expressions database was developed by (Goodfellow et al. 2013). This database has 35,887 images, all resized to 48 × 48 pixels and converted to grayscale, covering 7 types of emotions: Anger (4.593), Disgust (547), Fear (5.121), Happy (8.989), Neutral (6.198), Sad (6.077) and Surprise (4.002). The authors used CNN and SVM to extract attributes and classify images of facial expressions. The method proposed by them reached 71.2% accuracy in the test set.

The FER database was also used by (Ng et al. 2015). In their approach, the database was used to train a CNN model. Initially, the images were cropped and adjusted for better visualization of the facial expression. Then, they were used to refine the training of a CNN, given the size and diversity of the dataset. Finally, the trained architecture was used to classify the images from the EmotiW database. The accuracies found by them were median, assuming values between 42 and 56%. It is important to mention that the authors did not report the execution of class balancing steps. Since the base is originally unbalanced, the lack of balance may have negatively affected the results. This balance makes CNN learn the patterns of some classes better than others, skewing the result. Five years later, (Kusuma et al. 2020) conducted an emotion recognition study using a pre-trained VGG-16 model. ImageNet image dataset was used to train the model. Then, the authors used the model to classify the images from the FER-2013 database. Finally, their method was able to differentiate 7 distinct emotions with an accuracy of 69.40%.

In Table 1 we present the main information from these related studies, such as their main goal, the computation techniques used and their main findings. At the last line of this table is our method, proving that our proposal is well contextualized in the state-of-the-art.

Material and Methods

Theoretical background

Affective Computing, a burgeoning field at the intersection of computer science and psychology, is concerned with the development of computational systems endowed with the ability to perceive, interpret, and respond to human emotions (Bota et al. 2019; Cambria et al. 2017). It encompasses a diverse array of methodologies, including signal processing, machine learning, and human–computer interaction, which collectively enable machines to discern and appropriately react to human affective states (Picard 2000; Hasnul et al. 2021; Calvo and D'Mello 2010; Kołakowska et al. 2020). Of particular significance is the domain of healthcare, where emotion recognition assumes a crucial role. The recognition and comprehension of emotions in healthcare settings have profound implications for patient care, facilitating the early detection of mental health disorders and fostering tailored interventions (Picard 2000; Kołakowska et al. 2020; Sinha 2021). For instance, in the realm of mental health, emotion recognition systems can be employed to monitor the emotional well-being of individuals with conditions such as depression or anxiety, enabling timely intervention and personalized treatment strategies (Picard 2000; Kołakowska et al. 2020; Sinha 2021). Furthermore, in domains like telemedicine and remote patient monitoring,

Table 1 Summary of related works

Work	Main goal	Method	Main results
(Soleymani et al. 2011)	(1) The development of a multimodal database for emotion recognition. (2) To recognize emotion in physiological signals	In addition to the experimental protocol for acquiring the database, the authors conducted a binary classification study (arousal versus valence). For this classification they used an SVM model with RBF kernel. Signals were represented by spectral and statistical features	The authors found an accuracy around 45% in the classification of peripheral physiological signals. In the analysis of EEG, they obtained accuracy close to 55%
(Goodfellow et al. 2013)	To recognize emotions in facial expressions	The authors used a CNN model combined to SVM to classify 7 types of emotions	Their method achieved a maximum accuracy of 71.2%
(Ng et al. 2015)	To identify emotions from images of facial expressions	The authors conducted a study to assess the facial expressions using a CNN architecture	They found an correct classified rate of up to 56%
(Wiem and Lachiri 2017)	To perform emotion recognition from peripheral physiological signals	The authors extracted statistical features from the signals. Then, they assessed a binary (arousal versus valence) classification performance using SVM with linear, polynomial, sigmoid, and gaussian kernels	SVM performance with the different kernels were similar. For arousal classification the authors achieved 64.23% of accuracy using the polynomial kernel. The gaussian kernel performed better for valence classification with an accuracy of 68.75%
(Wei et al. 2018)	To perform emotion recognition from peripheral and central physiological signals (EEG, ECG, GSR, and respiration amplitude)	The authors extracted features from time and frequency domains. They used SVM with RBF kernel to perform classification into 5 emotion classes (i.e. sadness, happiness, disgust, neutral, and fear)	The best performance was achieved combining the signal modalities, with accuracy up to 84.62%. By using the signals individually the authors found the best accuracy (74%) with EEG signals
(Livingstone and Russo 2018)	The development of a validated database for emotion recognition from speech and video data	The authors recorded data from 24 professional actors expressing 8 emotional states (i.e. calm, happy, sad, angry, fearful, surprise, disgust, and neutral). Data was validated from human specialists	When using only speech data they reported a maximum accuracy of 67% with Kappa of 0.52. Angry and neutral states were easily classified
(Issa et al. 2020)	To perform emotion recognition from speech signals	The authors used spectral features to describe 8 emotions in speech data from RAVDESS database. A CNN model was used to classify these emotions	They found an overall accuracy of 71.61% using the proposed architecture. Strong emotions such as angry were better classified. There was more confusion between close emotions
(Kusuma et al. 2020)	To perform emotion recognition from facial expressions	The authors used a pre-trained VGG-16 architecture to classify the images of FER-2013 according to their respective emotion tags	The authors found an accuracy of 69.40% to classify these emotions
(Luna-Jiménez et al. 2021)	To perform emotion recognition from speech signals	The authors used CNN pre-trained models to describe 8 emotions in speech data from RAVDESS database. An SVM architecture with RBF kernel was used to classify these emotions	Their method achieved maximum accuracy of 76.58%. Strong emotions such as angry were better classified
Our approach	To perform emotion recognition from multimodal data	We used EEG and other physiological data from MAHNOB-HCI database; speech data from RAVDESS database; and facial expressions from FER-2013. From the signals we extracted explicit features from statistical, and time–frequency domains. We used a transfer learning architecture to classify the images from facial expressions	The proposed method achieved 99% of accuracy for physiological data, an accuracy of 80% for speech, and of 83% for facial expressions

emotion recognition technologies hold the potential to assess patients' emotional states during virtual consultations, thus providing a comprehensive understanding of their needs and experiences (Picard 2000; Hasnul et al. 2021; Kołakowska et al. 2020; Sinha 2021).

The paramount importance of emotion recognition in healthcare extends beyond patient care to the well-being of healthcare practitioners (Pujol et al. 2019). Professionals in highstress occupations, including doctors, nurses, and caregivers, confront significant emotional challenges that can profoundly impact their mental health and job performance (Pujol et al. 2019). Here, emotion recognition systems offer an avenue for assessing the emotional states of healthcare workers, enabling timely intervention and support. For example, wearable devices equipped with sensors can continuously monitor physiological signals, such as heart rate variability and skin conductance, thereby providing valuable insights into emotional states such as stress or burnout (Hasnul et al. 2021; Kołakowska et al. 2020; Saganowski et al. 2020; Marcos et al. 2021; Ayata et al. 2020; Dhuheir et al. 2021). Such information can be utilized to deliver real-time feedback and interventions to healthcare professionals, thereby ensuring their well-being and overall job satisfaction (Hasnul et al. 2021; Kołakowska et al. 2020; Saganowski et al. 2020; Marcos et al. 2021; Ayata et al. 2020; Dhuheir et al. 2021). Moreover, the integration of emotion recognition technologies has the potential to foster the development of intelligent systems that respond empathetically to the emotional needs of healthcare providers, thereby cultivating a supportive work environment and augmenting their overall work experience (Hasnul et al. 2021; Kołakowska et al. 2020; Saganowski et al. 2020; Marcos et al. 2021; Ayata et al. 2020; Dhuheir et al. 2021).

Alzheimer's disease and other forms of dementia impose a significant burden on individuals worldwide (Cobos and Rodríguez, M.d.M.M. 2012; Olanrewaju et al. 2015; Castro et al. 2021; Livingston et al. 2017; Shafqat 2008). The prevalence of these conditions has reached alarming levels, with an estimated 50 million people currently affected globally (Zhang et al. 2021; Barnes and Yaffe 2011). Alzheimer's disease, in particular, is the most common form of dementia, characterized by progressive cognitive decline and memory loss (Cobos and Rodríguez 2012; Olanrewaju et al. 2015; Castro et al. 2021; Livingston et al. 2017; Shafqat 2008). Patients with Alzheimer's experience a range of symptoms, including impaired thinking, disorientation, language difficulties, and changes in behavior and mood. These debilitating effects severely impact the quality of life of patients and their families, leading to a decline in functional abilities, loss of independence, and diminished social engagement (Cobos and Rodríguez 2012; Olanrewaju et al. 2015; Castro et al. 2021; Livingston et al. 2017; Shafqat 2008). Moreover, the global impact of Alzheimer's and dementia extends beyond the individual level, placing an immense strain on healthcare systems, economies, and society as a whole (Cobos and Rodríguez 2012; Olanrewaju et al. 2015; Castro et al. 2021; Livingston et al. 2017; Shafqat 2008). Several works have faced the problem of providing in vivo diagnosis for Alzheimer's disease and other dementias by using image diagnosis optimized by machine learning and evolutionary computing (Souza et al. 2021; Santos et al. 2008a, 2009a, 2008a, 2009b, 2008b; Silva et al. 2019). Given the profound consequences of these diseases, there is an urgent need for interventions that can alleviate symptoms and enhance the well-being of patients (Sörensen et al. 2006; Haan and Wallace 2004).

Music therapy and other art-based interventions have shown promising potential in assisting individuals with Alzheimer's disease, particularly in the early stages of the illness. Music, with its ability to evoke emotional responses and retrieve memories, holds a unique position in therapeutic approaches for dementia patients. Engaging in music therapy sessions can stimulate various cognitive functions, including memory recall and emotional processing, leading to the emergence of positive affective memories (Matziorinis and Koelsch 2022; Brotons and Marti 2003; Guess 2017; Steen et al. 2018; Leggieri et al. 2019). The power of positive affective memories evoked through music therapy is significant, as it can enhance the overall quality of life for Alzheimer's and dementia patients. These memories may evoke emotions, trigger reminiscence, and foster connections with personal experiences, promoting a sense of identity and emotional well-being (Steen et al. 2018; Blackburn and Bradshaw 2014; Guetin et al. 2013). By leveraging the emotional and memory-related benefits of music therapy, individuals with Alzheimer's can experience improved mood, reduced agitation, enhanced communication, and increased social interaction. Importantly, these effects can extend beyond the duration of therapy sessions, creating a positive impact on daily life and social interactions for patients and their caregivers (Steen et al. 2018; Blackburn and Bradshaw 2014; Guetin et al. 2013).

Furthermore, the integration of emotion recognition tools in music therapy holds great potential for improving its efficacy (Kim and André 2008). With a multimodal approach, several different datasets, provided by different information, can be used to improve emotion recognition by machine learning. It is possible to combine video, audio, face, and physiological signals to improve emotion recognition accuracy, generating a precise way to estimate emotions in a biofeedback-based approach (Santana et al. 2021; Muyuan et al. 2004). By providing biofeedback to therapists, these tools can assist in assessing the emotional responses and levels of engagement of patients during music therapy sessions. This feedback enables therapists to fine-tune their interventions, tailoring the therapeutic approach to optimize the emergence of positive affective memories. With real-time information

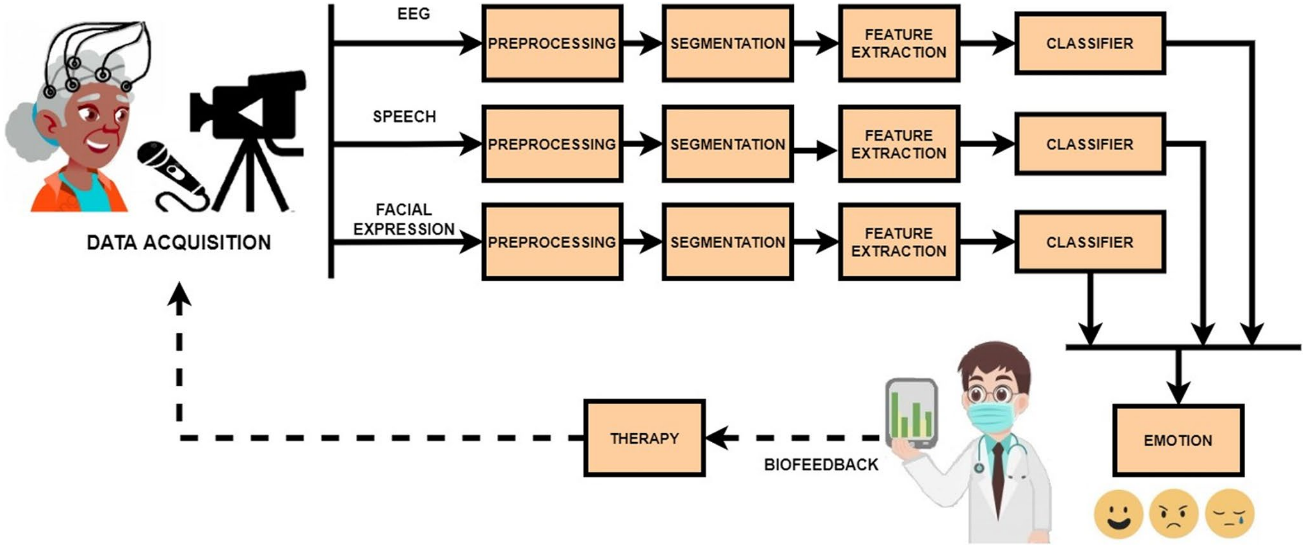


Fig. 1 General proposal of the HMI system used in this study. First, the EEG, speech, and facial expressions. After processing, the therapy is modulated according to the emotion recognition results. The patient is then benefited by the personalized therapeutic intervention

on the effectiveness of the therapeutic techniques employed, therapists can adjust the selection of music, tempo, or delivery method to maximize the desired emotional responses in patients (Sourina et al. 2012; Kim and André 2008; Muyuan et al. 2004; Lin et al. 2009; Yang et al. 2009). The incorporation of emotion recognition tools in music therapy enhances its precision and efficiency, ultimately leading to more targeted and personalized interventions, and potentially slowing the progression of Alzheimer's disease while improving the well-being and quality of life of patients (Sourina et al. 2012; Kim and André 2008; Muyuan et al. 2004; Lin et al. 2009; Yang et al. 2009).

Proposal

Considering the importance of emotion recognition in the therapeutic context (Marinoiu et al. 2018; Schipor et al. 2011; Sourina et al. 2012; Aranha et al. 2017; Arroyo-Palacios and Slater 2016), we propose an emotion recognition approach from multimodal data. It will be part of a human–machine interface (HMI) to support therapy of elderly people with dementia. Overall, the interface works as a biofeedback of emotions. This way, a therapist may change the intervention based on the patient's emotional response. In this context, we conducted some experiments to find the classification model that will integrate the system. We used public available databases of EEG and peripheral physiological signals, speech signals, and images of facial expressions.

Figure 1 illustrates the operation of the HMI. Data from 3 different sources (EEG, speech, and facial expression) are acquired from the patient. This data is then processed in the pre-processing and segmentation steps. These steps are followed by extracting features from the data. Then, we submit the feature vector to the classification step. The goal in classification is to identify the emotion felt by the patient. Finally, the therapist may assess this emotional response and use it to adapt the intervention applied to the patient.

Considering the EEG and audio signals, the pre-processing consists of applying a notch filter, to minimize the influence of the electrical network frequency (60 Hz or 50 Hz, depending on the country) and a bandpass filter, to keep the signals within the expected range. Then, the signals were segmented into windows, for later attribute extraction. As we use public databases, these two signal filtering steps had already been performed. Considering the images of facial expressions, the only preprocessing used was face segmentation, using the Haar-Cascade classifier. However, since public image databases were used, this step had already been performed, so we used the complete image of the face as available.

Therefore, this system seeks to improve the therapy of elderly people affected by dementia. It will provide emotional feedback for customized therapeutic interventions. The next sections present the tools and methods adopted to achieve this goal. The experiments and the findings shown here consist on a proof of concept for the development of the proposed HMI.

Datasets

During this study we used 3 different databases, all seeking to relate human data to their respective emotions. The first one has peripheral and central physiological signals, which are associated with 6 different classes of emotions. In the second database, speech signals from people expressing 8 different emotions were acquired. Finally, the third database consists of images of 7 emotions expressed in faces.

Further information regarding each databases are at the following topics. All three databases were initially submitted to a stage of splitting the data into trainingvalidation and test subsets. At this point, the original amount of data was randomly divided into 70% for the set used for training and validating the models and 25% for the test set. In the validation stage, to find the best ranking configuration, we use tenfold cross-validation. After defining the best classifier architecture, the entire database from the validation stage was used in the test stage as a training set to train the chosen model and test it in the previously separated test set. Figure 2 illustrates this sets' preparation. We left 5% of the datasets out to ensure there is no intersection between training/validation and test sets. Furthermore, it is important to emphasize that the test set did not participate in the training of any model, being used to assess the performance of the best model found in the training and validation stage. It is also worth mentioning the importance of designing these training-validation and test sets. This step ensures that they present the same statistical behavior even though being made of different instances.

Physiological data

The database used for emotion assessment from physiological signals was the Multimodal Database for Affect Recognition and Implicit Tagging (MAHNOB-HCI), developed by (Soleymani et al. 2011). The database has central and peripheral physiological data collected using a multimodal approach. Among the collected data there are Electroencephalogram (EEG), Electrocardiogram (ECG), Galvanic Skin Response (GSR), respiration amplitude, and skin temperature. All these data were used in our approach.

The instrumental arrangement for data collection was as follows. For GSR data, the authors used two electrodes in the distal phalanges of the middle and index fingers (Soleymani et al. 2011). To acquire ECG information they placed 3 electrodes at the upper right and left corners of the chest and abdomen. EEG was recorded from 32 active silver chloride electrodes, including 2 references, positioned according to the international 10–20 system. Finally, they measured skin temperature from a skin sensor. The authors started acquiring data from 30 volunteers with different cultural backgrounds and of both male and female genders. However, 6 subjects did not participate in all acquisition steps.

Emotional response to visual and auditory stimuli in MAHNOB-HCI was carried out in two stages. At the first one they played 20 short videos to evoke emotions while recording the physiological response of each participant. At the end of each video a neutral clip was played to minimize the emotional bias activated by the previous video and ease participants self-assessed after watching the videos. The emotional evaluation was performed using a discrete scale with values between 1 to 9 (where 1 is the most pleasant emotion and 9 is the most unpleasant). This assessment was based on five different questions: i) What emotion was presented?; ii) What level of pleasure?; iii) What level of activation?; iv) What level of dominance? and v) What is the level of predictability?. To classify the videos, the 3D model of inferences of affective states PAD (Pleasure-Activation-Dominance) was used and to answer which emotion each video is supposed to evoke. Then, the participants had to rate the stimuli in a scale from 1 to 9, corresponding to the following emotional states: neutral, anxiety, fun, sadness, happiness, disgust, anger, surprise, and fear. The data employed in our methodology was obtained during the acquisition phase. Although the authors have meticulously devised the MAHNOB-HCI database to encompass data pertaining to nine distinct classes of emotions, it is worth noting that specific data for joy, fun, and fear classes was not made available. Consequently, we leveraged the dataset comprising the remaining six emotions to fulfill our research objectives.

Additionally, at the second acquisition stage participants were asked to perform a digital content labeling tasks based on their emotional responses. During this second stage the authors acquired the reactions expressed on the face from video cameras. Data from this step were not used in our current study.

Therefore, the database we use here has 285 signals, being 55 related to Hapiness emotion, 14 to Sadness, 84 to Neutral, 40 to Disgust, 65 to Amusement, and 27 to Anger. Each signal originally has 47 channels, however, 9 were empty and 38 had some information. Among these channels, 1 to 32 had EEG signals. On channels 33, 34 and 35 were the ECG. Channel 41 was dedicated to GSR, channel 45 to respiration rate, and channel 46 to skin temperature.

Speech data

To perform the recognition of emotions through voice we used the Ryerson AudioVisual Database of Emotional Speech and Song (RAVDESS) (Livingstone and Russo 2018). This is a Canadian base composed of the voices of 24 professional actors, in English with an American accent. This database has 7356 audio and video files, totaling 25 GB of data. The amount of participants is equally divided in 12 men and 12 women.

During the recordings, each individual introduced himself speaking the following expressions in English: "Kids are talking by the door" and "Dogs are sitting by the door". They spoke these phrases in order to represent 8 emotions: neutral, calm, joy, sadness, anger, fear, surprise and disgust.

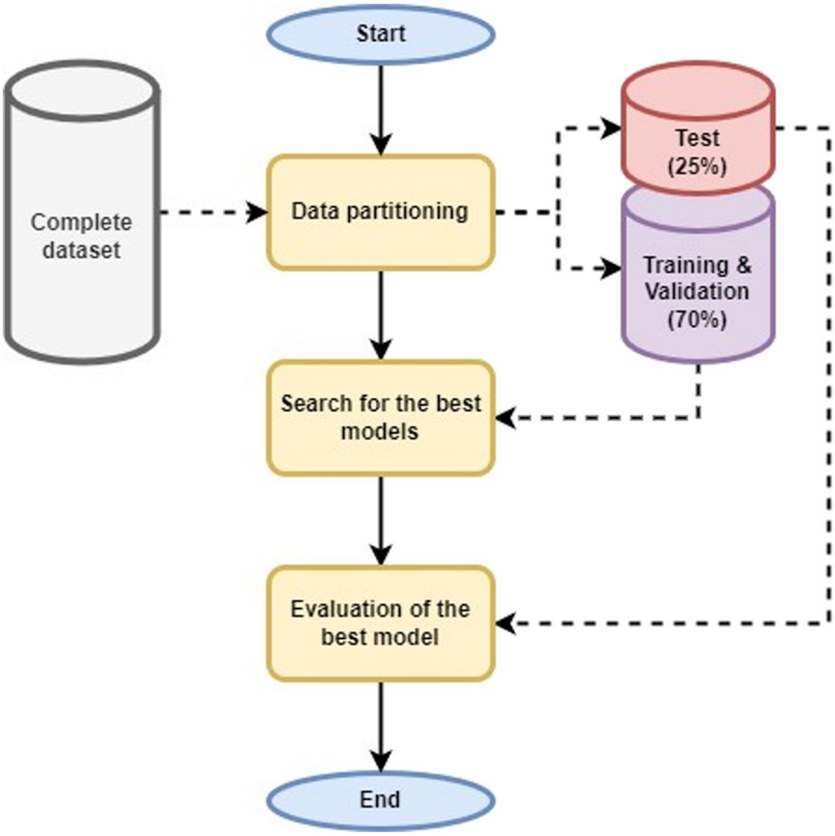

Start
Complete dataset
Data partitioning
Test (25%)
Training & Validation (70%)
Search for the best models
Evaluation of the best model
End

Fig. 2 General design of datasets

Likewise, they recorded the 2 phrases in singing tones to represent 6 emotions: neutral, calm, joy, sadness, anger and fear. The idea behind this method was to get as close as possible to the desired expression, in an induced way. For this, each actor used different techniques in order to achieve the final goal. They recorded each sentence at 2 different levels of intensity: normal and strong, plus neutral expressions. In the end, each contracted actor recorded, on average, for 4 h, with a microphone placed 20 cm in front of him.

As we would not use video signals and the focus of the work was not the singing, we selected the portion of files containing only spoken excerpts. This reduced our base to a set of 1440 exclusively audio files in.WAV format, with 48 kHz and 16bit.

The files are originally named according to Table 2, following the order of identification: modality, channel, emotion, intensity, declaration, repetition and actor. Thus, each label is composed of 7 numbers of 2 digits each (eg 02–01-06–01-02–01-12.wav).

Originally the database is divided by the actor/actor. However, as our objective is to classify emotions, we used this labeling of the files to reorganize the base according to emotion classes. Thus, we created 8 folders containing each emotion class. In this way, we can work better with the files in their characteristic groups, unlike the previous way, separated by actors.

As for the distribution of instances among the 8 classes of emotions, the base is unbalanced. Each emotion has a total of 192 files. However, Neutral emotion has only 96 audio files (Livingstone and Russo 2018).

Facial expressions

For the emotion recognition experiment in facial expressions, we used the Facial Expression Recognition 2013 (FER-2013) database, introduced in the ICML 2013.

Challenges in Representation Learning (Goodfellow et al. 2013). The FER-2013 database consists of 35,887 images, all resized to 48 × 48 pixels and converted to shades of gray, covering 7 types of emotions, namely: Anger (4,593), Disgust (547), Fear (5,121), Happy (8,989), Neutral (6,198), Sad (6,077) and Surprise (4,002). This database is currently considered the largest publicly available facial expression database for researchers who want to train machine learning models, mainly Deep Neural Networks (DNNs). Figure 3 shows examples of images from this database.

Processing and Classification

For data processing we used an approach to deal with physiological and voice signals and another approach to images of facial expressions. For the signals, features regarding their temporal, frequency and statistical distributions were extracted. As for the images, we propose an architecture for feature extraction and classification based on deep networks and the Random Forest algorithm. The procedures adopted for each database are detailed below.

Signal data

The recognition of emotions through physiological and speech signals was performed following the steps shown in the diagram in Fig. 4.

Initially, we submit the signals to a feature extraction step. In this step, we used the GNU/Octave mathematical computing software, version 4.0.3 (Eaton et al. 2015), to extract the 34 features mathematically described in the Table 3. In this way, each instance of the signal is represented by some of its statistical characteristics and in the time and frequency domains. Such attributes proved to be relevant and effective in the representation of EEG and peripheral physiological signals in previous studies with physiological and voice signals (Santana et al. 2020a; Espinola et al. 2021a, 2021b).

Statistical information was extracted by the attributes in the left column of the Table 3: mean, variance, standard deviation, root mean square, average amplitude changes, difference absolute deviation, integrated absolute value, logarithm detector, simple square integral, mean absolute value, mean logarithm kernel, skewness, kurtosis, maximum amplitude, and 3rd, 4th, and 5th moments.

The attributes related to the time–frequency domain of the signals are the right column of the Table 3: waveform length, zero crossing, slope sign changes, Hjorth parameters (activity, mobility, and complexity), mean frequency, median frequency, mean power, peak frequency, power spectrum ratio, total power, variance of central frequency, Shannon's entropy, and 1st, 2nd, and 3rd spectral moments.

During the process of extracting attributes from the peripheral and central physiological signals, we perform the windowing of these signals. We used a 5 s window with 1 s overlap between windows. This procedure aims to increase the spectral characteristics of the sample. From this windowing, we generate an unbalanced dataset with 8.097 instances. The Happiness class now has 1.704 instances, 1.114 in the Neutral class, 500 in Sadness, 1.222 in Disgust, 2.650 in Fun and 907 in Anger. Finally, each of the 38 channels of these instances was subjected to the extraction of the 34 attributes in the Fig. 3.

Since voice signals are shorter than physiological ones and it is important to analyze them in all their context, we did not perform windowing on these signals. However, some of the audios were mono (1 channel) and others stereo (2 channels), so, to avoid any kind of incompatibility, we duplicated the mono signals to the equivalent stereo. Finally, we extract the features in Fig. 3 from each channel of the signals.

After extracting attributes, we designed the training/validation and test sets as illustrated in Fig. 2. This step was performed for both the physiological and speech signal

Table 2 Description of RAVDESS filenames. The 7 identifier and respective codes

Modality	01 = Audio–video, 02 = Video-only, 03 = Audio-only
Channel	01 = Speech, 02 = Song
Emotion	01 = Neutral, 02 = Calm, 03 = Happy, 04 = Sad, 05 = Angry, 06 = Fear, 07 = Disgust, 08 = Surprised
Intensity	01 = Normal, 02 = Strong
Statement	= "Kids are talking by the door", = "Dogs are sitting by the door"
Repetition	01 = First repetition, 02 = Second repetition
Actor	01 = First actor, …, 24 = Twenty-fourth actor

Identifier, Coding description of factor levels

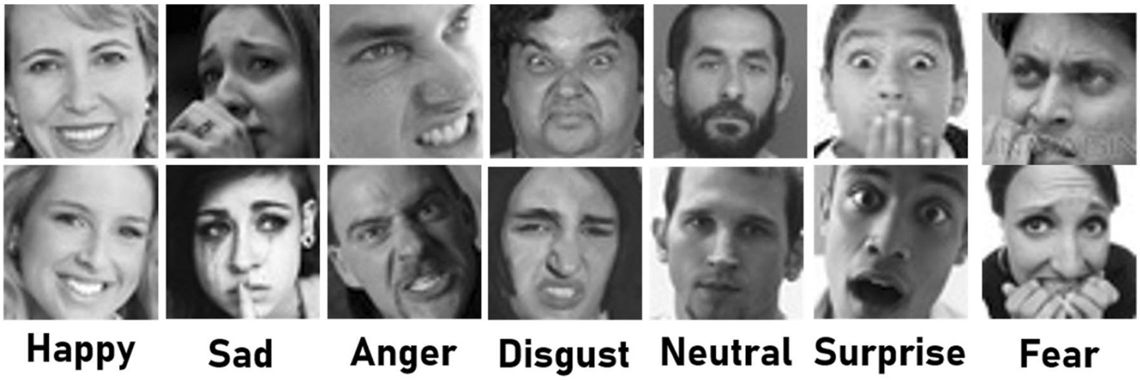


Fig. 3 Samples of images from FER-2013 database. The image shows two examples for each class on the database

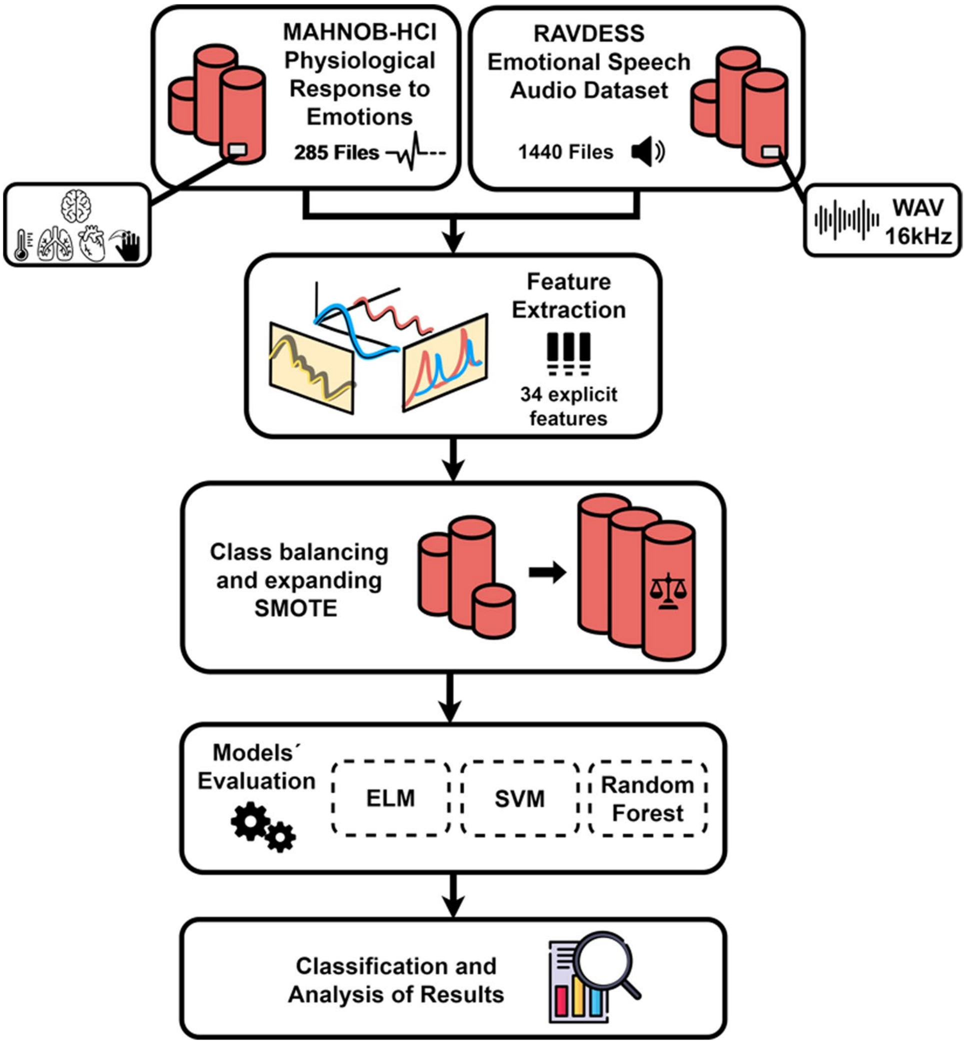


Fig. 4 General proposal for signal data assessment. We submitted MAHNOB-HCI e RAVDESS signal databases to the following steps: feature extraction; class balancing; training and validation with intelligent models; and results assessment

databases. The test sets (25%) were set aside to be used after finding the most suitable model for the classification of each type of signal. Therefore, the steps described below were performed only with the training/validation set, which corresponds to 70% of the instances of each database.

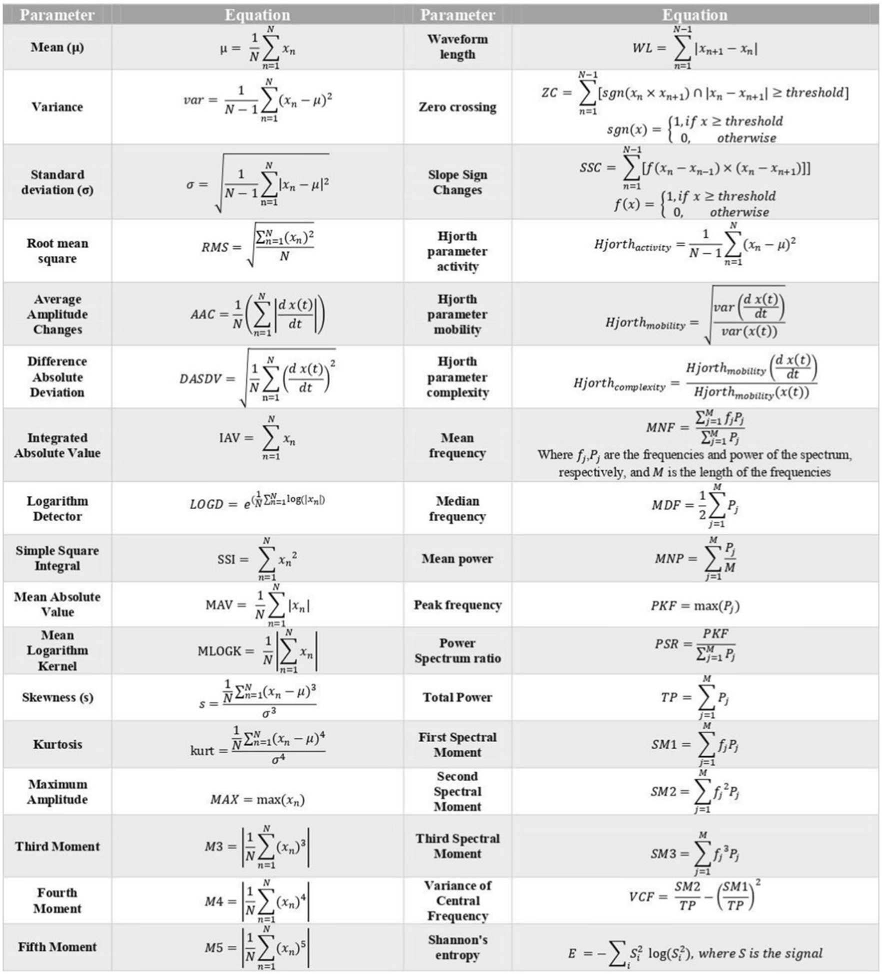
As mentioned before, for both physiological and speech signals, there is an unbalanced distribution of instances in the respective classes. If not adjusted, this class imbalance can generate biased learning, favoring classes with more representatives (instances). To avoid this unfair learning, we performed a class balancing step. Here we balanced classes by adding synthetic instances to minority classes using the Synthetic Minority Over-sampling Technique (SMOTE) (Blagus and Lusa 2013; Chawla et al. 2002). This algorithm creates synthetic instances based on the real instances of a given class. Minority classes are balanced by taking each instance and adding synthetic samples along the line segments joining their k nearest neighbors. In our approach, we configure SMOTE with $k = 3$ neighbors. For this step, we used the Waikato Environment for Knowledge Analysis (Weka) software, version 3.8 (Witten and Frank 2005).

In the physiological signals database, the balancing of the classes resulted in 2.649 instances in the Sadness class, 2.658 in the Happiness class, 2.651 in the Disgust and Neutral classes, 2.650 in the Fun class and the Anger class with 2.658 instances. Therefore, this set started to have a more balanced distribution of the instances in the classes.

In the context of speech signals, it was still necessary to apply SMOTE to expand the total number of instances of each class by 50%. We performed this procedure after balancing the classes in order to improve the distribution of the number of instances along the set, since there is a large number of classes in this problem (8). This expansion was an adjustment in the dimensionality of the set, that is, better balancing the relationship between the number of instances, attributes and classes. Thus, each emotion now has 288 instances.

Finally, the balanced sets of both databases were submitted to the training/validation stage. In this step, we evaluated

Table 3 List of the 34 features with their mathematical representations



Parameter	Equation	Parameter	Equation
Mean (μ)	$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n$	Waveform length	$WL = \sum_{n=1}^{N-1} \lvert x_{n+1} - x_n \rvert$
Variance	$var = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu)^2$	Zero crossing	$ZC = \sum_{n=1}^{N-1}[sgn(x_n \times x_{n+1}) \cap \lvert x_n - x_{n+1} \rvert \geq threshold]$ $sgn(x) = \begin{cases} 1, if\ x \geq threshold \\ 0, \quad otherwise \end{cases}$
Standard deviation (σ)	$\sigma = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N} \lvert x_n - \mu \rvert^2}$	Slope Sign Changes	$SSC = \sum_{n=1}^{N-1}[f(x_n - x_{n-1}) \times (x_n - x_{n+1})]]$ $f(x) = \begin{cases} 1, if\ x \geq threshold \\ 0, \quad otherwise \end{cases}$
Root mean square	$RMS = \sqrt{\frac{\sum_{n=1}^{N}(x_n)^2}{N}}$	Hjorth parameter activity	$Hjorth_{activity} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu)^2$
Average Amplitude Changes	$AAC = \frac{1}{N}\left(\sum_{n=1}^{N}\left\lvert\frac{d\,x(t)}{dt}\right\rvert\right)$	Hjorth parameter mobility	$Hjorth_{mobility} = \sqrt{\frac{var\left(\frac{d\,x(t)}{dt}\right)}{var(x(t))}}$
Difference Absolute Deviation	$DASDV = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\frac{d\,x(t)}{dt}\right)^2}$	Hjorth parameter complexity	$Hjorth_{complexity} = \frac{Hjorth_{mobility}\left(\frac{d\,x(t)}{dt}\right)}{Hjorth_{mobility}(x(t))}$
Integrated Absolute Value	$\mathrm{IAV} = \sum_{n=1}^{N} x_n$	Mean frequency	$MNF = \frac{\sum_{j=1}^{M} f_j P_j}{\sum_{j=1}^{M} P_j}$ Where f_j,P_j are the frequencies and power of the spectrum, respectively, and M is the length of the frequencies
Logarithm Detector	$LOGD = e^{(\frac{1}{N}\sum_{n=1}^{N}\log(\lvert x_n \rvert)}$	Median frequency	$MDF = \frac{1}{2}\sum_{j=1}^{M} P_j$
Simple Square Integral	$\mathrm{SSI} = \sum_{n=1}^{N} {x_n}^2$	Mean power	$MNP = \sum_{j=1}^{M}\frac{P_j}{M}$
Mean Absolute Value	$\mathrm{MAV} = \frac{1}{N}\sum_{n=1}^{N} \lvert x_n \rvert$	Peak frequency	$PKF = \max(P_j)$
Mean Logarithm Kernel	$\mathrm{MLOGK} = \frac{1}{N}\left\lvert\sum_{n=1}^{N} x_n\right\rvert$	Power Spectrum ratio	$PSR = \frac{PKF}{\sum_{j=1}^{M} P_j}$
Skewness (s)	$s = \frac{\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^3}{\sigma^3}$	Total Power	$TP = \sum_{j=1}^{M} P_j$
Kurtosis	$\mathrm{kurt} = \frac{\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^4}{\sigma^4}$	First Spectral Moment	$SM1 = \sum_{j=1}^{M} f_j P_j$
Maximum Amplitude	$MAX = \max(x_n)$	Second Spectral Moment	$SM2 = \sum_{j=1}^{M} {f_j}^2 P_j$
Third Moment	$M3 = \left\lvert\frac{1}{N}\sum_{n=1}^{N}(x_n)^3\right\rvert$	Third Spectral Moment	$SM3 = \sum_{j=1}^{M} {f_j}^3 P_j$
Fourth Moment	$M4 = \left\lvert\frac{1}{N}\sum_{n=1}^{N}(x_n)^4\right\rvert$	Variance of Central Frequency	$VCF = \frac{SM2}{TP} - \left(\frac{SM1}{TP}\right)^2$
Fifth Moment	$M5 = \left\lvert\frac{1}{N}\sum_{n=1}^{N}(x_n)^5\right\rvert$	Shannon's entropy	$E = -\sum_i S_i^2 \log(S_i^2)$, *where S is the signal*

the performance of 3 classic classifier algorithms: Random Forest, Support Vector Machine (SVM) and Extreme Learning Machine. Random Forest is an algorithm structured by a committee of decision trees. Individually, these trees act as experts in identifying the patterns associated with the problem (Breiman 2001; Jackins et al. 2021; Pal 2005). SVM is a method that stands out for its good generalization performance in classification problems (Platt 1998; Cortes and Vapnik 1995; Zeng et al. 2021). It is based on the modeling of hyperplanes that serve as decision boundaries for problem solving. ELM stands out for its great generalization power and its reduced training time due to the random initialization of the input layer weights and the analytical calculation of the subsequent weights (Santana et al. 2018; Huang et al. 2004; Silva and Krohling 2016).

We investigated SVM because it is a well-established classification architecture in the context of biomedical signal and image classification. Additionally, ELM was used because it can also be considered as a fast algorithm for two-layer multilayer perceptron networks, another classic approach to biomedical problems. We also investigated Random Forest for its potential good performance in problems where generalization is difficult. Due to its ensemble behavior, Random Forest is robust to class imbalance and can deal well with problems that are expressed through multiple rules, unlike SVM and ELM which, in turn, look for general rules in the form of combinations of polynomials and other mathematical functions.

We tested different models for each of these methods, varying their main hyperparameters in the configurations presented in Table 4. It is worth mentioning that the k-fold cross-validation method with $k = 10$ was used during the experiments to avoid overfitting (Jung and Hu 2015). In this method, the dataset was randomly divided into k subsets, with $k - 1$ used for training and the remaining subset used for validation. In this way, successive training steps are performed until the performance is validated for all k sets. Furthermore, in order to obtain statistical information regarding the performance of the algorithms, each configuration was evaluated for 30 repetitions. These experiments were also conducted in Weka, version 3.8.

Table 4 Experimental settings for the classifiers applied to signal data

Classifiers	Settings
Random Forest	trees: 10, 20, 50, 100, 150, 200, 250, 300, and 350 maxDepth: unlimited numDecimalPlaces: 2 numExecutionSlots: 1
SVM	Kernel functions: linear, polynomial (d = 2, d = 3, and d = 4), RBF ($\gamma = 0.50$) C = 1.0
ELM	Kernel functions: linear, polynomial (d = 2, d = 3, and d = 4), RBF and sigmoid Neurons in the hidden layer: 500

The Weka environment was chosen because it is easy to prototype, separating the choice of the machine learning model from the final prototyping in the application, thus allowing users to build complex machine learning models for different applications. Weka also allows the chosen models to be saved in a file, for later application in the final emotion recognition solution.

Image data

For the recognition of emotions through images of facial expressions we propose a new architecture based on deep network with a Random Forest classifier in the output. As illustrated in the Fig. 5, in this architecture we use a transfer learning approach to extract attributes from images from the

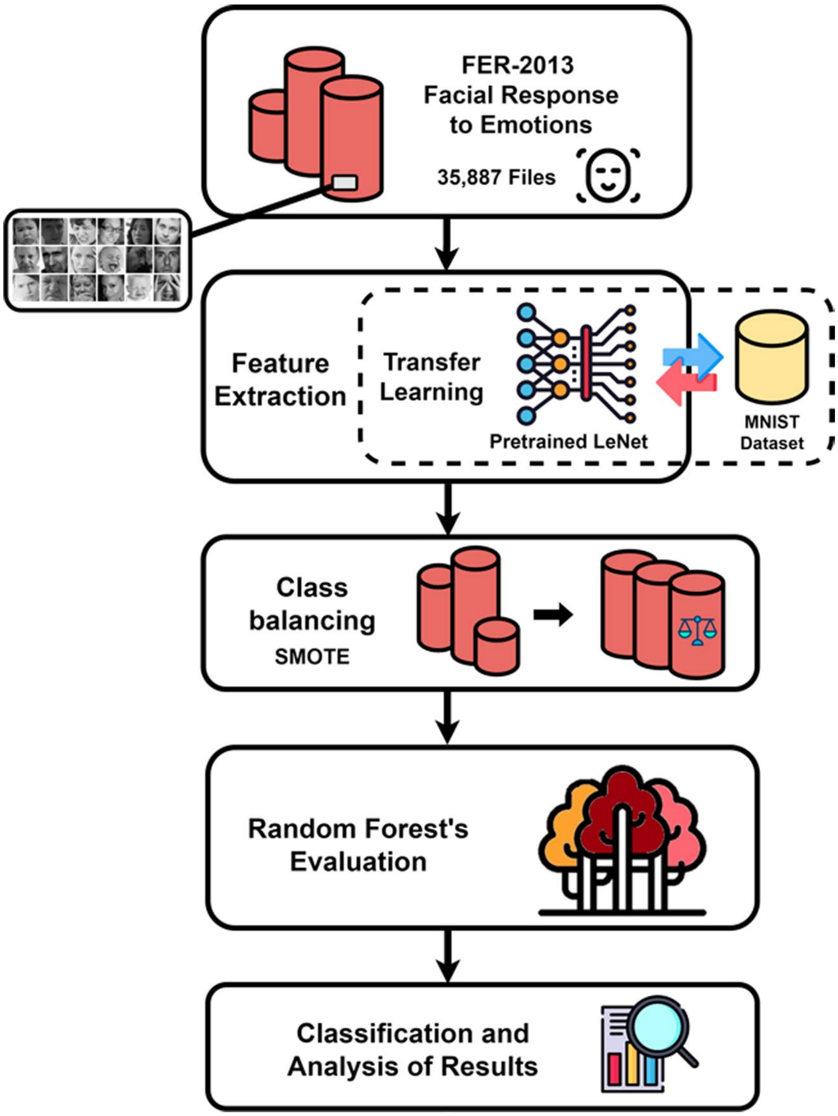


Fig. 5 General proposal for image data assessment. We submitted FER2013 image database of emotion driven facial expression to the following steps: feature extraction; class balancing; training and validation with random forest; and results assessment

FER2013 database. For this, we applied a pre-trained LeNet network with the MNIST dataset, composed of a training set of 60.000 images of handwritten digits (Deng 2012). LeNet was one of the first Convolutional Neural Networks. It was proposed by (LeCun et al. 1998) and has 7 layers, being 3 convolutional layers, 2 downsampling layers, and 2 fully connected layers. The convolution filters are used to extract spatial features from the images. Therefore, this network extracted 500 features from each image from the database.

After extracting attributes, we designed the training/validation and test sets (Fig. 2). The test set was only used after finding the most suitable model for image classification. Therefore, the steps described below were performed only with the training/validation set, which corresponds to 70% of the instances.

Since the database we used has a notable imbalance between classes, we also submit the feature vectors of this set to the SMOTE (Chawla et al. 2002) method for balancing. As for the data from the other databases, we configured SMOTE with $k = 3$ close neighbors. Balancing increased the pool to 62269 instances. These instances became better distributed among the classes happy (8.989), fear (8.961), neutral (8.987), disgust (8.861), sad (8.872), anger (8.915), and surprise (8.684).

Finally, the balanced set was submitted to classification with Random Forest algorithms. This algorithm was chosen to compose this architecture because it is versatile, fast-executing and deals well with large datasets. Methods based on Random Forest have also been successful in sets with missing data, with poor balance and with little variability (Andrade et al. 2020; Oliveira et al. 2020; Gomes et al. 2020; Freitas Barbosa et al. 2021).

We conducted experiments with different models of Random Forest, varying the number of trees between 10, 20, 50, 100, 150, 200, 250, 300, 350, 400 and 500. The Table 5 details the settings. In these experiments we also used the k-fold cross-validation method with $k = 10$ to avoid overfitting (Jung and Hu 2015). In addition, each configuration was evaluated for 30 repetitions to verify the statistical behavior of the models. Both the class balancing and classification steps were carried out by Weka, version 3.8.

Table 5 Experimental settings for the classifiers applied do image data

Classifier	Settings
Random Forest	trees: 10, 20, 50, 100, 150, 200, 250, 300, 350, 400, and 500 maxDepth: unlimited numDecimalPlaces: 2 numExecutionSlots: 1

Test stage

In the test stage, we used the subsets with 25% of each dataset (physiological signals, speech signals and images of facial expressions) that did not participate in the evaluation stages of the classification models. This step is important to verify the generalization capacity of the evaluated models. A good generalization is desirable, as it implies a good performance of the model to classify new data. "New data" is data that did not take part in the training and is therefore unknown to the algorithm.

From the training and validation stages, we assessed the performance of the tested models. This analysis allowed us to identify the most suitable models to deal with each of the three data sources. After identifying these models, we trained each one with the entire training/validation set. Finally, we use the trained models to estimate the classes of the data in the test sets.

Metrics

To evaluate the performance of the models in both the training/validation and the test stages, we used five metrics: accuracy, kappa index, sensitivity, specificity and area of the ROC curve. It is important to note that for the analysis of the models we take into account this set of performance metrics and not just one of them. The Table 6 presents the mathematical description of these metrics.

Accuracy is a metric that indicates how efficient the classifier is at correctly predicting the class of each instance. It is an index directly proportional to the true positives (TP) and true negatives (TN) rates. The kappa statistic is a metric similar to accuracy. However, kappa takes into account the random hit chance (Artstein and Poesio 2008). When predictions are purely random, kappa index assumes 0 (zero) or negative values. Sensitivity is

Table 6 Mathematical expressions for the metrics used to evaluate the classification performance. TP, TN, FP and FN are the quantity of True Positives, True Negatives, False Positives and False Negatives, respectively. TPR and FPR are the True Positive Rate and False Positive Rate, respectively

Metric	Mathematical expression
Accuracy	$Acc = \frac{TP + TN}{TP + TN + FP + FN}$
Sensitivity	$Sens = \frac{TP}{TP + FN}$
Specificity	$Esp = \frac{TN}{TN + FP}$
Kappa	$\kappa = \frac{\rho_0 - \rho_e}{1 - \rho_e}$, where ρ_0 is the observed agreement rate and ρ_e is the expected agreement rate, given by: $\rho_e = \frac{(TP + FP)(TP + FN) + (FN + TN)(FP + TN)}{(TP + FP + FN + TN)^2}$
Area under the ROC curve	$AUC = \int TPR \, d\,(FPR)$

the metric used to assess the classifier's performance in identifying the true positives. Sensitivity is commonly called the true positive rate (TPR), but it is also known as recall. Specificity, as opposed to sensitivity, it is used to assess performance in identifying the true negatives. Thus, it is known as the true negative rate (TNR). The area under the ROC (Receiver Operating Characteristics) curve, widely known as Area Under Curve (AUC), is also a metric used to assess how well the model performs in the prediction. ROC curve is a probabilistic curve and the area under it represents the chance the model has to correctly predicts the data (Hanley and McNeil 1982). The curve is built from the false positive rate (FPR) on the x-axis, and the sensitivity (TPR) on the y-axis. In the case of multiclass problems, such as in this work, AUC can be evaluated in two ways: one vs one; one vs all. In the first one, all the curves of the combination of all classes with each other are plotted in pairs. In the second, the curves of the combination of one class versus all others are plotted.

Results

In this section, we show the results of the training-validation and test stages for the three databases. Initially, we present the results for the physiological signals base, followed by the voice signals database. Finally, there are the results for the identification of emotions in facial expressions.

Physiological data

Table 7 and Fig. 6 show the performance of the classifiers in the training-validation stage. In Table 7 are the average and standard deviation values for all performance metrics. In addition, we highlight in this table the settings of each classifier family with the best performance. Sometimes, many configurations of the same classifier family obtained similar results. Thus, we performed a joint analysis of the metrics and the intervals defined by the standard deviations to establish the best setting.

In this context, using the physiological signals, Random Forest with 300 trees presented the best performance. This model achieved high average values of accuracy (98.48%), kappa (0.9817), sensitivity (0.9957), specificity (0.9977) and AUC (0.9999). In the SVM family, we found the best results using the polynomial kernel of 2nd degree. These results were a little lower than those with Random Forest, with an accuracy of 95.29%, kappa index of 0.9435, sensitivity of 0.9887, 0.9942 specificity and AUC of 0.9968. The ELM configurations had the worst performances when compared to the other families. ELM with 4th degree polynomial kernel showed the best results among the ELMs. This ELM model achieved an accuracy of 44.28%, kappa of 0.3314, sensitivity of 0.4429, specificity of 0.8886 and 0.6657 for AUC. As mentioned earlier, we assessed all settings for 30 repetitions to verify their statistical behavior. Therefore, in the graphs of Fig. 6 we present the behavior of all metrics for these best configurations of each classifier family.

Table 7 Classification performance for the dataset from physiological signals

Classifier		Accuracy (%)	Kappa statistic	Sensitivity	Specificity	AUC
Random Forest	10 trees	98.08 ± 0.35	0.9770 ± 0.0042	0.9959 ± 0.0037	0.9966 ± 0.0016	0.9996 ± 0.0007
	20 trees	98.28 ± 0.33	0.9793 ± 0.0040	0.9956 ± 0.0040	0.9972 ± 0.0015	0.9998 ± 0.0005
	50 trees	98.40 ± 0.31	0.9808 ± 0.0037	0.9955 ± 0.0039	0.9975 ± 0.0014	0.9999 ± 0.0001
	100 trees	98.44 ± 0.29	0.9813 ± 0.0035	0.9956 ± 0.0039	0.9976 ± 0.0013	0.9999 ± 0.0001
	150 trees	98.46 ± 0.29	0.9815 ± 0.0034	0.9956 ± 0.0039	0.9977 ± 0.0013	0.9999 ± 0.0001
	200 trees	98.47 ± 0.29	0.9816 ± 0.0035	0.9957 ± 0.0039	0.9977 ± 0.0013	0.9999 ± 0.0001
	250 trees	98.47 ± 0.29	0.9816 ± 0.0034	0.9958 ± 0.0039	0.9977 ± 0.0013	0.9999 ± 0.0001
	300 trees	98.48 ± 0.29	0.9817 ± 0.0034	0.9957 ± 0.0039	0.9977 ± 0.0013	0.9999 ± 0.0001
	350 trees	98.47 ± 0.28	0.9817 ± 0.0034	0.9958 ± 0.0039	0.9976 ± 0.0013	0.9999 ± 0.0001
SVM	linear	86.79 ± 0.76	0.8414 ± 0.0092	0.9859 ± 0.0075	0.9751 ± 0.0043	0.9886 ± 0.0022
	poly 2	95.29 ± 0.52	0.9435 ± 0.0062	0.9887 ± 0.0063	0.9942 ± 0.0021	0.9968 ± 0.0014
	poly 3	93.09 ± 1.76	0.9171 ± 0.0212	0.9795 ± 0.0274	0.9906 ± 0.0044	0.9956 ± 0.0029
	poly 4	80.81 ± 2.66	0.7697 ± 0.0319	0.9491 ± 0.0416	0.9727 ± 0.0093	0.9892 ± 0.0048
	RBF	95.11 ± 0.52	0.9414 ± 0.0063	0.9826 ± 0.0074	0.9919 ± 0.0025	0.9868 ± 0.0046
ELM	linear	29.14 ± 7.46	0.1497 ± 0.0896	0.2914 ± 0.2365	0.8583 ± 0.0892	0.5748 ± 0.1134
	poly 2	40.32 ± 5.34	0.2838 ± 0.0640	0.4032 ± 0.2365	0.8806 ± 0.0611	0.6419 ± 0.1144
	poly 3	41.86 ± 7.15	0.3023 ± 0.0858	0.4186 ± 0.2453	0.8837 ± 0.0600	0.6512 ± 0.1218
	poly 4	44.28 ± 5.57	0.3314 ± 0.0669	0.4429 ± 0.2392	0.8886 ± 0.0593	0.6657 ± 0.1165
	RBF	16.67 ± 0.01	0.0000 ± 0.0000	0.1667 ± 0.3728	0.8333 ± 0.3728	0.5000 ± 0.0000
	sigmoid	38.87 ± 6.94	0.2664 ± 0.0832	0.3887 ± 0.2373	0.8777 ± 0.0676	0.6332 ± 0.1171

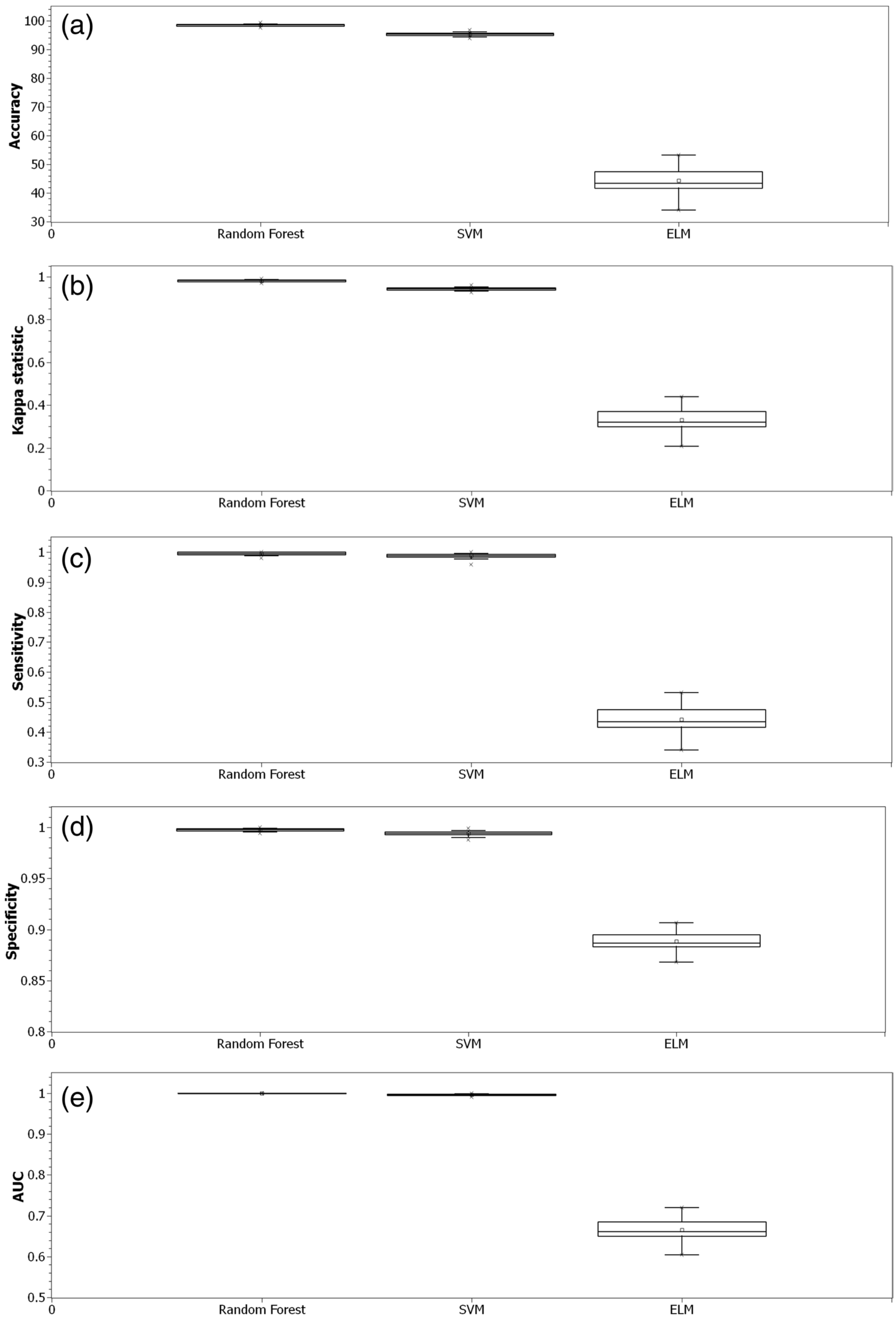


Fig. 6 Emotion classification performance from physiological signals. Each classifier family was assessed based on (**a**) accuracy, (**b**) kappa statistic, (**c**) sensitivity, (**d**) specificity, and (**e**) AUC

Considering these results, we used the previously trained Random Forest model with 300 trees to perform the prediction of emotions in the test set. The confusion matrix in Fig. 7 shows how this model distributed the test data in the 6 classes. The density of instances is also indicated by the color, so that the closer to green, the greater the number of instances in that region. Thus, you may see that the instances were almost all concentrated along the main diagonal of the matrix. This phenomena indicates that most of the data were correctly classified. Furthermore, in Table 8 we present the values for the rating performance evaluation metrics. Therefore, the Random Forest model was able to classify the test set with 99.159% accuracy, kappa of 0.989, sensitivity of 0.992, 0.998 specificity, and AUC of 1.

Speech data

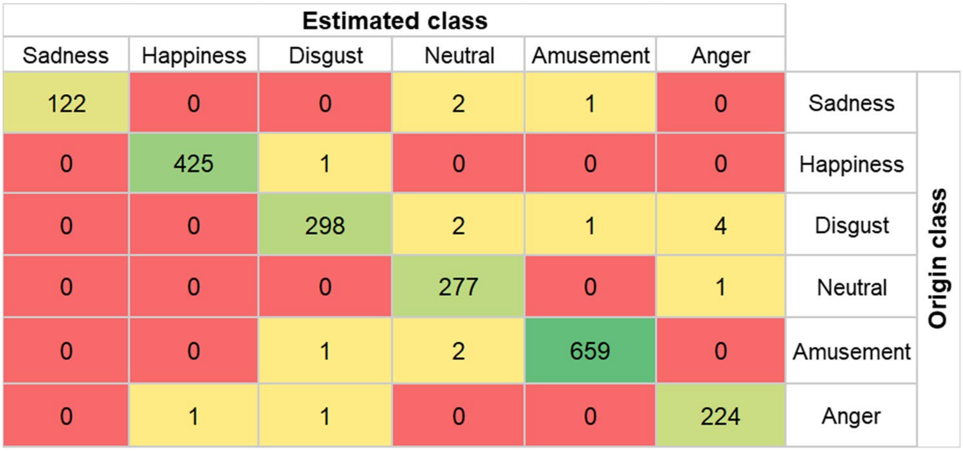
In classifying emotions in voice signals, the outcome of the 3 types of architectures (Table 9) shows that for this type of data we can still go a long way. With Random Forests, the 10-tree configuration had the worst result, reaching 38.03% average accuracy. As expected, as we increased the number of trees in the model, we noticed an increase in overall performance. Yet, this increase was not significant, to the point that the setting with the best performance was the Random Forest with 300 trees. This model achieved an average accuracy of 43.01%, with a kappa statistic of 0.3488, sensitivity of 0.5848, 0.9073 of specificity and AUC of 0.8958. The performances of SVMs were similar to Random Forests, being slightly worse. The best performing SVM configuration was a 4th-degree polynomial kernel, with an accuracy of 42.78%, 0.3460 kappa, 0.6035 sensitivity, 0.9110 specificity, and 0.8785 AUC. Among the ELM configurations, the 3rd-degree polynomial kernel stood out positively. However, the ELM results were even worse than those obtained with the other architectures. The highest accuracy obtained with this architecture was 40.79%, which is associated with a kappa of 0.3233, sensitivity of 0.4079, specificity of 0.9154 and AUC of 0.6616. The graphs in Fig. 8 present the behavior of all metrics for the best settings of the three classifier families for the speech data.

From these results, we found that Random Forest with 300 trees achieved the best performance among the models. Therefore, we used this architecture to create the model for classifying the test data. The matrix in Fig. 9 shows the distribution of instances along the 8 classes of the problem. Although there is still confusion between the classes, it is observed that most instances were correctly classified. The accuracy associated with this classification was 79,888% (Table 10). When applied to the test set, this model also resulted in good results for kappa (0.769), sensitivity (0.799), specificity (0.971) and AUC (0.965).

Facial expressions

For the training-validation stage with the facial expression data we obtained the results in Table 11 and Fig. 10. The plots in Fig. 10 illustrate the general behavior of the evaluated Random Forest configurations. By increasing the number of trees, there was also an improvement in performance until reaching a certain plateau, where performance became almost constant. This certain point of stability was reached from the configuration with 350 trees. After this configuration, the increase in the number of trees did not lead to significant increases in any of the performance metrics. This behavior was repeated for all metrics. Yet, there was a little more data dispersion in the sensitivity and specificity metrics. Random Forest with 350 trees resulted in an average accuracy of 75.29%, kappa of 0.7117, sensitivity of 06116, 0.9233 of specificity and 0.8858 of AUC. So this was the setting used to create the model to classify the test set.

Fig. 7 Confusion matrix regarding the classification of emotions on the test set of physiological signals with a Random Forest of 300 trees



Estimated class							
Sadness	Happiness	Disgust	Neutral	Amusement	Anger		
122	0	0	2	1	0	Sadness	Origin class
0	425	1	0	0	0	Happiness	
0	0	298	2	1	4	Disgust	
0	0	0	277	0	1	Neutral	
0	0	1	2	659	0	Amusement	
0	1	1	0	0	224	Anger	

Table 8 Test results of the best overall method (Random Forest with 300 trees) in the classification of physiological data

Accuracy	99.159%
Kappa	0.989
Sensitivity	0.992
Specificity	0.998
AUC	1.000

Model Random Forest (300 trees)

Table 12 presents the classification performance of the test set. For this database, the model classified the instances with 82,752% of accuracy, kappa of 0.791, 0.828 of sensitivity, 0.962 of specificity and AUC of 0.975. The distribution of these instances in the 7 classes of the problem is shown in Fig. 11. Once again, the confusion matrix shows that most instances were correctly classified.

Discussion

Using the physiological signals database, we found interesting results in classifying the emotions (i.e. sadness, happiness, disgust, neutral, amusement, and anger). The results in training-validation phase pointed to a good performance of Random Forest and SVM algorithms. Sensitivity, specificity and AUC values for these methods were similar to each other. However, Random Forest presented slightly higher results of accuracy and kappa statistic. Both algorithms also showed high reliability in the results, since there was almost no data dispersion. On the other hand, the ELM configurations presented much lower results than those obtained with the other classifiers. The performance of the best ELM setting showed greater dispersion. It was also statistically dissociated from the others for all evaluated metrics (see Fig. 6). From these findings, we selected the Random Forest model with 300 trees as the one with the best performance. This last model was later used to classify the test set.

In the test step with physiological data most instances were correctly classified. The matrix in Fig. 7 shows this result. There was low confusion between classes, with few instances placed outside their origin class. The biggest confusion rate was between the disgust and anger classes, with 4 instances of disgust placed in anger. Furthermore, except for the classes happiness and anger, all other classes had instances classified as neutral. On the other hand, the neutral and happiness classes showed the lowest rate of confusion, with only 1 instance being classified outside its class of origin. In total, of the 2022 instances of the test set, only 17 were incorrectly classified, representing an error of 0.84%. This test performance is confirmed by the high values of all evaluation metrics presented in Table 8.

Considering the speech data, the results of the proposed method were not all positive. Overall, the results for the Random Forest, SVM and ELM configurations were similar to each other. They achieved median values for most of the evaluated metrics. The results in Fig. 8 prove that there

Table 9 Classification performance for the dataset from speech signals

Classifier		Accuracy (%)	Kappa statistic	Sensitivity	Specificity	AUC
Random Forest	10 trees	38.03 ± 3.54	0.2918 ± 0.0406	0.5591 ± 0.1096	0.8889 ± 0.0272	0.8464 ± 0.0449
	20 trees	39.91 ± 3.52	0.3133 ± 0.0403	0.5716 ± 0.1149	0.8977 ± 0.0275	0.8733 ± 0.0354
	50 trees	41.91 ± 3.78	0.3362 ± 0.0432	0.5879 ± 0.1125	0.9034 ± 0.0264	0.8883 ± 0.0309
	100 trees	42.44 ± 3.59	0.3423 ± 0.0411	0.5840 ± 0.1144	0.9065 ± 0.0252	0.8933 ± 0.0295
	150 trees	42.69 ± 3.71	0.3451 ± 0.0425	0.5841 ± 0.1129	0.9072 ± 0.0247	0.8942 ± 0.0294
	200 trees	42.71 ± 3.73	0.3454 ± 0.0427	0.5825 ± 0.1160	0.9072 ± 0.0242	0.8949 ± 0.0296
	250 trees	42.89 ± 3.79	0.3474 ± 0.0434	0.5828 ± 0.1154	0.9066 ± 0.0249	0.8953 ± 0.0296
	300 trees	43.01 ± 3.77	0.3488 ± 0.0431	0.5848 ± 0.1147	0.9073 ± 0.0245	0.8958 ± 0.0295
	350 trees	43.06 ± 3.79	0.3494 ± 0.0433	0.5858 ± 0.1149	0.9079 ± 0.0245	0.8960 ± 0.0293
SVM	linear	36.42 ± 3.33	0.2734 ± 0.0381	0.5253 ± 0.1150	0.9075 ± 0.0247	0.8491 ± 0.0363
	poly 2	37.65 ± 3.33	0.2874 ± 0.0381	0.5634 ± 0.1158	0.9143 ± 0.0237	0.8683 ± 0.0359
	poly 3	39.95 ± 3.240	0.3137 ± 0.0370	0.5927 ± 0.1034	0.9106 ± 0.0243	0.8778 ± 0.0334
	poly 4	42.78 ± 3.79	0.3460 ± 0.0433	0.6035 ± 0.1128	0.9110 ± 0.0252	0.8785 ± 0.0338
	RBF	37.35 ± 3.45	0.2840 ± 0.0394	0.5370 ± 0.1180	0.9135 ± 0.0235	0.8564 ± 0.0368
ELM	linear	34.47 ± 8.74	0.2511 ± 0.0999	0.3447 ± 0.2758	0.9064 ± 0.0793	0.6256 ± 0.1144
	poly 2	39.74 ± 8.74	0.3113 ± 0.0999	0.3974 ± 0.2579	0.9139 ± 0.0692	0.6556 ± 0.1126
	poly 3	40.79 ± 9.96	0.3233 ± 0.1024	0.4079 ± 0.2400	0.9154 ± 0.0688	0.6616 ± 0.1040
	poly 4	40.20 ± 10.11	0.3165 ± 0.1155	0.4020 ± 0.2562	0.9146 ± 0.0686	0.6583 ± 0.1115
	RBF	13.22 ± 0.99	0.0083 ± 0.0114	0.1322 ± 0.3298	0.8760 ± 0.3245	0.5041 ± 0.0170
	sigmoid	37.50 ± 8.09	0.2857 ± 0.0924	0.3750 ± 0.2456	0.9107 ± 0.0737	0.6428 ± 0.1042

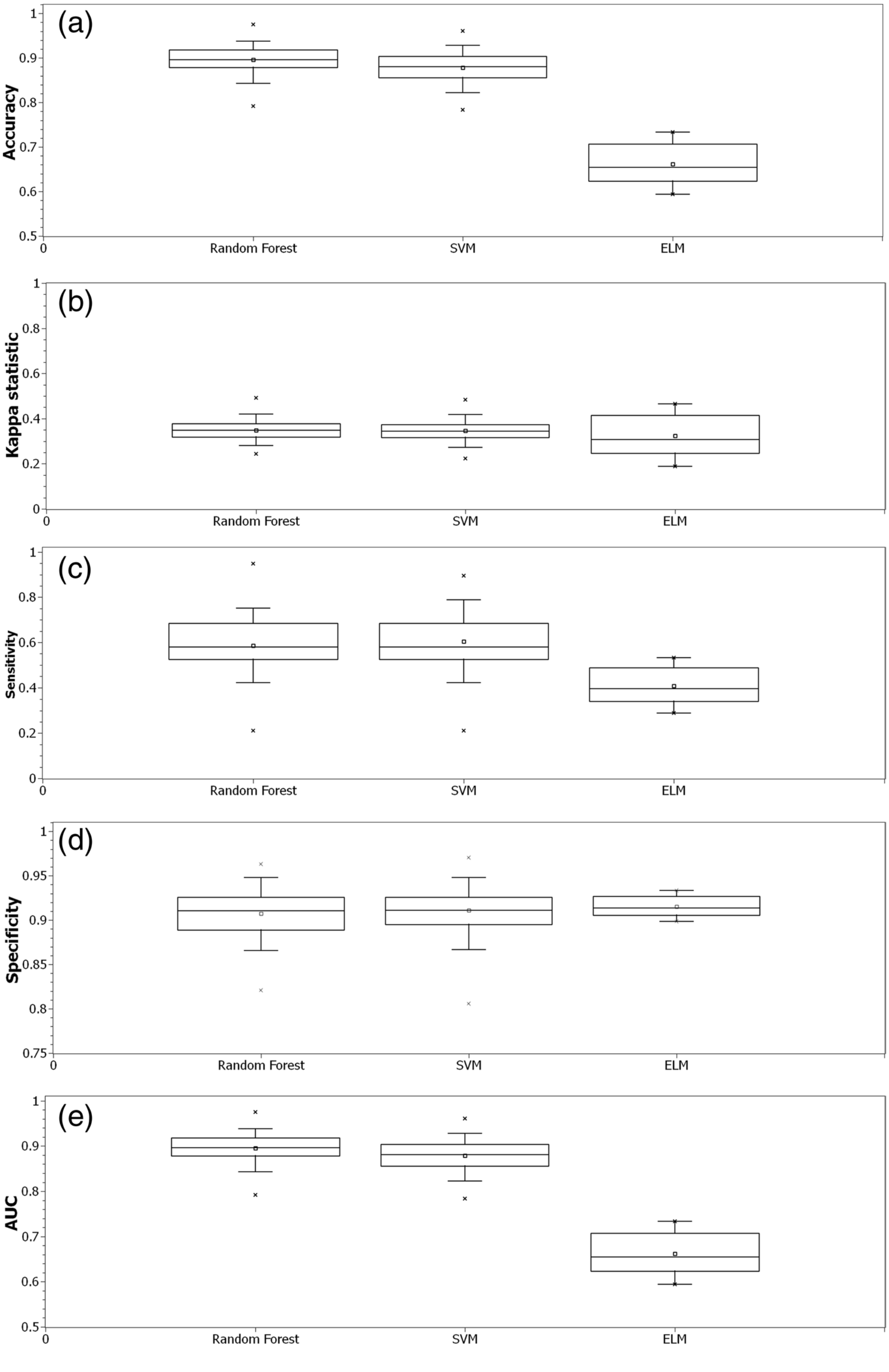
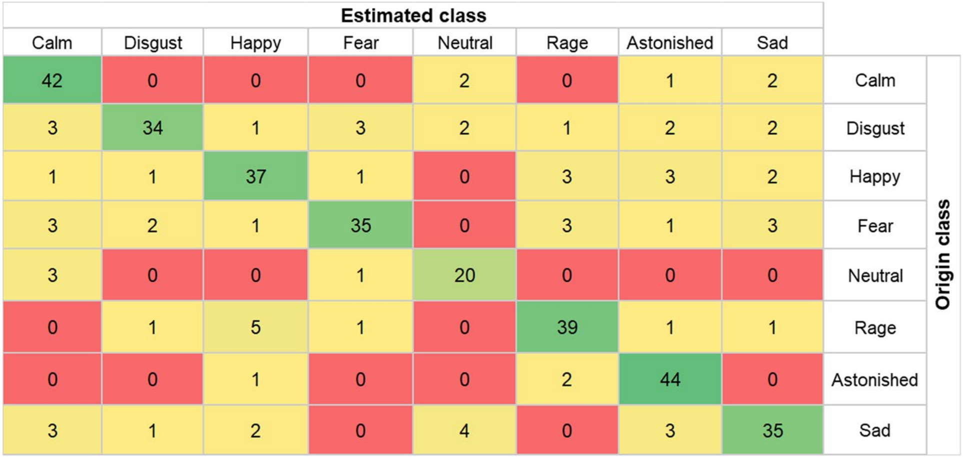


Fig. 8 Emotion classification performance from speech data. Each classifier family was assessed based on (a) accuracy, (b) kappa statistic, (c) sensitivity, (d) specificity, and (e) AUC

was no statistical difference between these architectures if we look to the results of kappa, sensitivity and specificity. However, there was difference for the metrics accuracy and AUC, with Random Forest and SVM performing better than ELM. In most cases ELM also presented greater data dispersion than the other architectures. It is also worth mentioning the good results of specificity and AUC. This phenomenon indicates that the algorithms were more capable of identifying the classes to which a given instance does not belong to than identifying which is the correct class of that instance. It is important to highlight the large number of emotion classes in this voice database. In addition, this



Estimated class									
Calm	Disgust	Happy	Fear	Neutral	Rage	Astonished	Sad		
42	0	0	0	2	0	1	2	Calm	Origin class
3	34	1	3	2	1	2	2	Disgust	
1	1	37	1	0	3	3	2	Happy	
3	2	1	35	0	3	1	3	Fear	
3	0	0	1	20	0	0	0	Neutral	
0	1	5	1	0	39	1	1	Rage	
0	0	1	0	0	2	44	0	Astonished	
3	1	2	0	4	0	3	35	Sad	

Fig. 9 Confusion matrix regarding the classification of emotions on the test set of speech signals with a Random Forest of 300 trees

Table 10 Test results of the best overall method (random forest with 300 trees) in the classification of speech data

Accuracy	79.888%
Kappa	0.769
Sensitivity	0.799
Specificity	0.971
AUC	0.965

Model Random Forest (300 trees)

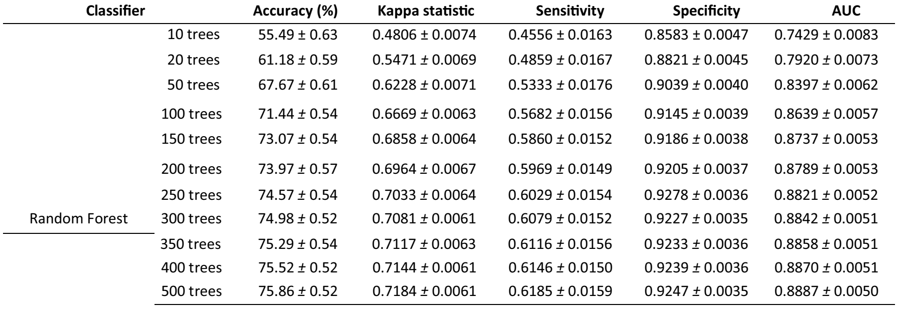
database has few data for each class, thus resulting in low variability. These factors difficult the training of the algorithms, reducing classification performance.

In the test stage with the voice dataset, we also used a Random Forest model with 300 trees. This model achieved the best performance during the training-validation stage. The performance of this model was superior during the classification of the test data, reaching almost 80% accuracy. This result shows that the training, validation and test sets were properly designed. The confusion matrix (Fig. 9) shows that most instances were correctly classified. Of the 358 instances of the validation set, 72 were incorrectly classified. Almost all instances of the calm, neutral and astonished classes were correctly classified. Additionally, contrary to what happened with physiological signals, there was little confusion between neutral class and the others. The highest confusion rate was between rage and happy classes, with 5 of the 48 instances of rage being classified as happy. The class that was most confused with the others was disgust, with 14 instances of the 48 being incorrectly classified outside the origin class.

In classifying emotions from facial expressions, we found a gradual increase in the performance of the algorithms. This increase was noticed as we increased the number of trees that build the Random Forest. From the evaluation of all metrics, we found that from Random Forest with 350 trees we may consider that there was no further increase in the classification performance. Overall, this algorithm achieved good results of accuracy, kappa statistic, specificity and AUC. However, the sensitivity values were a little below the other metrics. This finding indicates that there is some difficulty in identifying the correct class of each instance in this data modality.

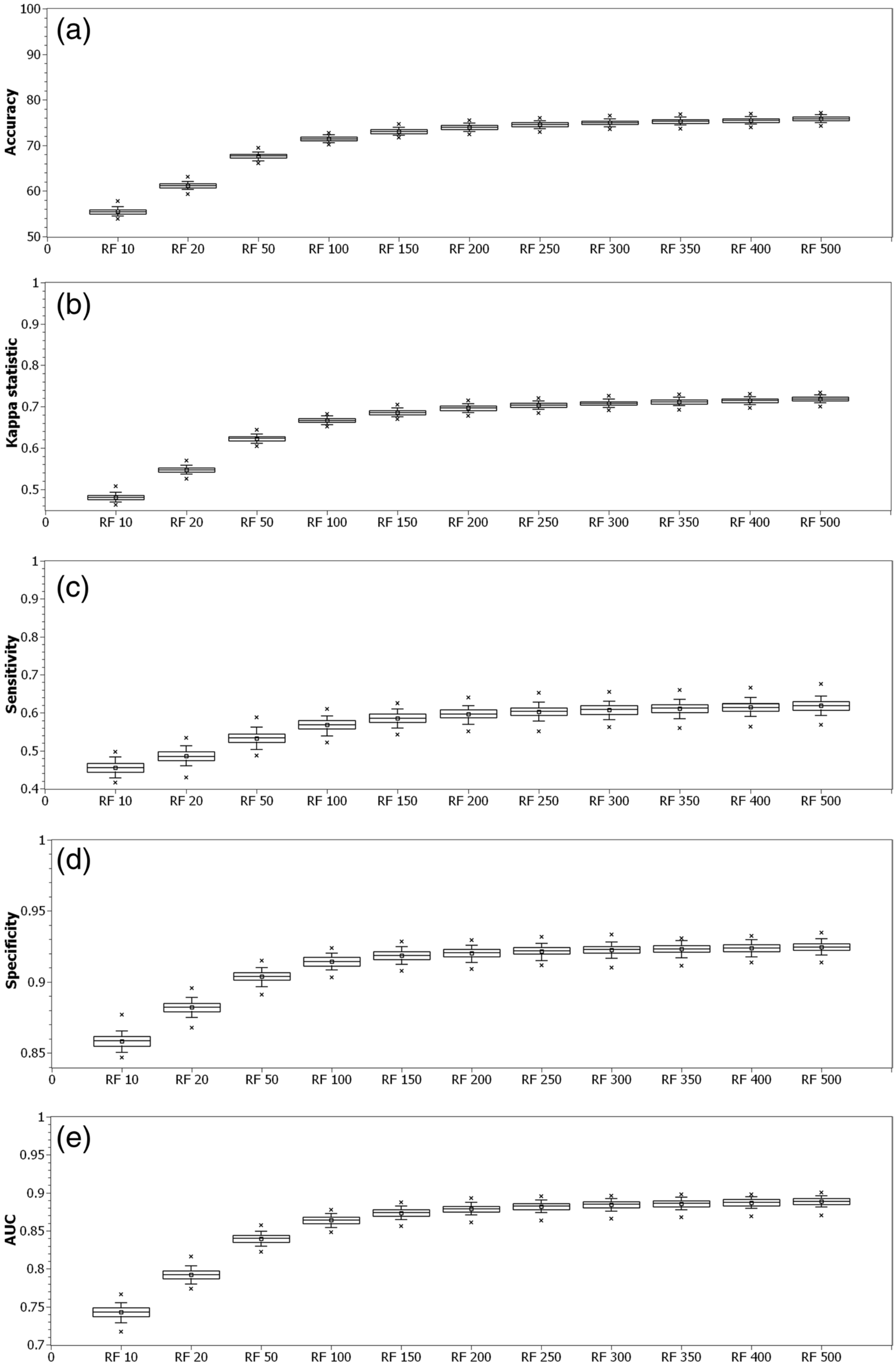
In addition, there is also a good classification performance of Random Forest with 350 trees for the test set. The matrix in Fig. 11 shows that most of the data were correctly classified. Of the 8969 instances in the test set, only 1547 were misclassified. Most errors occurred between neutral, happy and sad classes. There was a low rate of confusion

Table 11 Classification performance for the dataset from facial expressions



Classifier		Accuracy (%)	Kappa statistic	Sensitivity	Specificity	AUC
Random Forest	10 trees	55.49 ± 0.63	0.4806 ± 0.0074	0.4556 ± 0.0163	0.8583 ± 0.0047	0.7429 ± 0.0083
	20 trees	61.18 ± 0.59	0.5471 ± 0.0069	0.4859 ± 0.0167	0.8821 ± 0.0045	0.7920 ± 0.0073
	50 trees	67.67 ± 0.61	0.6228 ± 0.0071	0.5333 ± 0.0176	0.9039 ± 0.0040	0.8397 ± 0.0062
	100 trees	71.44 ± 0.54	0.6669 ± 0.0063	0.5682 ± 0.0156	0.9145 ± 0.0039	0.8639 ± 0.0057
	150 trees	73.07 ± 0.54	0.6858 ± 0.0064	0.5860 ± 0.0152	0.9186 ± 0.0038	0.8737 ± 0.0053
	200 trees	73.97 ± 0.57	0.6964 ± 0.0067	0.5969 ± 0.0149	0.9205 ± 0.0037	0.8789 ± 0.0053
	250 trees	74.57 ± 0.54	0.7033 ± 0.0064	0.6029 ± 0.0154	0.9278 ± 0.0036	0.8821 ± 0.0052
	300 trees	74.98 ± 0.52	0.7081 ± 0.0061	0.6079 ± 0.0152	0.9227 ± 0.0035	0.8842 ± 0.0051
	350 trees	75.29 ± 0.54	0.7117 ± 0.0063	0.6116 ± 0.0156	0.9233 ± 0.0036	0.8858 ± 0.0051
	400 trees	75.52 ± 0.52	0.7144 ± 0.0061	0.6146 ± 0.0150	0.9239 ± 0.0036	0.8870 ± 0.0051
	500 trees	75.86 ± 0.52	0.7184 ± 0.0061	0.6185 ± 0.0159	0.9247 ± 0.0035	0.8887 ± 0.0050

Fig. 10 Emotion classification performance from facial expressions data. Each classifier configuration was assessed based on (**a**) accuracy, (**b**) kappa statistic, (**c**) sensitivity, (**d**) specificity, and (**e**) AUC



between the disgust class and the others. This phenomenon is evidenced by the red quadrants of the matrix. The fact that there are fewer instances of this class in the set may have favored this result. There was still low confusion between the astonished class and the others.

The findings from all data modalities show that the Random Forest algorithm was sufficiently robust to generalize the results obtained in the training-validation steps to the test data. In addition, this model presented low data variability. It points to a good reliability and repeatability of the

Table 12 Test results of the best overall method (random forest with 350 trees) in the classification of facial data

Accuracy	82.752%
Kappa	0.791
Sensitivity	0.828
Specificity	0.962
AUC	0.975

Model Random Forest (350 trees)

performances obtained here. These findings were common to the three databases, even though they came from very different sources. This means that the proposed method is promising for the classification of emotions in multimodal data.

Conclusion

Considering the importance of emotions for the regulation of social interactions and the growing demand for tools that help in their identification, we propose an approach for automatic recognition of emotions. Since the expression of emotions is commonly affected by neurodegenerative pathologies such as dementia, our approach aims to use emotional feedback to support the personalization of therapies for elderly people with dementia. In the therapeutic context, the customization of interventions has been shown to be effective in adapting the conduct to better meet the individual needs of patients. In this approach, we evaluated the performance of a computational methodology in the recognition of emotions in data from different modalities.

Since previous studies prove that data association benefits the identification of emotions, we used data that have a proven relationship with the expression of different human emotions. Therefore, we accessed data from facial expressions, speech signals, and central and peripheral physiological signals. We obtained this data from public databases. Data analysis was performed using artificial intelligence computational tools. In this sense, two approaches were proposed: one for images (facial expressions) and another for signals (speech and physiological).

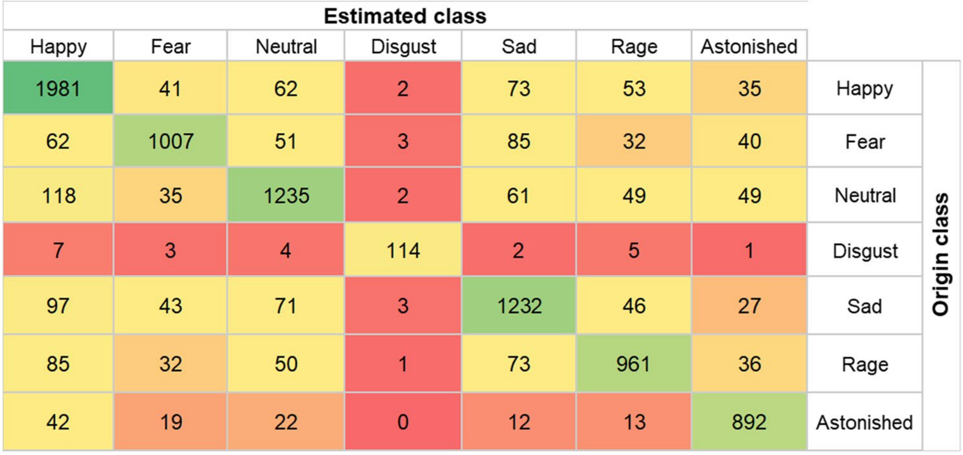
For the investigation of emotion patterns in images we propose a hybrid architecture based on pre-trained deep neural network and Random Forest for feature extraction and classification. The signals were described by statistical attributes and in time and frequency domains. Next, we evaluated the performance of different configurations of ELM, SVM and Random Forest classifiers in differentiating emotions in these signals.

The performance assessment of these methods took place in two stages, one of training-validation and the other of test. In the training-validation stage, we evaluated the different architectures to identify the most suitable setting to deal with each type of data. Then, the chosen model was used in the test stage of each data modality.

In the context of signals, the approach adopted allowed a classification with high accuracy (98.48%), kappa index (0.9817), sensitivity (0.9957), specificity (0.9977) and AUC (0.9999) for the physiological signals. However, the performances with speech signals were much worse, with maximum accuracy of 43% and kappa of 0.35. For both cases Random Forest with 300 trees showed the best performance. The low performance with speech signals can be explained by the high complexity of this data associated with the low number of records per class. Even though the database contains a good amount of signals, there are also many classes of emotions (8). Therefore, the amount of data per class ends up becoming insufficient to enable the identification of the patterns of each class.

In the image analysis, the architecture combining LeNet and Random Forest with 350 trees presented the best classification performance. This algorithm was able to identify the 7 emotions of the problem with an accuracy of 75.29%, kappa of 0.7117, sensitivity of 0.6116, specificity of 0.9233 and AUC of 0.8858.

Fig. 11 Confusion matrix regarding the classification of emotions on the test set of facial expression data with a Random Forest of 350 trees



Estimated class								
Happy	Fear	Neutral	Disgust	Sad	Rage	Astonished		
1981	41	62	2	73	53	35	Happy	Origin class
62	1007	51	3	85	32	40	Fear	
118	35	1235	2	61	49	49	Neutral	
7	3	4	114	2	5	1	Disgust	
97	43	71	3	1232	46	27	Sad	
85	32	50	1	73	961	36	Rage	
42	19	22	0	12	13	892	Astonished	

It is important to highlight that in both types of data (signals and images) the best classifications were achieved by Random Forests with a large number of trees. This indicates that the identification of emotions in these data modalities requires a certain degree of computational complexity. In the test step, we use the architectures with the best performance for each type of data to classify instances that were not used in the training step. From the test we noticed a high generalization capacity of the architectures for the three scenarios. With the physiological data we reached an accuracy of 99.16%. For speech signals, the percentage of correct classification was 79.89%. There was also an accuracy of 82.75% in the identification of emotions in images of facial expressions. In addition, a good performance in differentiating emotions was also perceived in all scenarios. This was explicit in the confusion matrices and in the high values of the sensitivity, specificity and AUC metrics.

The good results in the test stage are encouraging and point to the possibility of adopting the method in the analysis of emotions in multimodal data. These findings are even more interesting due to the large amount and variety of emotions. However, some improvements can be incorporated, such as the use of a greater amount and diversity of data, especially for speech signals. The limited availability of data in the speech signal database made it necessary to use strategies to synthesize data from the original dataset, which may end up attributing a certain degree of redundancy to the instances of each class. In this sense, the incorporation of new data in the architecture training can make the solution more robust and improve its performance in a context of daily use. For the development of a final solution, more in-depth studies are also needed on the computational, time and memory costs associated with processing. Future works may also investigate the performance of architectures in classifying emotions in multimodal signals coming from the same individual.

Author contributions M. S.: Conceptualization, methodology, investigation, writing.

F. F.: Conceptualization, methodology, investigation, writing.

A. T.: Conceptualization, methodology, investigation, writing.

W. P.: Conceptualization, methodology, investigation, writing, scientific supervision.

Funding This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001. The authors are also grateful to the Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco, FACEPE, Brazil, under the code IBPG-0013–1.03/20, and to Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq, Brazil, for the partial financial support of this research.

Data availability Not applicable.

Declarations

Conflict of interest The authors declare no conflicts of interest.

References

Abdullah SMSA, Ameen SYA, Sadeeq MA, Zeebaree S. Multimodal emotion recognition using deep learning. J Appl Sci Technol Trends. 2021;2(02):52–8.

Alarcao SM, Fonseca MJ. Emotions recognition using eeg signals: A survey. IEEE Trans Affect Comput. 2017;10(3):374–93.

Andrade MK, Santana MA, Moreno G, Oliveira I, Santos J, Rodrigues MCA, dos Santos WP: An EEG Brain-Computer Interface to Classify Motor Imagery Signals, 83–98. Springer, Singapore (2020)

Aranha RV, Silva LS, Chaim ML, Nunes FDLDS.: Using Affective Computing to Automatically Adapt Serious Games for Rehabilitation. In: 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS), 55–60 (2017). https://doi.org/10.1109/CBMS.2017.89

Arroyo-Palacios J, Slater M. Dancing with Physio: A Mobile Game with Physiologically Aware Virtual Humans. IEEE Trans Affect Comput. 2016;7(4):326–36. https://doi.org/10.1109/TAFFC.2015.2472013.

Artstein R, Poesio M. Inter-coder agreement for computational linguistics. Comput Linguist. 2008;34(4):555–96. https://doi.org/10.1162/coli.07-034-R2.

Ayata D, Yaslan Y, Kamasak ME. Emotion recognition from multimodal physiological signals for emotion aware healthcare systems. J Med Biol Eng. 2020;40:149–57.

Barbosa VAF, Santana MA, Andrade MKS, Lima RCF, Santos WP. Deep-wavelet neural networks for breast cancer early diagnosis using mammary termographies. In: Das H, Pradhan C, Dey N, editors. Deep Learning for Data Analytics: Foundations, Biomedical Applications, and Challenges. 1st ed. London: Academic Press; 2020.

Barnes DE, Yaffe K. The projected effect of risk factor reduction on alzheimer's disease prevalence. The Lancet Neurology. 2011;10(9):819–28.

Behere R, Arasappa R, Jagannathan A, Varambally S, Venkatasubramanian G, Thirthalli J, Subbakrishna D, Nagendra H, Gangadhar B. Effect of yoga therapy on facial emotion recognition deficits, symptoms and functioning in patients with schizophrenia. Acta Psychiatr Scand. 2011;123(2):147–53.

Blackburn R, Bradshaw T. Music therapy for service users with dementia: A critical review of the literature. J Psychiatr Ment Health Nurs. 2014;21(10):879–88.

Blagus R, Lusa L. Smote for high-dimensional class-imbalanced data. BMC Bioinformatics. 2013;14(1):106.

Bota PJ, Wang C, Fred AL, Da Silva HP. A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals. IEEE Access. 2019;7:140990–1020.

Breiman L. Random Forests. Machine Learning. 2001;45(1):5–32.

Brotons M, Marti P. Music therapy with alzheimer's patients and their family caregivers: a pilot project. J Music Ther. 2003;40(2):138–50.

Calvo RA, D'Mello S. Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Trans Affect Comput. 2010;1(1):18–37.

Cambria E, Das D, Bandyopadhyay S, Feraco A.: Affective computing and sentiment analysis. A practical guide to sentiment analysis, 1–10 (2017)

Castro CB, Costa L, Dias CB, Chen J, Loo R, Sohrabi HR, Brown BM, Martins RN. Multi-domain interventions for dementia prevention–a systematic review. Alzheimer's & Dementia. 2021;17:058289.

Chaturvedi V, Kaur AB, Varshney V, Garg A, Chhabra GS, Kumar M.: Music mood and human emotion recognition based on physiological signals: a systematic review. Multimed Syst, 1–24 (2021)

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.

Cobos FJM, Rodríguez MDMM. A review of psychological intervention in alzheimer s disease. Int J Psychol Psychol Therapy. 2012;12(3):373–88.

Cortes C, Vapnik V. Support-Vector Networks. Machine Learning. 1995;20(3):273–97.

Cruz T, Cruz T, Santos W. Detection and classification of lesions in mammographies using neural networks and morphological wavelets. IEEE Lat Am Trans. 2018;16(3):926–32.

da Saúde OM (2021) Ageing and health. Available in: https://www.who.int/news−room/fact−sheets/detail/ageing−and−health

da Silva CAS, Krohling RA: Classificação de grandes bases de dados utilizando algoritmo de máquina de aprendizado extremo. In: Simpósio Brasileiro de Pesquisa Operacional - SBPO (2016)

de Oliveira E, Jaques PA. Classificação de emoções básicas através de imagens capturadas por webcam. Revista Brasileira De Computação Aplicada. 2013;5(2):40–54.

de Oliveira APS, Santana MA, Andrade MKS, Gomes JC, Rodrigues MC, dos Santos WP. Early diagnosis of parkinson's disease using eeg, machine learning and partial directed coherence. Res Biomed Eng. 2020;36(3):311–31.

de Santana MA, de Lima CL, Torcate AS, Fonseca FS, dos Santos WP. Affective computing in the context of music therapy: a systematic review. Res Soc Dev. 2021;10(15):392101522844–392101522844.

De Silva LC, Miyasato TM, Nakatsu R: Facial emotion recognition using multi-modal information. In: Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications (Cat., 1, 397–401 (1997). IEEE

Delmastro F, Martino FD, Dolciotti C.: Physiological Impact of VibroAcoustic Therapy on Stress and Emotions through Wearable Sensors. In: 2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), 621–626 (2018). https://doi.org/10.1109/PERCOMW.2018.8480170

Deng L. The mnist database of handwritten digit images for machine learning research. IEEE Signal Process Mag. 2012;29(6):141–2.

Deshpande A, Kumar M. Artificial Intelligence for Big Data: Complete Guide to Automating Big Data Solutions Using Artificial Intelligence Techniques. United Kingdom: Packt Publishing; 2018.

Dhuheir M, Albaseer A, Baccour E, Erbad A, Abdallah M, Hamdi M: Emotion recognition for healthcare surveillance systems using neural networks: A survey. In: 2021 International Wireless Communications and Mobile Computing (IWCMC), 681–687 (2021). IEEE

Doma V, Pirouz M. A comparative analysis of machine learning methods for emotion recognition using eeg and peripheral physiological signals. Journal of Big Data. 2020;7(1):1–21.

Dorneles SODSO, Barbosa DNF, Barbosa JLV: Sensibilidade ao contexto na identificação de estados afetivos aplicados à educação: um mapeamento sistemático. Renote 18(1) (2020)

Eaton JW, Bateman D, Hauberg S, Wehbring R.: Gnu octave version 4.0. 0 manual: a high-level interactive language for numerical computations. 2015. URL http://www.gnu.org/software/octave/doc/interpreter 8, 13 (2015)

Espinola CW, Gomes JC, Pereira JMS, dos Santos WP. Detection of major depressive disorder using vocal acoustic analysis and machine learning—an exploratory study. Res Biomed Eng. 2021a;37(1):53–64.

Espinola CW, Gomes JC, Pereira JMS, dos Santos WP. Vocal acoustic analysis and machine learning for the identification of schizophrenia. Res Biomed Eng. 2021b;37(1):33–46.

Ferreira CD, Torro-Alves N.: Reconhecimento de emoções faciais no envelhecimento: uma revisão sistemática. Universitas Psychologica 15(5) (2016)

de Freitas Barbosa VA, Gomes JC, de Santana MA, Jeniffer EdA, de Souza RG, de Souza RE, dos Santos WP.: Heg.IA: An intelligent system to support diagnosis of Covid-19 based on blood tests. Res Biomed Eng 1–18 (2021)

García-Casal JA, Goñi-Imizcoz M, Perea-Bartolomé M, Soto-Pérez F, Smith SJ, Calvo-Simal S, Franco-Martín M. The efficacy of emotion recognition rehabilitation for people with alzheimer's disease. J Alzheimer's Disease. 2017;57(3):937–51.

Gomes JC, Barbosa VADF, Santana MA, Bandeira J, Valença MJS, de Souza RE, Ismael AM, dos Santos WP: IKONOS: An intelligent tool to support diagnosis of covid-19 by texture analysis of x-ray images. Res Biomed Eng 1–14 (2020)

Goodfellow IJ, Erhan D, Carrier PL, Courville A, Mirza M, Hamner F, Li R, Wang X, Athanasakis D, Shawe-Taylor J, Milakov M, Park J, Ionescu R, Popescu M, Grozea C, Bergstra J, Xie J, Romaszko L, Xu B, Chuang Z, Bengio Y.: Challenges in representation learning: A report on three machine learning contests. In: International Conference on Neural Information Processing, 117–124 (2013). Springer

Guess H.: Alzheimer's disease and the impact of music therapy a systematic literature review (2017)

Guetin S, Charras K, Berard A, Arbus C, Berthelon P, Blanc F, Blayac J-P, Bonte F, Bouceffa J-P, Clement S, Ducourneau G, Gzil F, Laeng N, Lecourt E, Ledoux S, Platel H, Thomas-Anterion C, Touchon J, Vrait F-X, Leger J-M. An overview of the use of music therapy in the context of alzheimer's disease: a report of a french expert group. Dementia. 2013;12(5):619–34.

Gupta V, Chopda MD, Pachori RB. Cross-subject emotion recognition using flexible analytic wavelet transform from eeg signals. IEEE Sens J. 2018;19(6):2266–74.

Haan MN, Wallace R. Can dementia be prevented? brain aging in a population-based context. Annu Rev Public Health. 2004;25:1–24.

Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology. 1982;143(1):29–36.

Harms MB, Martin A, Wallace GL. Facial emotion recognition in autism spectrum disorders: a review of behavioral and neuroimaging studies. Neuropsychol Rev. 2010;20(3):290–322.

Hasnul MA, Aziz NAA, Alelyani S, Mohana M, Aziz AA. Electrocardiogram-based emotion recognition systems and their applications in healthcare—a review. Sensors. 2021;21(15):5015.

Huang G-B, Zhu Q-Y, Siew C-K: Extreme learning machine: a new learning scheme of feedforward neural networks. In: 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541), 2: 985–9902 (2004). https://doi.org/10.1109/IJCNN.2004.1380068

Issa D, Demirci MF, Yazici A. Speech emotion recognition with deep convolutional neural networks. Biomed Signal Process Control. 2020;59:101894.

Izard CE. Human Emotions. New York: Springer; 1977.

Izard CE. The Psychology of Emotions. New York, USA: Springer; 1991.

Jackins V, Vimal S, Kaliappan M, Lee MY. Ai-based smart prediction of clinical disease using random forest classifier and naive bayes. J Supercomput. 2021;77(5):5198–219.

Jung Y, Hu J. A K-fold averaging cross-validation procedure. J Nonparametric Stat. 2015;27(2):167–79.

Khalili Z, Moradi MH: Emotion recognition system using brain and peripheral signals: using correlation dimension to improve the results of eeg. In: 2009 International Joint Conference on Neural Networks, 1571–1575 (2009). IEEE

Kim J, André E. Emotion recognition based on physiological changes in music listening. IEEE Trans Pattern Anal Mach Intell. 2008;30(12):2067–83.

Kohler CG, Turner TH, Bilker WB, Brensinger CM, Siegel SJ, Kanes SJ, Gur RE, Gur RC. Facial emotion recognition in schizophrenia: intensity effects and error pattern. Am J Psychiatry. 2003;160(10):1768–74.

Kołakowska A, Szwoch W, Szwoch M. A review of emotion recognition methods based on data acquired via smartphone sensors. Sensors. 2020;20(21):6367.

Kusuma GP, Jonathan APL, Lim A. Emotion recognition on fer-2013 face images using fine-tuned vgg-16. Adv Sci Technol Eng Syst J. 2020;5(6):315–22.

Lawrence K, Campbell R, Skuse D. Age, gender, and puberty influence the development of facial emotion recognition. Front Psychol. 2015;6:761.

LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86(11):2278–324.

Leggieri M, Thaut MH, Fornazzari L, Schweizer TA, Barfett J, Munoz DG, Fischer CE. Music intervention approaches for alzheimer's disease: A review of the literature. Front Neurosci. 2019;13:132.

Lenze SN, Pautsch J, Luby J. Parent–child interaction therapy emotion development: A novel treatment for depression in preschool children. Depress Anxiety. 2011;28(2):153–9.

Lin Y-P, Wang C-H, Wu T-L, Jeng S-K, Chen J-H.: Eeg-based emotion recognition in music listening: A comparison of schemes for multiclass support vector machine. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, 489–492 (2009). IEEE

Livingston G, Sommerlad A, Orgeta V, Costafreda SG, Huntley J, Ames D, Ballard C, Banerjee S, Burns A, Cohen-Mansfield J, Cooper C, Fox N, Gitlin LN, Howard R, Kales HC, Larson EB, Ritchie K, Rockwood K, Sampson EL, Samus Q, Schneider LS, Selbaek G, Teri L, Mukadam N. Dementia prevention, intervention, and care. The Lancet. 2017;390(10113):2673–734.

Livingstone SR, Russo FA. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. PLoS ONE. 2018;13(5):0196391.

Luna-Jiménez C, Griol D, Callejas Z, Kleinlein R, Montero JM, Fernández-Martínez F. Multimodal emotion recognition on ravdess dataset using transfer learning. Sensors. 2021;21(22):7665.

Marcos S, García Peñalvo FJ, Vázquez Ingelmo A.: Emotional ai in healthcare: A pilot architecture proposal to merge emotion recognition tools. In: Ninth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'21), 342–349 (2021)

Marinoiu E, Zanfir M, Olaru V, Sminchisescu C.: 3d human sensing, action and emotion recognition in robot assisted therapy of children with autism. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2158–2167 (2018)

Matziorinis AM, Koelsch S. The promise of music therapy for alzheimer's disease: A review. Ann N Y Acad Sci. 2022;1516(1):11–7.

McIntosh LG, Mannava S, Camalier CR, Folley BS, Albritton A, Konrad PE, Charles D, Park S, Neimat JS. Emotion recognition in early parkinson's disease patients undergoing deep brain stimulation or dopaminergic therapy: a comparison to healthy participants. Front Aging Neurosci. 2015;6:349.

Organização Mundial da Saúde, O.M.S (218) Ageing. Available in: https://www.who.int/health−topics/ageingtab=tab1

Muyuan W, Naiyao Z, Hancheng Z: User-adaptive music emotion recognition. In: Proceedings 7th International Conference on Signal Processing, 2004. Proceedings. ICSP'04. 2004. 2 1352–1355 (2004). IEEE

Ng H-W, Nguyen VD, Vonikakis V, Winkler S: Deep learning for emotion recognition on small datasets using transfer learning. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, 443–449 (2015)

Olanrewaju O, Clare L, Barnes L, Brayne C. A multimodal approach to dementia prevention: a report from the cambridge institute of public health. Alzheimer's & Dementia: Transl Res Clin Interv. 2015;1(3):151–6.

Pal M. Random forest classifier for remote sensing classification. Int J Remote Sens. 2005;26(1):217–22.

Picard RW: Affective Computing. MIT press, ??? (2000)

Platt J.: Fast training of support vector machines using sequential minimal optimization. In: Schoelkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning. MIT Press, ??? (1998). http://research.microsoft.com/texttildelowjplatt/smo.html

Pujol FA, Mora H, Martínez A.: Emotion recognition to improve ehealthcare systems in smart cities. In: Research & Innovation Forum 2019: Technology, Innovation, Education, and Their Social Impact 1, 245–254 (2019). Springer

Reyes BN, Segal SC, Moulson MC. An investigation of the effect of race-based social categorization on adults' recognition of emotion. PLoS ONE. 2018;13(2):0192418.

Russell JA. A Circumplex Model of Affect. J Pers Soc Psychol. 1980;39(6):1161–78. https://doi.org/10.1037/h0077714.

Saganowski S, Dutkowiak A, Dziadek A, Dziezyc M, Komoszy˙nska J, Michalska W, Polak A, Ujma M, Kazienko P.: Emotion recognition using wearables: A systematic literature review-work-in-progress. In: 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), 1–6 (2020). IEEE

Santana MA, Pereira JMS, Silva FLD, Lima NMD, Sousa FND, Arruda GMSD, Lima RDCFD, Silva WWAD, Santos WPD. Breast cancer diagnosis based on mammary thermography and extreme learning machines. Res Biomed Eng. 2018;34:45–53.

Santana MA, Gomes JC, de Souza GM, Suarez A, Torcate AS, Fonseca FS, Moreno GMM, dos Santos WP. Reconhecimento automático de emoçoes a partir de sinais multimodais e inteligência artificial. Anais Do IV Simpósio De Inovação Em Engenharia Biomédica-SABIO. 2020a;2020:43.

Santana MA, Pereira JMS, Lima RCF, Santos WP. Breast lesions classification in frontal thermographic images using intelligent systems and moments of haralick and zernike. In: dos Santos WP, de Santana MA, da Silva WWA, editors. Understanding a Cancer Diagnosis. 1st ed. New York: Nova Science; 2020b. p. 65–80.

dos Santos WP, de Assis FM, de Souza RE, dos Santos Filho PB.: Evaluation of alzheimer's disease by analysis of mr images using objective dialectical classifiers as an alternative to adc maps. In: 2008a 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 5506–5509 (2008a). IEEE

dos Santos WP, de Souza RE, Santos Filho PB, Neto FBL, de Assis FM.: A dialectical approach for classification of DW-MR Alzheimer's images. In: 2008b IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence), 1728–1735 (2008b). IEEE

dos Santos WP, De Assis FM, De Souza RE, Mendes PB, de Souza Monteiro HS, Alves HD.: A dialectical method to classify alzheimer's magnetic resonance images. In: dos Santos, W.P. (ed.) Evolutionary Computation 473. IntechOpen, ??? (2009a)

dos Santos WP, de Assis FM, de Souza RE, dos Santos Filho PB.: Dialectical classification of MR images for the evaluation of Alzheimer's disease. In: Naik, G.R. (ed.) Recent Advances in Biomedical Engineering. IntechOpen, ??? (2009b)

Saxena A, Khanna A, Gupta D. Emotion recognition and detection methods: A comprehensive survey. J Artif Intell Syst. 2020;2(1):53–79.

Schipor O, Pentiuc S, Schipor M. The utilization of feedback and emotion recognition in computer based speech therapy system. Elektronika Ir Elektrotechnika. 2011;109(3):101–4.

Schuller B, Rigoll G, Lang M.: Hidden markov model-based speech emotion recognition. In: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03)., 2, 1 (2003). IEEE

Shafqat S. Alzheimer disease therapeutics: perspectives from the developing world. J Alzheimer's Disease. 2008;15(2):285–7.

Shu Y, Wang S.: Emotion recognition through integrating eeg and peripheral signals. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2871–2875 (2017). IEEE

Silva WWA, Santana MA, Silva Filho AG, Lima SML, Santos WP. Morphological extreme learning machines applied to the detection and classification of mammary lesions. In: Gandhi TK, Bhattacharyya S, De S, Konar D, Dey S, editors. Advanced Machine Vision Paradigms for Medical Image Analysis. London: Elsevier; 2020.

Silva IR, Silva GS, de Souza RG, dos Santos WP, Fagundes RADA.: Model based on deep feature extraction for diagnosis of alzheimer's disease. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–7 (2019). IEEE

Sinha N.: Affective computing and emotion-sensing technology for emotion recognition in mood disorders. Enhanced Telemedicine and e-Health: Advanced IoT Enabled Soft Computing Framework, 337–360 (2021)

Soleymani M, Lichtenauer J, Pun T, Pantic M. A multimodal database for affect recognition and implicit tagging. IEEE Trans Affect Comput. 2011;3(1):42–55.

Sörensen S, Duberstein P, Gill D, Pinquart M. Dementia care: mental health effects, intervention strategies, and clinical implications. The Lancet Neurology. 2006;5(11):961–73.

Sourina O, Liu Y, Nguyen MK. Real-time eeg-based emotion recognition for music therapy. J Multimodal User Interfaces. 2012;5(1):27–35.

de Souza RG, dos Santos Lucas e Silva G, dos Santos WP, de Lima ME, Initiative, A.D.N 2021 Computer-aided diagnosis of Alzheimer's disease by MRI analysis and evolutionary computing. Res Biomed Eng 37, 455–483

Van der Steen JT, Smaling HJ, Van der Wouden JC, Bruinsma MS, Scholten RJ, Vink AC.: Music-based therapeutic interventions for people with dementia. Cochrane Database of Systematic Reviews (7) (2018)

Vijayakumar S, Flynn R, Murray N.: A comparative study of machine learning techniques for emotion recognition from peripheral physiological signals. In: 2020 31st Irish Signals and Systems Conference (ISSC), 1–6 (2020). IEEE

Wei W, Jia Q, Feng Y, Chen G.: Emotion recognition based on weighted fusion strategy of multichannel physiological signals. Comput Intell Neurosci 2018 (2018)

Wiem MBH, Lachiri Z.: Emotion classification in arousal valence model using mahnob-hci database. Int J Adv Comput Sci App 8(3) (2017)

Witten IH, Frank E. Data Mining: Pratical Machine Learning Tools and Technique. San Francisco, CA, USA: Morgan Kaufmann Publishers; 2005.

Yang Y-H, Lin Y-C, Chen H.: Personalized music emotion recognition. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 748–749 (2009)

Zeng J, Roussis PC, Mohammed AS, Maraveas C, Fatemi SA, Armaghani DJ, Asteris PG.: Prediction of peak particle velocity caused by blasting through the combinations of boosted-chaid and svm models with various kernels. Applied Sciences 11(8) (2021). 10.3390/ app11083705

Zhang X-X, Tian Y, Wang Z-T, Ma Y-H, Tan L, Yu J-T. The epidemiology of alzheimer's disease modifiable risk factors and prevention. J Prev Alzheimer's Disease. 2021;8:313–21.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.