IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 September 2024, accepted 24 October 2024, date of publication 29 October 2024, date of current version 7 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3487932

RESEARCH ARTICLE

NeuroSense: A Novel EEG Dataset Utilizing Low-Cost, Sparse Electrode Devices for Emotion Exploration

TOMMASO COLAFIGLIO[1,2], ANGELA LOMBARDI[2], PAOLO SORINO[2], ELVIRA BRATTICO[3,4], DOMENICO LOFÙ[2], (Member, IEEE), DANILO DANESE[2], EUGENIO DI SCIASCIO[2], TOMMASO DI NOIA[2], (Member, IEEE), AND FEDELUCIO NARDUCCI[2]

[1]Department of Computer, Control, and Management Engineering (DIAG), Sapienza University of Rome, 00185 Rome, Italy
[2]Department of Electrical and Information Engineering, Politecnico di Bari, 70125 Bari, Italy
[3]Center for Music in the Brain, Department of Clinical Medicine, Aarhus University, Aarhus C, 8000 Aarhus, Denmark
[4]Department of Education, Psychology, Communication Sciences, Università degli Studi di Bari, 70122 Bari, Italy

Corresponding author: Angela Lombardi (angela.lombardi@poliba.it)

This work was supported by the "SECURE SAFE APULIA," itaLian system wIde Frailty nEtwork (LIFE), Casa delle Tecnologie Emergenti di Matera (CTEMT), rete Integrata meDiterranea per l'osservazione ed Elaborazione di percorsi di Nutrizione (IDENTITA), and OncologIA under Grant PNRR-MAD-2022-12376656. This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Local Ethical Committee of the University of Bari.

ABSTRACT Emotion recognition is crucial in affective computing, aiming to bridge the gap between human emotional states and computer understanding. This study presents NeuroSense, a novel electroencephalography (EEG) dataset utilizing low-cost, sparse electrode devices for emotion exploration. Our dataset comprises EEG signals collected with the portable 4-electrodes device Muse 2 from 30 participants who, thanks to a neurofeedback setting, watch 40 music videos and assess their emotional responses. These assessments use standardized scales gauging arousal, valence, and dominance. Additionally, participants rate their liking for and familiarity with the videos. We develop a comprehensive preprocessing pipeline and employ machine learning algorithms to translate EEG data into meaningful insights about emotional states. We verify the performance of machine learning (ML) models using the NeuroSense dataset. Despite utilizing just 4 electrodes, our models achieve an average accuracy ranging from 75% to 80% across the four quadrants of the dimensional model of emotions. We perform statistical analyses to assess the reliability of the self-reported labels and the classification performance for each participant, identifying potential discrepancies and their implications. We also compare our results with those obtained using other public EEG datasets, highlighting the advantages and limitations of sparse electrode setups in emotion recognition. Our results demonstrate the potential of low-cost EEG devices in emotion recognition, highlighting the effectiveness of ML models in capturing the dynamic nature of emotions. The NeuroSense dataset is publicly available, inviting further research and application in human-computer interaction, mental health monitoring, and beyond.

INDEX TERMS Emotion recognition, EEG dataset, low-cost EEG devices, machine learning, human-computer interaction, Russell's model.

I. INTRODUCTION

Emotion recognition is emerging as a pivotal area in affective computing, aiming to bridge the gap between human emotional states and computer recognition. This interdisciplinary field leverages electroencephalogram (EEG) devices to capture the nuanced electrical activities of the brain that correspond to different emotional states. The motivation for employing EEG in emotion recognition stems from its potential to provide direct, physiological indicators of

The associate editor coordinating the review of this manuscript and approving it for publication was Orazio Gambino.

emotional experiences, bypassing the subjective nature of self-reports and the ambiguity of behavioral observations [1], [2], [3].

Recent advancements in the field have highlighted the effectiveness of EEG in capturing the dynamic nature of emotions [4]. However, the widespread adoption of EEG-based emotion recognition remains limited by the cost and complexity of traditional EEG systems, which typically involve multiple electrodes and specialized equipment. This has spurred interest in exploring low-cost, sparse electrode EEG devices as a viable alternative. These devices hold the promise of making emotion recognition technologies more accessible for research and applications in human-computer interaction (HCI), mental health monitoring, and beyond [5], [6], [7], [8], [9].

Machine learning (ML) plays a crucial role in translating EEG data into meaningful insights about emotional states. ML algorithms enable the automatic detection and classification of complex patterns within EEG signals, facilitating the identification of specific emotional states [10], [11], [12]. The impact of ML in this domain is profound, as it offers the potential to enhance the emotional intelligence of HCI systems, improve the accuracy of mental health assessments, and support the development of personalized user experiences [13], [14], [15].

However, the variability of individual brain patterns poses significant challenges for ML models. Each individual's brain activity is unique and influenced by physiological factors, personal experiences, and cognitive processes. This variability necessitates sophisticated validation schemes to ensure that ML models are accurate and capable of generalizing across different subjects [16]. The leave-one-subject-out (LOSO) validation scheme is a critical method for addressing this challenge [17]. The LOSO validation scheme tests the model's ability to generalize by training it on data from all but one subject and then testing it on the unseen data of the left-out subject. This process is repeated for each subject in the dataset, ensuring that the model is validated across all possible subject-specific scenarios. Such an approach is essential in emotion recognition research, where the goal is to develop models that can accurately interpret the emotional states of any individual, not just those from whom the training data is sourced [18].

This work presents a novel EEG dataset, NeuroSense, as a pioneering contribution to emotion exploration. This is the first open-source EEG dataset designed explicitly for this purpose, using only four electrodes. By leveraging the potential of low-cost, sparse electrode devices, we advance the accessibility and inclusivity of emotion recognition technologies. Utilizing a validated protocol and state-of-the-art ML techniques, we demonstrate the feasibility of accurately categorizing emotions across Russell's affective space with our dataset [19]. The open-source nature of NeuroSense democratizes access to cutting-edge research tools and fosters a collaborative environment for innovation in the field.

- We verify the performance of ML models using the *NeuroSense* dataset, demonstrating that high classification accuracy can be achieved with only four electrodes.
- We perform statistical analyses to assess the reliability of the self-reported labels and the classification performance for each participant, identifying potential discrepancies and their implications.
- We compare our results with those obtained using other public EEG datasets, highlighting the advantages and limitations of sparse electrode setups in emotion recognition.

The paper is structured as follows: Section II provides an overview of the most widely used EEG databases. Section III describes the procedure for selecting stimuli. Section IV details the experimental setup. Section V presents the system architecture. Section VI outlines the results, while Section VII discusses the findings. Finally, Section VIII concludes the work.

II. RELATED WORK

Different open-source EEG emotion databases have been explicitly devised for EEG emotion recognition. The following section presents a concise overview of currently available emotion databases. The main characteristics of the databases are also briefly listed in Table 1.

The DEAP (Database for Emotion Analysis using Physiological Signals) dataset [20] is a comprehensive repository designed to analyze human emotions through physiological signals. Developed primarily for emotion-related research, the DEAP dataset encompasses multimodal data, incorporating electroencephalogram (EEG), peripheral physiological signals (such as electrocardiogram and galvanic skin response), facial electromyogram (EMG), and subjective ratings. The dataset features recordings of emotional responses elicited by audio-visual stimuli, with subjects exposed to music videos. Notably, the DEAP dataset comprises data from 32 participants, contributing to its richness and diversity in capturing emotional states. In the experiment, each participant views 40 one-minute music video clips. Participants self-evaluate their subjective emotional experiences during the induction experiments, utilizing assessment scales encompassing four emotional dimensions: arousal, valence, dominance, and liking. Data acquisition involves using a Biosemi ActiveTwo system to record EEG and peripheral physiological signals. The dataset encompasses 40 channels, of which the initial 32 capture EEG signals with a sampling rate of 512 Hz and the remaining eight record peripheral signals.

The MAHNOB (Multimodal Affective Human-Nonverbal Behavioral) dataset [21] is designed for research in affective computing and HCI. It comprises multimodal data collected with 32 electrodes from 30 subjects (17 women and 13 men, age range: 19-40), including video recordings, physiological signals, and subjective annotations. The dataset also includes recordings of participants' facial expressions and nonverbal

TABLE 1. Summary of published EEG databases for emotion recognition/classification. Note that "ch" stands for "channel".

Dataset	EEG Electrodes	Participants	Stimuli	Emotion Labels	Devices
DEAP [20]	32	32	Music videos	Arousal, Valence, Dominance, Liking	Biosemi ActiveTwo
MAHNOB [21]	32	30	Movie clips	Disgust, Amusement, Fear, Sadness, Joy	32ch EEG + ECG + EDA
SEED [2]	62	15	Film clips	Positive, Negative, Neutral	62ch EEG
MPED [22]	62	23	Emotion videos	Joy, Anger, Fear, Disgust, Sadness, Neutrality	62ch EEG
DREAMER [6]	14	23	Audio-visual clips	Valence, Arousal, Dominance	Emotiv EPOC

behaviors captured using six cameras. The physiological signals comprise electroencephalogram (EEG), electrocardiogram (ECG), and electrodermal activity (EDA), providing insights into participants' physiological responses. The participants viewed 20 different emotional video clips selected from movies and video websites. The clips stimulate five emotions: disgust, amusement, fear, sadness, and joy. The duration of the videos ranges from 35 to 117 seconds. Annotations are provided for facial expressions, physiological signals, and subjective emotional experiences. After watching each video clip, participants provide subjective assessments or self-reports about their emotional experiences.

The SJTU Emotion EEG Dataset (SEED) [2] is a collection of EEG datasets that contains data with 62 electrodes from 15 subjects (7 men and 8 women) recorded while they watch carefully selected 15 film clips designed to induce different types of emotions, including positive, negative, and neutral emotions. The participants are given a 5-second prompt before each video. Following this, they complete a 45-second self-assessment and then have a 15-second rest period. The SEED-IV dataset is an evolution of the original SEED dataset, providing EEG signals and eye movement features. The film clips are chosen to induce happiness, sadness, fear, and neutral emotions.

The Multi-Modal Physiological Emotion Database (MPED) [22] is an extensive dataset developed for studying human emotions. It gathers four physiological signals: electroencephalogram (EEG) with 62 channels, galvanic skin response, respiration, and electrocardiogram (ECG). The purpose of creating this dataset is to improve the extraction of distinguishing features and the effectiveness of useful sequences, which in turn assists in recognizing discrete emotions (joy, funny, anger, fear, disgust, sadness, and neutrality). The MPED dataset comprises data from 23 participants (10 men and 13 women), each having one session. The data are gathered while the participants are exposed to 28 videos from an emotion-elicitation material database.

The DREAMER Database [6] for Emotion Analysis using Physiological and Electroencephalographic Responses is a collection designed to analyze emotions. It contains EEG (14 channels) and ECG recordings taken while participants are exposed to 18 audio-visual clips intended to provoke emotional responses. Following each stimulus, this dataset encompasses data from 23 individuals who also self-evaluate their emotional states, including valence, arousal, and dominance. The data are gathered using consumer-grade, mobile, and wireless devices that are affordable and readily available, such as the Emotiv EPOC EEG[1] headset and the Shimmer ECG sensor.[2]

This short review of the existing databases shows that the majority contain EEG data obtained in a time-consuming and costly way from multiple channels. In contrast, only one is gathered using easy-to-wear, mobile, wireless, and economical devices. Moreover, most reviewed EEG emotion datasets include fewer participants than NeuroSense. For example, the SEED dataset comprises only 15 participants, and DREAMER includes 23 participants, even though they use more electrodes and higher sampling rates. In contrast, NeuroSense utilizes a portable 4-electrode device and includes 30 participants, showcasing the potential of such setups in emotion recognition tasks.

III. STIMULI SELECTION

In this work, we adopt the same approach validated in the DEAP protocol to elicit emotional responses from the users. The DEAP protocol was selected for this study due to its widespread recognition and standardization in emotion analysis using physiological signals. By employing a validated and reliable protocol such as DEAP, we ensure the robustness and comparability of the results. The protocol provides comprehensive multidimensional evaluations of emotional states, including arousal, valence, and dominance, which align with our research objectives. Additionally, the DEAP protocol facilitates the use of audio-visual stimuli to evoke authentic emotional responses, thereby enhancing the ecological validity of our findings.

The stimuli selection process is divided into several macro steps organized into key phases, including video stimuli selection, detection of one-minute highlights, relevance vector machine application, and online subjective annotation. The macro steps are briefly summarized below and further detailed in the following subsections. An overview is provided in Figure 1.

1) Video clip selection: a total of 120 videos are selected in two modalities: 60 videos through affective tags research selection using a comprehensive list of emotional keywords expanded from Parrott's work [23], and 60 videos through a manual process. This involves leveraging the Last.fm[3] database to identify relevant

[1] https://www.emotiv.com/
[2] https://shimmersensing.com/
[3] www.last.fm

musical content tags, selecting songs frequently tagged with each emotion, and manually curating songs to represent each of Russel's emotional quadrants [19].
2) Detection of one-minute highlights: following dataset creation, one-minute video segments with the highest emotional content are identified using the DEAP protocol and the Relevance Vector Machine (RVM) [24] model to compute arousal and valence scores.
3) Video and audio features are also extracted.
4) Relevance Vector Machine: the RVM [24] is trained using features extracted from 21 annotated movies to predict valence and arousal scores for each video. Segments with higher emotional highlight scores are selected, with a manual override option for segments characteristic of the song and expected to elicit emotional responses.
5) Online subjective annotation: the final selection of 40 test video clips is based on a web-based subjective emotion assessment by volunteers who rate music videos on valence, arousal, and dominance. Videos are chosen based on the intensity of emotion they elicit.

A. VIDEO-CLIP SELECTION

First, 120 videos are chosen in two different modalities. Specifically, i) 60 videos are selected using an affective tags research selection; ii) a manual process selects another 60 videos. Related to the first step, the entire pipeline is highlighted below:

- Initially, a comprehensive list of emotional keywords is compiled based on Parrott's work [23], which is then expanded to include variations and synonyms, resulting in 304 unique keywords.
- Subsequently, leveraging the Last.fm database, relevant tags are identified for each keyword, facilitating the association of keywords with musical content.
- Following this, for each emotional tag, an automated process is employed to select the ten songs most frequently tagged with that emotion, thereby creating a dataset consisting of 1084 songs.
- Additionally, in alignment with Russel's model, fifteen songs are manually curated to represent each of the four emotional quadrants. Selection criteria encompass the accurate portrayal of emotional content, availability of a corresponding music video, and appropriateness for the target audience, predominantly European or North American students.

The second primary phase involves manually selecting an additional 60 videos. Specifically, 15 videos are chosen to represent each quadrant of Russell's model derived from the intersection of valence and arousal values as depicted in Figure 2.

Consequently, the dataset comprises 120 emotional videos. Panel A in Figure 1 summarizes the whole pipeline.

B. DETECTION OF ONE-MINUTE HIGHLIGHTS

Following the dataset's creation, the subsequent phase involves the sub-selection of one-minute video segments to identify the segments with the highest emotional content. Following the DEAP protocol, the RVM model is employed to compute arousal and valence scores for each film segment, as outlined by Soleymani et al. [24] and described in Section III-C. Subsequently, video and audio features are extracted from the 120 videos using a designated pipeline.

The video features extraction process is summarized below:

- the video content is encoded into MPEG-1 format, as described by Symes [25], to facilitate the extraction of motion vectors and I-frames for subsequent feature extraction.
- Shot segmentation is performed on the video stream using a method proposed by Kelm et al. [26], enabling the selection of various shot domains within the video.
- Lighting and color variance, recognized as significant elements in eliciting emotion, are extracted, considering factors such as the movie director's vision. Specifically:
  - lighting keys are calculated by multiplying the average value V in the HSV color space [27] by the standard deviation of V for each frame.
  - Color variance is computed in the CIE LUV color space [28] by determining the coefficients of the covariance matrix for L (luminosity), U (color coordinate of green-red), and V (color coordinate of blue-yellow) for each frame.
- Video rhythm and emotional impact, as emphasized in the works of Hanjalic and Xu [29], are characterized by extracting the average shot change rate and shot length variance.
- Dynamic scenes and object movements across successive frames are identified as effective contributors to excitement. This parameter is quantified by delineating the motion component, computed as the aggregate motion within successive frames by summing the magnitudes of motion vectors across all B- and P-frames, as described by Richardson [30].
- The proportion of colors is calculated through a 20-bin color histogram of hue and lightness values in the HSV color space for each I-frame, subsequently averaged across all frames.
- To determine the median lightness of a frame, the median of the L value in the HSL color space is computed.

Panel B in Figure1 summarizes the whole pipeline.

C. SELECTION OF AUDIO FEATURES

The audio features extraction process is detailed below and summarized in Panel C in Figure 1:

- the audio channels from the video dataset are extracted and converted into mono MPEG-3 format with a sampling rate of 44.1 kHz.

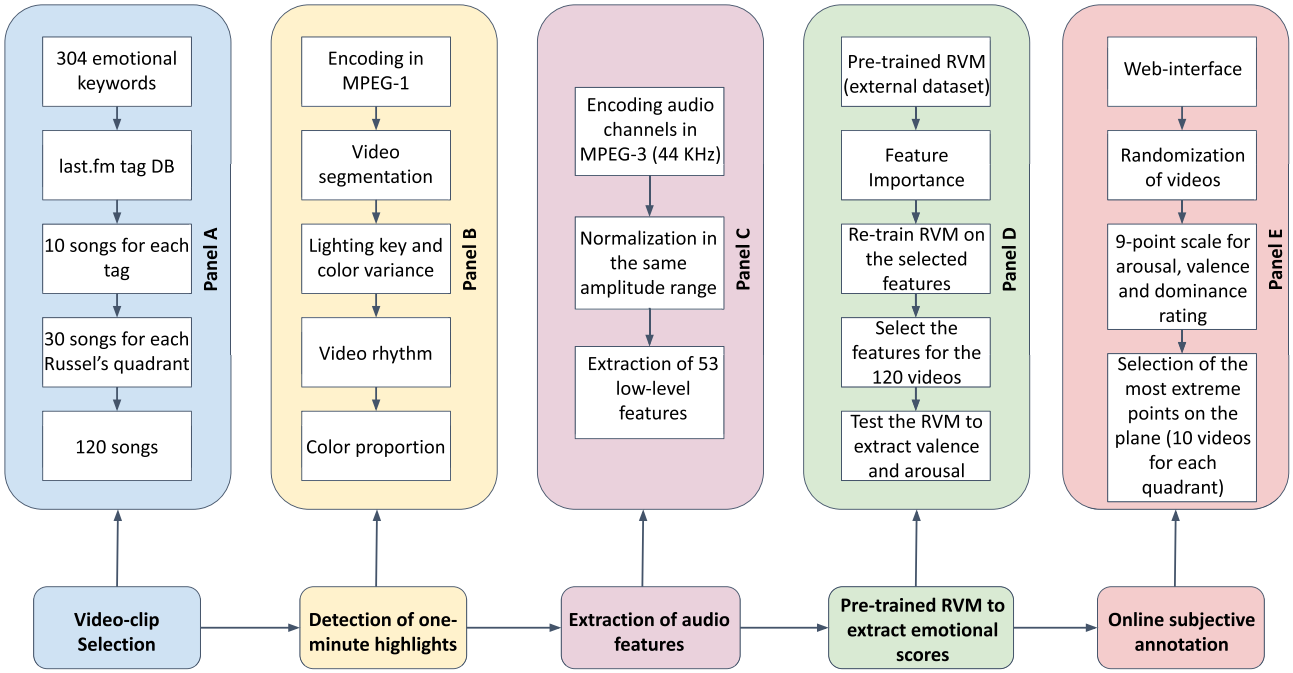


FIGURE 1. Overview of the stimuli selection process. The procedure includes five key steps: (1) video clip selection through affective tags and manual curation to represent emotional quadrants; (2) identification of one-minute highlights using the DEAP protocol; (3) extraction of video and audio features; (4) application of the RVM model to predict valence and arousal scores; and (5) final video selection through online subjective annotation based on volunteer ratings of valence, arousal, and dominance.

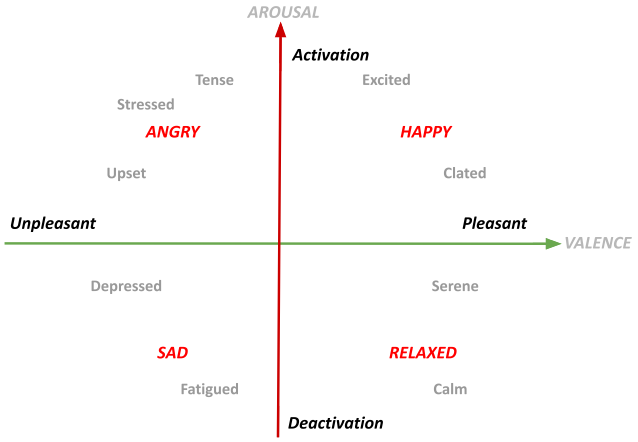


FIGURE 2. Representation of Russel's circumplex model [19].

- All audio signals are normalized to the same amplitude range before further processing.
- A total of 53 low-level audio features, including Mel-Frequency Cepstral Coefficients (MFCC), energy formants, time-frequency characteristics, pitch, zero crossing rate, and silence ratio, are extracted from each audio signal, as outlined by Gorjian et al. [31].

D. RELEVANCE VECTOR MACHINE

To predict the valence and arousal scores for each video segment among the selected 120 videos, the Relevance Vector Machine (RVM) [32] is trained using all shots from 21 annotated movies within the dataset presented in [24]. Additionally, the RVM identifies the importance of features. It is thus employed in the DEAP protocol to select a subset of features from the entire set of extracted features. The final prediction pipeline is delineated below:

- the music videos are segmented into one-minute intervals with a 55-second overlap.
- Content features are extracted from each segment.
- Prediction scores for arousal and valence are obtained from the trained RVM.
- A final score is computed for each segment using the equation: $e_i = sqrt(a_i^2 + v_i^2)$, where a_i and v_i represent centered arousal and valence scores.

The graphical representation shown in Panel D of Figure 1 summarizes the steps involved in the pipeline. Segments with higher emotional highlight scores are then selected. A manual override of the affective highlight detection is performed for some clips. This manual intervention is applied to segments deemed characteristic of the song, recognized by the public and expected to elicit emotional responses. Following this iterative process, a collection of 120 one-minute videos featuring high emotional content is obtained.

E. ONLINE SUBJECTIVE ANNOTATION

This section describes the process for selecting the final 40 test video clips. The selection methodology involves a web-based subjective emotion assessment wherein

participants viewed music videos and provided ratings on a 9-point scale for valence, arousal, and dominance. Key aspects of this process include:

1) participants used a web interface for rating videos.
2) They were allowed to watch as many videos as they wanted and end the rating process at any time.
3) The order of video clips was randomized.
4) Participants did not see the same video twice.
5) All 120 videos received ratings from 14 volunteers.

For each video, a score representing the intensity of elicited emotion was computed based on the ratings provided by volunteers. This score was calculated by dividing the mean rating (μ_x) by the standard deviation (σ_x), yielding $\frac{\mu_x}{\sigma_x}$. Subsequently, for each quadrant where the videos were positioned, the videos situated at the extreme corners of the quadrant were selected. Through this iterative process, 40 videos were ultimately selected. We defined the corresponding labels of these videos as external labels. Panel E of Figure 1 illustrates the main steps.

IV. EXPERIMENTAL SETUP

A. MATERIALS AND SETUP

The experiments were conducted in laboratory settings with controlled illumination in the Department of Electrical and Information Engineering premises at the Polytechnic University of Bari. EEG signals were recorded using a Muse 2 device,[4] equipped with four electrodes, connected to a dedicated recording computer, a MacBook Pro (Retina, 15-inch, Mid 2015).

Stimuli presentation was facilitated through a dedicated PC monitor (HP), and the software for stimuli presentation was developed using Max/MSP.[5] This system featured a graphical user interface (GUI) with corresponding numerical identifiers linked to the videos to be played. Upon the operator pressing the button associated with a specific video, a marker was transmitted to a server using the LabStreamingLayer (LSL) protocol.[6] Labrecorder software[7] was employed to record EEG signals and synchronize markers.

The music videos were displayed on a 17-inch screen with a 1280 × 1024 pixel resolution. To minimize eye movements, the videos were presented at a reduced resolution of 800 × 600 pixels, filling approximately two-thirds of the screen. Participants were seated approximately one meter away from the screen. Yamaha-HS 8 speakers were used for audio playback, with the volume set relatively loud. However, participants were consulted beforehand regarding their comfort level with the volume, and adjustments were made accordingly.

[4]https://choosemuse.com/products/muse-2
[5]https://cycling74.com/products/max
[6]https://labstreaminglayer.org
[7]https://github.com/labstreaminglayer/App-LabRecorder

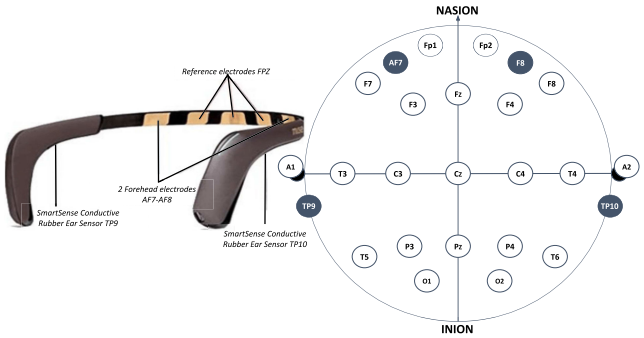


FIGURE 3. Muse 2 EEG Device with electrode placement according to the 10-20 System. The figure illustrates the specific positions of the four electrodes (AF7, AF8, TP9, and TP10) used in the experiment. The Fpz electrode serves as the reference point, and the device provides a non-invasive, portable solution for EEG data acquisition optimized for comfort and ease of use during the emotional response experiment.

B. EEG ACQUISITION DEVICE

This study employs the Muse 2 EEG device for signal acquisition. The Muse 2 EEG device is a cutting-edge and adaptable headset designed for capturing EEG signals. Equipped with four dry sensors, it detects brain activity with high signal quality while minimizing the need for scalp preparation associated with traditional gel electrodes. Its ergonomic design provides a comfortable fit on the user's head, facilitated by adjustable headbands and soft earpieces, enhancing the user experience during prolonged use. Powered by a rechargeable battery and Bluetooth connectivity, the device offers exceptional portability and versatility, making it suitable for various environments and applications.

The electrodes are strategically positioned at Fpz, AF7, AF8, TP9, and TP10 locations according to the international 10-20 system. The Fpz electrode serves as a reference, and the sampling frequency is 256 Hz. Figure 3 illustrates the Muse device and its relative positions based on the international 10-20 system [33].

C. EXPERIMENTAL PROTOCOL

A total of 30 healthy participants (50% female), aged between 19 and 30 years (mean age 23.5), were recruited for the experiment. Participants for this study were recruited through university-wide advertisements and social media platforms to ensure a diverse and representative sample. All participants provided informed consent before participating, and no monetary compensation was offered. Ethical permission was obtained from the Local Ethical Committee of the University of Bari. The participants received comprehensive information regarding the experimental protocol, including detailed explanations of the various self-assessment scales. Once the protocol was thoroughly explained, the EEG device was positioned, and signal quality was meticulously assessed using the MuseLSL library for EEG signal window [34].

The experiment began with the presentation of 40 videos across 40 separate trials, each lasting 1 minute. Each trial followed a standardized sequence of events comprising:

- a 2-second display indicating the current trial number to inform participants of their progress.
- A 5-second recording of baseline activity, represented by a fixation cross.
- Presentation of a 1-minute music video.
- Subsequent rating of the participants' arousal, valence, liking, and dominance levels.

Upon completion of 20 trials, participants were given a brief break. During this intermission, the facilitator assessed signal quality and electrode placement to ensure accuracy before instructing participants to proceed with the second part of the test.

D. PARTICIPANT SELF-ASSESSMENT

Participants were tasked with assessing their levels of arousal, valence, liking, and dominance after each trial. To facilitate this process, Self-Assessment Manikins (SAMs) [35] were employed, providing visual representations of the scales. For example, the liking scale featured symbols of thumbs-down and thumbs-up, positioned at the center of the screen with numbers 1 to 9 displayed below them. Figure 4 illustrates an example of the various scales utilized in the questionnaire. Participants indicated their self-assessment levels horizontally, moving the mouse beneath the numbers and clicking on their chosen level. They were informed of the flexibility to click anywhere below or between the numbers, effectively creating a continuous scale for self-assessment.

The valence scale ranged from unhappy or sad to happy or joyful, allowing participants to rate the emotional tone of their experience. The arousal scale spanned from calm or bored to stimulated or excited, enabling participants to gauge their level of stimulation or excitement. The dominance scale ranged from submissive (indicating a lack of control) to dominant (indicating a sense of control or empowerment).

A fourth scale was also included to assess participants' personal liking for the video. It was essential to distinguish this scale from the valence scale, as it measured preferences rather than emotional responses. For instance, participants could like videos that elicited feelings of sadness or anger.

Following the experiment, participants were requested to rate their familiarity with each song on a scale from 1 (''Had never heard it before the experiment'') to 5 (''Knew the song very well'').

The questionnaire structure is outlined below:

- User ID (i.e., the numerical code representing each participant).
- Video ID (i.e., the numerical identifier associated with each of the 40 videos).
- Valence score (Likert scale 1-9).
- Arousal score (Likert scale 1-9).
- Dominance score (Likert scale 1-9).
- Liking (3 options: liked, neutral, dislike).
- Familiarity with the song heard (Likert scale 1-5, ranging from ''Never heard of it before the experiment'' to ''I know it very well'').

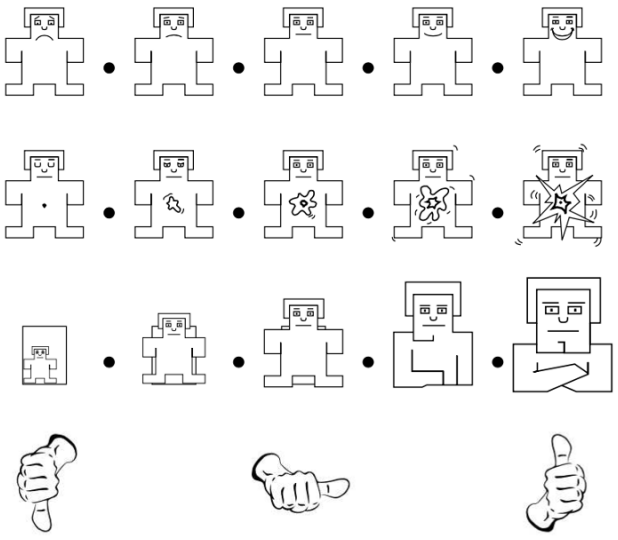


FIGURE 4. Self-Assessment Manikins (SAMs) used in the experiment. The figures show the SAMs used by participants to rate their emotional responses across three dimensions: valence (ranging from unpleasant to pleasant), arousal (ranging from calm to excited), and dominance (ranging from submissive to dominant). Additionally, participants rated their liking for each video using a three-option scale: liked, neutral, or dislike.

V. SYSTEM ARCHITECTURE

We developed an ML framework to test the effectiveness of ML on the collected EEG data.

The architecture of our ML framework, as illustrated in Figure 5, is composed of six principal components, each playing a crucial role in the processing and analysis of EEG data for emotion recognition:

- acquisition module: this component collects EEG data, utilizing the Muse device as the primary source.
- Data creation: this module systematically collects and organizes EEG data from each participant, ensuring precise alignment with the timestamps of baseline periods and stimuli presentations.
- Preprocessing module: upon receiving EEG signals from the Muse device, this module applies a series of processing techniques to refine the data for analysis. The preprocessing pipeline enhances signal quality by eliminating noise and artefacts commonly associated with EEG data. The steps involved are as follows:
  1) creation of an EEG epochs structure, utilizing the MNE framework[8] for efficient data segmentation and organization.
  2) Application of a Finite Impulse Response (FIR) filter with a cutoff frequency range of $1 - 45$ Hz with fir window='hamming' to isolate the relevant frequency bands for emotion recognition.
  3) Implementation of ringing artifact reduction techniques, specifically devised to remove non-brain

[8]https://mne.tools/stable/index.html

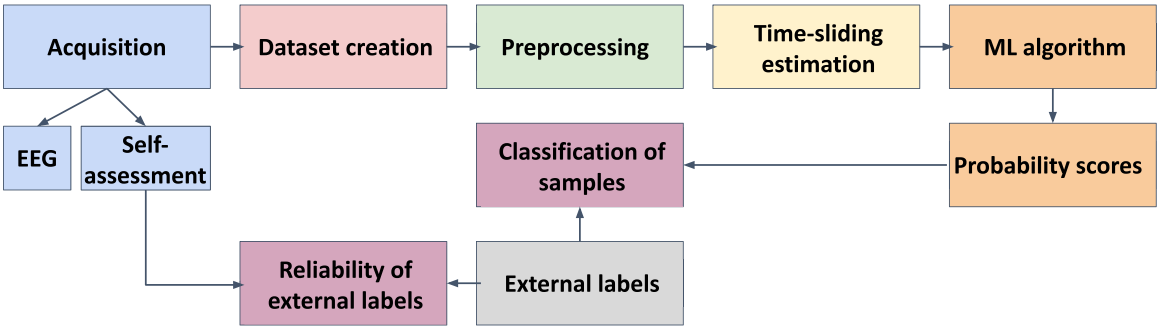


FIGURE 5. Illustration of the workflow representing all the key steps in our proposed system for emotion recognition using EEG data. The figure outlines the following stages: data acquisition via the Muse 2 EEG device, data creation and organization, preprocessing of EEG signals (including noise reduction, filtering, and epoch segmentation), time-sliding estimation for identifying ROIs, machine learning-based classification of emotional states, and statistical analyses for correlation and participant reliability assessment.

artifacts without compromising the integrity of the EEG signal.

4) Division of the processed EEG data into sub-epochs of 5 seconds each, facilitating a granular analysis of emotional responses over time.

- Time-sliding estimation: a dynamic strategy to pinpoint the most informative regions of interest (ROI) within the EEG signals as they evolve over time.
- ML module: this module is at the heart of our framework, where different ML algorithms are employed to interpret the preprocessed EEG signals. Leveraging state-of-the-art models, this component predicts the emotional state represented by the input data.
- Statistical analyses involve two main steps. First, the Pearson correlation between self-assessment scores and external labels for arousal and valence is calculated to evaluate label reliability. Second, the self-assessment and ML probability scores are analyzed to identify non-trustworthy or poorly performing participants.

A description of the operations performed is provided in the Algorithm 1. It details all the previously described steps.

Moreover, additional details about each step are provided in the following subsections.

Algorithm 1 Pseudocode for Subject-Specific Data Processing and Classification

```
Initialize a data table containing user data, with each user having a unique subject ID
for each unique subject ID do
    Create subject-specific training dataset
    Create subject-specific testing dataset
    for each file in the subject's training dataset do
        Extract trial epochs and corresponding labels from the training dataset
        Extract baseline epochs from the training dataset
    end for
    for each file in the subject's testing dataset do
        Extract trial epochs and corresponding labels from the testing dataset
        Extract baseline epochs from the testing dataset
    end for
end for
Initialize a processing pipeline with predefined configurations (e.g., feature extraction, normalization, classification)
Define hyperparameter grids for optimization (e.g., number of features, maximum dilations per feature, regularization parameters)
Initialize grid search for hyperparameter optimization based on the processing pipeline
Train the model using the training data (X_train, y_train)
Generate predictions and classification probabilities using the testing data (X_test, y_test)
Store the results in a dictionary for further analysis
```

A. DATASET CREATION

The development of our dataset formed the foundational phase of our research. We meticulously collected EEG data files for each participant, augmented by timestamps indicating the onset of baseline periods and specific stimuli. This compilation was methodically structured into a dictionary format, integrating the EEG recordings with corresponding timestamps and markers for the pertinent stimuli.

After data collection, we delineated EEG epochs for each experimental condition, including baseline and stimulus-exposure intervals. These epochs were then subjected to a series of preprocessing steps, as detailed in Sections V-B and V-C. The dataset is publicly available at https://sisinflab.poliba.it/neurosense-dataset-request/.

B. PREPROCESSING

The preprocessing of the EEG signals is crucial due to their inherent susceptibility to various forms of noise and artifacts. While standard pipelines exist for denoising signals, it is essential to tailor these methods to the specific characteristics of each dataset. In our study, participants were given specific instructions during the acquisition phase to minimize head

movements and maintain focus on the stimuli, effectively mitigating various types of artefacts and general noise.

The initial step in preprocessing involved extracting two EEG epochs corresponding to the segments of interest. The first segment encompasses the baseline and the stimuli EEG period, marked by the fixation cross and video start markers. However, accurately determining the timing of the epochs for the stimuli was challenging.

Initially, complete epochs were extracted solely for denoising purposes, utilizing the MNE framework. The maximum epoch length was set to 5 seconds relative to the maximum baseline duration. Subsequently, a filter operation was applied using an FIR filter with a frequency range of $1 - 45$ Hz in MNE.

Given the limited number of electrodes (four in our study), denoising techniques posed a challenge. To address this issue, we adopted a straightforward approach, treating ocular artifacts as outliers relative to the standard distribution of EEG signals. The Megkit framework[9] provided a specific method for removing these artifacts through Ringing Artifact Reduction Techniques. An optimal threshold value was determined by training a simple K-NN algorithm with a contamination hyperparameter specifically set to 0.1 to identify the outlier array index value. If a sample exceeded the detected threshold, the artifact interval was interpolated using samples from the entire trial.

C. TIMESLIDING ESTIMATION

We adopted a dynamic approach to identify the most informative ROIs within the EEG signals over time. This technique involved the dynamic application of a multivariate predictive model across different time points to evaluate its performance continually as new epochs were introduced. Central to this methodology was the use of the MNE framework [36] alongside the SlidingEstimator technique, which required input in the form of feature-objective pairs, with the precondition that the feature dataset extended beyond two dimensions.

This technique was mainly designed to handle the intricacies of EEG data, characterized by its organization into epochs, channels, and sequential time points. By incorporating the temporal dimension directly into the feature set, we enabled the fitting of a distinct estimator at each time point, thus facilitating a nuanced analysis of temporal dynamics within the EEG signal. This temporal decoding strategy draws conceptual parallels to techniques prevalent in fMRI research, aiming to extract maximal differentiation between experimental conditions through the temporal segmentation of the data [37].

Our application of this method within the realm of EEG analysis aimed to leverage the temporal granularity of the EEG data and enhance our understanding of the temporal evolution of emotional responses. Specifically, we sought to pinpoint the optimal differentiation between experimental conditions.

In our study, we identified the best ROI for each target emotion relative to the baseline condition. The entire pipeline to determine the best ROI is outlined as follows:

- Selection of a 5-second sub-epoch for each user.
- Preprocessing each stimulus and baseline epoch using the pipeline described in Section V-B.
- Selection of a time interval ranging from 0 to 25 seconds for each Russell's emotional quadrant, with a step size of 5 seconds.
- Selection of a subsection of trials for each user and emotion to construct training and testing datasets.
- Construction of the dataset comprising stimuli trials related to the selected emotion target and their corresponding baseline EEG.
- Conducting a grid search to identify the best model to retrieve an ROI.
- Application of the SVM classifier within the SlidingEstimator framework.
- Computation of the accuracy score sample by sample using the best model within the selected time interval range.
- Calculation of the 95th percentile for the accuracy distribution of each participant.
- Consideration of only those accuracy scores exceeding the threshold obtained from the 95th percentile.
- Selection of the time segment exhibiting the highest accuracy scores surpassing the threshold.
- Repetition of the entire pipeline for the four emotion classes.

D. MACHINE LEARNING ALGORITHM

In this study, we adopted a binary logic framework to train our models to pinpoint the models that exhibit the highest performance across the four emotional quadrants delineated by Russell's circumplex model. This model is a foundational theoretical framework that categorizes human emotions into a structured system, facilitating a more systematic exploration of emotional states. Specifically, emotions are mapped onto a two-dimensional space defined by the arousal (high vs. low) and valence (positive vs. negative) dimensions, and our classification task is aimed at decoding these emotional states from EEG data. To explore these emotional states, we framed our classification task as a multi-class problem within the dimensional space, where each quadrant represents a unique combination of arousal and valence levels. The ability to accurately classify the EEG signals into these quadrants serves as a proxy for exploring how well the EEG data reflect underlying emotional states. Therefore, the classification task is not merely a matter of predicting labels but also of uncovering the relationship between brain activity patterns and self-reported emotional experiences.

First of all, we implemented a LOSO validation strategy. Upon completing the training phase, we conducted a rigorous

[9]https://nbara.github.io/python-meegkit/

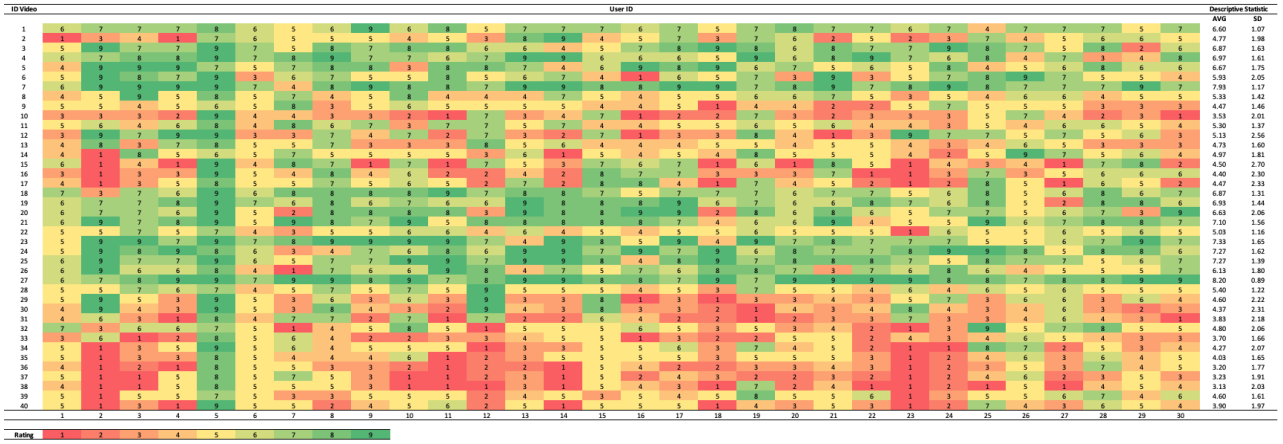


FIGURE 6. Representation of participants' self-assessed *Valence* scores resulting from the SAM questionnaire.

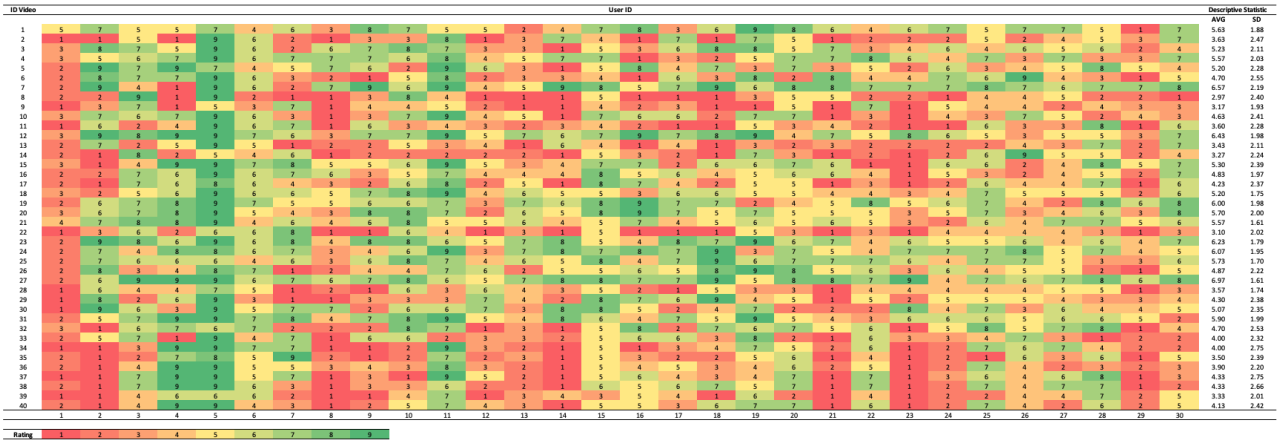


FIGURE 7. Representation of participants' self-assessed *Arousal* scores resulting from the SAM questionnaire.

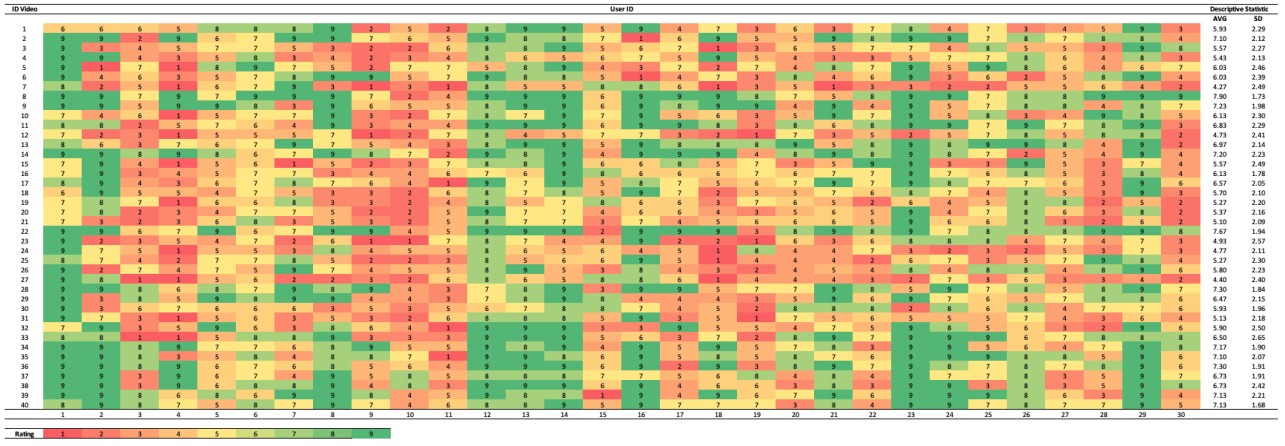


FIGURE 8. Representation of participants' self-assessed *Dominance* scores resulting from the SAM questionnaire.

analysis to identify the most effective models. The criteria for this selection are based on a set of predefined performance metrics, leading to the identification of the top 30 models. These models are recognized for their superior predictive capabilities and are subsequently chosen to conduct the final analyses and predictions.

TABLE 2. External scores estimated by using the DEAP dataset.

ID Video	VAQ estimate	AVG valence	STD valence	AVG arousal	STD arousal	AVG dominance	STD dominance
1	1	6.86	1.30	5.86	2.20	6.00	1.56
2	1	5.93	2.05	6.93	1.98	5.50	2.41
3	1	7.14	1.19	4.86	1.46	5.21	1.31
4	1	6.93	2.32	6.47	1.93	5.80	2.01
5	2	5.93	1.79	3.36	1.34	4.93	1.79
6	2	6.57	1.40	4.21	2.51	5.43	2.06
7	2	7.07	1.43	4.73	2.11	5.33	2.08
8	2	6.47	1.36	4.00	1.79	4.93	1.91
9	3	4.20	1.42	3.73	1.81	3.93	1.95
10	3	3.33	1.19	4.47	2.00	3.20	1.37
11	3	5.13	1.09	2.40	1.62	4.47	2.12
12	3	3.33	1.35	2.93	1.69	4.67	2.15
13	3	4.20	1.64	3.60	1.20	4.60	1.82
14	3	4.20	1.80	3.00	1.51	3.33	1.78
15	4	3.67	1.49	5.47	2.06	4.60	1.74
16	4	4.67	1.49	6.40	1.93	4.93	1.57
17	4	3.93	2.05	6.13	2.09	5.53	1.89
18	1	7.00	1.71	5.93	1.98	6.07	1.98
19	1	7.20	1.76	7.33	1.58	6.53	2.06
20	1	6.13	1.50	6.20	1.60	5.60	1.96
21	1	6.67	1.89	6.47	2.06	5.93	2.32
22	2	6.07	1.61	3.00	1.51	4.80	2.14
23	2	7.13	1.41	3.87	1.86	5.00	2.19
24	2	7.53	1.26	4.47	1.59	5.73	1.48
25	2	7.27	1.34	6.07	1.77	6.67	1.62
26	2	6.27	1.18	4.13	1.71	4.67	1.78
27	2	7.07	1.84	6.40	1.99	6.80	2.17
28	3	4.33	2.02	3.13	1.78	4.80	1.97
29	3	3.25	1.29	2.75	1.44	2.94	1.43
30	3	3.44	1.32	3.63	1.83	3.94	2.30
31	3	3.20	1.76	3.67	1.89	3.27	1.88
32	4	4.79	1.93	6.36	1.95	4.93	2.19
33	4	4.14	1.81	4.21	1.52	4.00	1.81
34	4	3.53	1.89	6.33	2.12	4.93	2.35
35	4	4.93	2.29	7.27	1.34	7.07	2.08
36	4	3.27	1.61	5.87	2.28	5.53	2.28
37	4	3.27	2.11	5.33	2.24	5.67	2.41
38	4	3.47	2.33	5.33	2.39	5.27	2.86
39	4	5.87	2.19	7.07	1.43	7.07	1.61
40	4	3.73	2.59	5.73	2.24	5.53	2.36

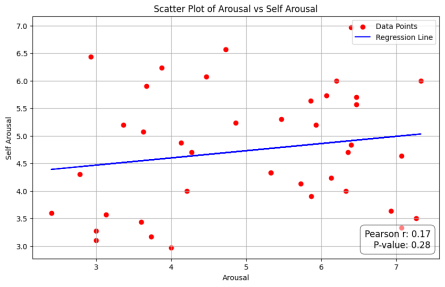
Our methodology incorporated convolutional feature extraction techniques. In particular, we utilized the MiniRocket algorithm, a streamlined version of the Rocket algorithm designed specifically for efficient feature extraction from time series data, as documented by Dempster et al. [38]. The implementation of MiniRocket is adopted from the sktime library, as outlined by Löning et al. [39].

The application of our ML strategy unfolded in two primary phases. Initially, we deployed the MiniRocket algorithm, which entails the convolution of each time series with a set of random convolutional kernels, followed by global max pooling. The proportion of positive outcomes derived from the pooling operation is then leveraged as features for each convolutional kernel. These generated features, capturing pivotal patterns in the multivariate EEG data, are instrumental in the subsequent classification task, typically employing an ML classifier such as the SVM.

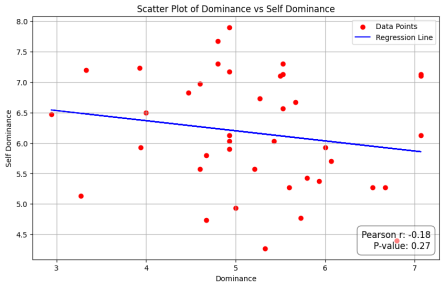
A critical aspect of our feature extraction endeavour is identifying distinctive patterns within EEG data segments, a task influenced by the selection of two key hyperparameters: the size and length of the kernels. We undertook hyperparameter optimization within Python Pipelines, engaging in a multi-stage optimization process. This process began with the adjustment of MiniRocket's kernel size and length parameters, followed by the selection of an optimal normalization technique from among options such as the MinMaxScaler, StandardScaler, and RobustScaler [40]. The optimization concluded with fine-tuning the SVM classifier, focusing on adjusting the C parameter, which governs the regularization strength.

To ascertain the efficacy of our methodological approach, we conducted our evaluations within a 3-fold cross-validation scheme within each LOSO round. Hyperparameter optimization is pursued through Random Search across 50 iterations.

(a) Correlation between external and self-assessed valence scores.



(b) Correlation between external and self-assessed arousal scores.



(c) Correlation between external and self-assessed dominance scores.

FIGURE 9. **Correlation between external and self-assessed indexes.**

These computational procedures are executed using the Python programming language, with the Scikit-learn library providing the necessary tools for ML model implementation and evaluation [40].

E. STATISTICAL ANALYSES

1) RELIABILITY OF EXTERNAL LABELS

We conducted an initial analysis to evaluate the correlation between the self-assessment scores and the external labels, providing a foundation for the subsequent ML model development. In particular, the Pearson correlation index was calculated for arousal and valence scores averaged across participants for each video.

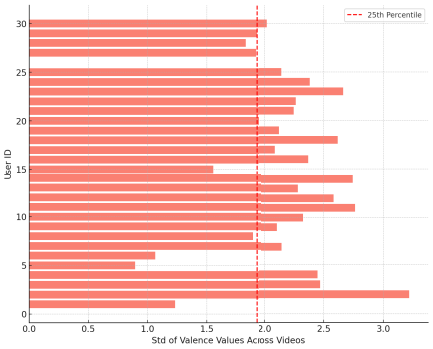
2) CLASSIFICATION OF SAMPLES

After training the ML model to predict the quadrant location of each video on a bivariate plane defined by arousal and valence scores, we used the decision probability scores to assess the average performance of each participant. The probability scores were averaged across the videos to obtain a single numeric performance score for each participant. Moreover, for each participant, we calculated the standard deviation (STD) of the self-assessment scores for arousal, dominance, and valence across the 40 videos. These STD

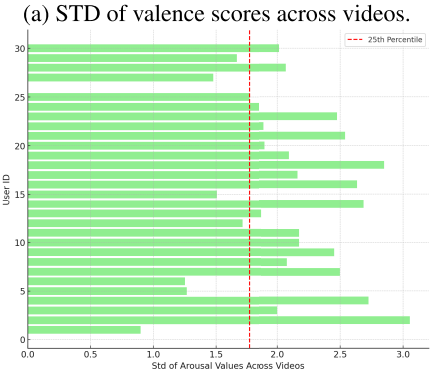
TABLE 3. **Time-sliding accuracy.**

Time Segment [*seconds*]	1° RQ	2° RQ	3° RQ	4° RQ
0-5	**63.36**	**62.60**	**62.12**	**58.24**
5-10	62.44	61.76	60.68	55.6
10-15	62.76	61.12	60.56	55.92
15-20	62.88	62.24	61.72	58.08
20-25	61.47	61.04	61.36	57.88

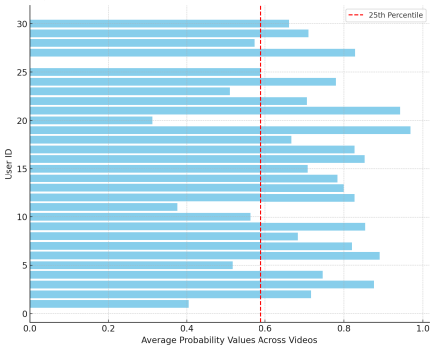
(a) In detail, columns two, three, four, and five represent the average frequency with which the accuracy exceeds the threshold in the first, second, third, and fourth Russell quadrants. RQ = Russel's quadrant.



(a) STD of valence scores across videos.



(b) STD of arousal scores across videos.



(c) Average probability scores across videos.

FIGURE 10. **Representation of the results for the STD analysis for the valence, arousal and probability scores for each participant.**

TABLE 4. Accuracy results for the first quadrant of Russell's Model.

user id	split0_test_score	split1_test_score	split2_test_score	mean_test_score	std_test_score
1	0.68	0.80	0.82	0.77	0.060
2	0.76	0.78	0.82	0.78	0.024
3	0.74	0.79	0.82	0.78	0.034
4	0.75	0.76	0.82	0.78	0.030
5	0.77	0.76	0.82	0.78	0.024
6	0.73	0.79	0.82	0.78	0.035
7	0.75	0.76	0.82	0.78	0.028
8	0.73	0.75	0.79	0.76	0.025
9	0.77	0.77	0.80	0.78	0.015
10	0.73	0.79	0.84	0.79	0.047
11	0.73	0.76	0.83	0.78	0.040
12	0.71	0.76	0.81	0.76	0.039
13	0.77	0.74	0.81	0.77	0.030
14	0.74	0.76	0.81	0.77	0.030
15	0.71	0.71	0.82	0.75	0.050
16	0.77	0.80	0.84	0.80	0.029
17	0.73	0.74	0.82	0.76	0.043
18	0.75	0.71	0.83	0.77	0.049
19	0.70	0.77	0.84	0.77	0.055
20	0.71	0.75	0.80	0.75	0.039
21	0.77	0.75	0.79	0.77	0.019
22	0.74	0.75	0.80	0.76	0.029
23	0.72	0.74	0.78	0.75	0.027
24	0.73	0.76	0.81	0.77	0.032
25	0.74	0.73	0.82	0.76	0.043
26	0.75	0.73	0.84	0.77	0.045
27	0.77	0.73	0.85	0.78	0.052
28	0.76	0.75	0.81	0.77	0.026
29	0.76	0.72	0.82	0.77	0.043
30	0.74	0.73	0.80	0.76	0.029

values served as proxies for the credibility of the participants' engagement and understanding of the task, given the expected high variability across the videos. We compared these STD values to the decision probabilities output by the ML model. By setting the 25^{th} percentile as the lower threshold, we identified participants with low variability in their scores and/or low classifier decision probabilities, flagging them as potentially non-credible or poorly performing participants.

VI. RESULTS

A. RELIABILITY OF EXTERNAL LABELS

Figures 6, 7, and 8 display the results of self-assessment concerning valence, arousal, and dominance scales, respectively. The X-axis represents the 30 participants (user IDs), while the Y-axis depicts the 40 video IDs. Table 2 summarizes the information retrieved from the DEAP dataset.

The correlation analysis reported in Figure 9 revealed varying degrees of alignment between self-assessment scores and external labels. Generally, there was a strong correlation for valence, while arousal and dominance correlations were comparatively weaker and not significant. This discrepancy underscores the challenge of matching subjective self-assessments with external labels obtained from a broader population database.

Nevertheless, external labels derived from a large-scale population study provided a standardized and objective measure of emotional response, essential for the consistency required in ML model training and evaluation. This approach mitigated individual variability and biases inherent in self-assessment data, allowing for more robust model performance. Hence, we opted to utilize the Valence-Arousal Quadrant Estimate available in the DEAP dataset.

B. REGION OF INTEREST COMPUTATION

Table 3 presents the outcomes of the ROI analysis in terms of scores associated with each of Russell's quadrants across different time segments. The results indicate that the most effective time segment for Russell's quadrants is the interval from 0 to 5 seconds. Hence, our pipeline enabled us to identify the EEG signal trial that optimally discriminates between EEG data corresponding to emotional stimulation and EEG data from the resting condition using binary classification methods.

C. CLASSIFICATION PERFORMANCE

Tables 4, 5, 6, and 7 present accuracy split scores, mean test scores, and mean standard deviations for each user in the LOSO cross-validation approach across the k-fold training strategy for each of Russell's quadrants.

TABLE 5. Accuracy results for the second quadrant of Russell's Model.

user id	split0_test_score	split1_test_score	split2_test_score	mean_test_score	std_test_score
0	0.80	0.75	0.78	0.78	0.024
1	0.80	0.78	0.81	0.80	0.013
2	0.77	0.79	0.80	0.79	0.012
3	0.79	0.77	0.81	0.79	0.014
4	0.79	0.81	0.81	0.80	0.011
5	0.78	0.77	0.77	0.78	0.005
6	0.79	0.77	0.80	0.79	0.012
7	0.82	0.77	0.80	0.80	0.022
8	0.77	0.75	0.78	0.77	0.013
9	0.75	0.78	0.82	0.78	0.029
10	0.78	0.80	0.77	0.79	0.013
11	0.76	0.81	0.80	0.79	0.024
12	0.78	0.75	0.79	0.78	0.017
13	0.79	0.77	0.77	0.78	0.011
14	0.82	0.76	0.81	0.80	0.026
15	0.76	0.79	0.80	0.78	0.019
16	0.80	0.75	0.77	0.78	0.021
17	0.78	0.74	0.76	0.76	0.017
18	0.77	0.79	0.77	0.78	0.007
19	0.81	0.72	0.81	0.78	0.044
20	0.81	0.73	0.76	0.77	0.036
21	0.82	0.74	0.81	0.79	0.038
22	0.80	0.72	0.77	0.76	0.034
23	0.80	0.71	0.79	0.77	0.041
24	0.82	0.72	0.79	0.78	0.044
25	0.76	0.74	0.81	0.77	0.030
26	0.78	0.81	0.80	0.80	0.013
27	0.81	0.72	0.79	0.77	0.038
28	0.82	0.73	0.79	0.78	0.039
29	0.78	0.80	0.78	0.79	0.009

Table 8 displays each model's mean accuracy and standard deviation in the LOSO training strategy for the binary classification between stimuli and baseline conditions.

D. CLASSIFICATION OF SAMPLES

Figure 10 shows the horizontal bar plots of the STD values of the self-assessed valence scores (10a), the STD values of the self-assessed arousal scores (9b), and the average probability scores across the 40 videos (10c). The participant with user ID 26 is not considered in the analysis due to a lack of acceptable trials after the preprocessing steps.

The STD analysis identified two participants (user IDs 1 and 5) with consistently low variability across the two metrics (arousal and valence), thus flagging them as participants with untrustworthy labels. Interestingly, some participants exhibited low variability in arousal and valence (user IDs 6 and 15) but maintained high mean decision probabilities, suggesting that their self-assessments did not align well with the classifier's performance. These cases likely indicate misunderstandings or inaccuracies in self-assessment, reinforcing the need for external labels.

Conversely, another group of participants showed high variability in their self-assessments but low classifier performance (user IDs 10, 11, 20, and 23), suggesting that the external labels may not accurately reflect their actual emotional states. This outcome highlights the complex interplay between subjective emotional experiences and objective measures. However, the overall use of external labels remains justified, as they provide a more stable reference point for the classifier, facilitating the identification of broader patterns in EEG data that the subjective nature of self-assessment scores might obscure. This approach ultimately supports the development of more generalizable and reliable emotion recognition models.

VII. DISCUSSION

A. RELIABILITY OF EXTERNAL LABELS

The findings from our study reveal significant insights into the reliability and validity of self-assessment questionnaires used in conjunction with EEG data to predict emotional states. The moderate to strong correlation observed for valence between self-assessment scores and externally derived labels suggests that, to some extent, subjective reports can align with broader population data. However, the weaker correlations for arousal and dominance point to potential limitations in the self-assessment method, which are well-documented in the literature.

One major issue with self-assessment questionnaires, especially when administered immediately after a cognitive task,

TABLE 6. Accuracy results for the third quadrant of Russell's Model.

user id	split0_test_score	split1_test_score	split2_test_score	mean_test_score	std_test_score
0	0.72	0.72	0.76	0.74	0.017
1	0.76	0.73	0.79	0.76	0.024
2	0.68	0.75	0.81	0.75	0.051
3	0.76	0.77	0.79	0.77	0.010
4	0.74	0.75	0.77	0.75	0.015
5	0.76	0.72	0.77	0.75	0.022
6	0.78	0.73	0.81	0.77	0.030
7	0.77	0.75	0.79	0.77	0.013
8	0.75	0.76	0.78	0.76	0.013
9	0.75	0.73	0.76	0.75	0.011
10	0.74	0.79	0.80	0.77	0.022
11	0.72	0.76	0.81	0.77	0.036
12	0.76	0.73	0.78	0.75	0.021
13	0.72	0.76	0.76	0.75	0.020
14	0.76	0.70	0.75	0.74	0.025
15	0.74	0.74	0.80	0.76	0.028
16	0.78	0.74	0.78	0.77	0.018
17	0.72	0.77	0.76	0.75	0.020
18	0.76	0.70	0.76	0.74	0.026
19	0.78	0.71	0.82	0.77	0.045
20	0.74	0.75	0.77	0.76	0.010
21	0.69	0.82	0.82	0.78	0.060
22	0.76	0.80	0.80	0.78	0.019
23	0.76	0.80	0.76	0.77	0.019
24	0.76	0.72	0.77	0.75	0.022
25	0.72	0.81	0.79	0.77	0.035
26	0.75	0.73	0.79	0.76	0.024
27	0.78	0.72	0.78	0.76	0.027
28	0.72	0.77	0.79	0.76	0.028
29	0.76	0.74	0.78	0.76	0.015

is the potential for response biases and inaccuracies. Studies such as those by Paulhus et al. [41] and Podsakoff et al. [42] have highlighted the impact of social desirability bias, where participants may consciously or unconsciously alter their responses to be viewed more favourably. Additionally, the immediate nature of the self-assessment can lead to hurried or less reflective responses, reducing the reliability of the data [43]. Furthermore, cognitive fatigue and emotional carryover effects from the task itself can distort self-assessment scores, as discussed by Schwarz [44]. The experiment lasted approximately 100 minutes for each participant, with a 15-minute break after the first 20 videos. These factors collectively contribute to the observed discrepancies between self-reported measures and objective data, such as the EEG-based classifier outcomes.

B. PERFORMANCE OF THE ML MODELS

The findings from our study demonstrate that our ML models achieved promising accuracy in classifying emotional states across different quadrants of Russell's model using EEG data. The mean accuracy and standard deviation for each user, as presented in Tables 4, 5, 6, and 7, indicate consistent performance across the LOSO cross-validation strategy. Additionally, Table 8 shows the average model accuracy and standard deviations for binary classification between stimuli and baseline conditions, highlighting the robustness of our approach.

Direct comparisons with other studies that utilized publicly available EEG datasets (as shown in Table 9) must be interpreted cautiously due to differences in experimental setups, the number of electrodes, the quality of data, and the specific classification tasks involved. While these studies used more electrodes and higher-resolution EEG devices, our work focuses on demonstrating that even with a low-cost, portable device using only four electrodes, it is possible to achieve competitive results in emotion classification tasks.

For example, the DEAP dataset, which uses 32 electrodes, provides high spatial resolution, leading to improved accuracy in emotion classification tasks. Studies utilizing the DEAP dataset have reported varying accuracies for emotion classification tasks. For instance, Singh et al. proposed a hybrid deep learning model for the quaternary classification of emotions, achieving significant accuracy improvements (accuracy 88.19%) [45]. Cui et al. achieved an accuracy of 83% for predicting the arousal state and 85% for the valence by using the group phase locking value of multichannel EEG [46].

Meng et al. used cascaded convolutional neural networks (CNNs) to achieve an average accuracy of 94.43%

TABLE 7. Accuracy results for the fourth quadrant of Russell's Model.

user id	split0_test_score	split1_test_score	split2_test_score	mean_test_score	std_test_score
0	0.75	0.79	0.81	0.78	0.03
1	0.77	0.79	0.83	0.80	0.03
2	0.78	0.80	0.85	0.81	0.03
3	0.79	0.80	0.85	0.81	0.03
4	0.78	0.74	0.87	0.80	0.06
5	0.77	0.80	0.84	0.80	0.03
6	0.77	0.80	0.84	0.80	0.03
7	0.79	0.81	0.84	0.82	0.02
8	0.77	0.81	0.82	0.80	0.02
9	0.76	0.82	0.84	0.81	0.03
10	0.75	0.85	0.83	0.81	0.04
11	0.77	0.81	0.84	0.81	0.03
12	0.78	0.80	0.83	0.80	0.02
13	0.77	0.82	0.84	0.81	0.03
14	0.76	0.77	0.83	0.79	0.03
15	0.80	0.79	0.83	0.81	0.01
16	0.78	0.81	0.86	0.82	0.03
17	0.75	0.79	0.84	0.80	0.04
18	0.76	0.76	0.82	0.78	0.03
19	0.77	0.80	0.85	0.81	0.03
20	0.79	0.80	0.83	0.81	0.02
21	0.78	0.82	0.83	0.81	0.02
22	0.79	0.81	0.81	0.80	0.01
23	0.75	0.78	0.85	0.79	0.04
24	0.79	0.80	0.86	0.82	0.03
25	0.77	0.79	0.83	0.80	0.02
26	0.79	0.75	0.83	0.79	0.03
27	0.75	0.77	0.83	0.78	0.03
28	0.78	0.83	0.82	0.81	0.02
29	0.75	0.82	0.84	0.80	0.03

TABLE 8. Average model accuracy scores and standard deviations for binary classification between stimuli and baseline conditions.

Models	Average Test score Accuracy	Std score
1	0.77	0.035
2	0.78	0.022
3	0.76	0.024
4	0.80	0.028

and 94.85% in arousal-based and valence-based classification [47]. Similarly, Saha et al. employed multi-band feature extraction for emotion classification, reporting average accuracies of 97.06% for valence and 96.93% for arousal [48]. Ye et al. proposed a deep spatio-temporal mutual learning model, demonstrating the effectiveness of spatio-temporal features in enhancing emotion recognition performance. The authors achieved 98.32% accuracy on four-class classification tasks [49].

In contrast, our dataset employs only four electrodes, significantly reducing the spatial resolution of the EEG data. Despite this limitation, our study reports mean accuracies ranging from 75% to 80% across different quadrants, suggesting that our approach, which integrates external labels for model training, can enhance the classifier's performance. This improvement could be attributed to the use of more representative emotional labels, which potentially offer better generalization capabilities for the model.

The SEED dataset is designed to evoke positive, neutral, and negative emotions, recorded using 62 electrodes [2]. The SEED dataset's multi-session recordings provide a robust evaluation of cross-session variability. Zheng and Lu [2] utilized deep belief networks (DBNs) to achieve an accuracy of approximately 83% for emotion classification. Meng et al. applied their method also on the SEED dataset, achieving an average accuracy of 94.16% [47]. Zhang et al. proposed an attention-based hybrid deep learning model with a final accuracy of 92.47% on SEED datasets [50]. In [51], a flexible analytic wavelet transform (FAWT) that decomposes the EEG signal into different sub-band signals is tested on the SEED, achieving an average classification accuracy of 90.48%.

The high electrode count in SEED allows for capturing detailed spatial information, contributing to high classification accuracy. Our study, with fewer electrodes, still demonstrates competitive accuracy.

The DREAMER dataset contains EEG and ECG recordings from 23 participants watching video stimuli, using 14 electrodes [6]. The dataset includes self-reported ratings of valence, arousal, and dominance. Katsigiannis and Ramzan employed various machine learning models,

TABLE 9. Performance comparison of various EEG datasets and machine learning methods for emotion recognition.

Reference	Dataset	N. of electrodes	Methods	Average Accuracy
[45]	DEAP	32	Hybrid Deep Learning Model	88.19%
[46]	DEAP	32	Group Phase Locking Value	83% (arousal), 85% (valence)
[47]	DEAP	32	Cascaded CNNs	94.43% (arousal), 94.85% (valence)
[48]	DEAP	32	Multi-band Feature Extraction	97.06% (valence), 96.93% (arousal)
[49]	DEAP	32	Deep Spatio-temporal Mutual Learning	98.32%
[2]	SEED	62	Deep Belief Networks	83%
[47]	SEED	62	Cascaded CNNs	94.16%
[50]	SEED	62	Attention-based Hybrid Deep Learning	92.47%
[51]	SEED	62	Flexible Analytic Wavelet Transform	90.48%
[6]	DREAMER	14	KNN and SVM	66% (valence), 63% (arousal)
[52]	DREAMER	14	Sparse DGCNN	68% (valence), 67% (arousal)
This Study	NeuroSense	4	MiniRocket and SVM	75%-80%

reporting accuracies around 66% for valence and 63% for arousal classification using KNN and SVM. More recently, a sparse DGCNN has been developed with different features and spectral bands, including EEG features in the time-frequency domain, with accuracy for valence at 68% and for arousal at 67% [52]. These performance metrics are lower than those achieved in our study, possibly due to differences in stimuli type, feature extraction methods, and the number of electrodes used.

Indeed, the number of electrodes used in EEG studies significantly impacts the quality and richness of the data. More electrodes provide better spatial resolution and capture more detailed neural activity [53]. However, our results demonstrate that even with only four electrodes, it is possible to achieve high classification accuracy. This suggests that the quality of emotional labels and the robustness of the model training approach are critical factors in the performance of emotion recognition systems.

C. CLASSIFICATION OF SAMPLES

Our analysis identified two participants with consistently low variability in their self-assessment scores, suggesting a lack of engagement or understanding of the task. Low variability in self-reported measures can often indicate disengagement or superficial task engagement, as participants may not fully engage with the self-assessment process, leading to less variation in their responses. Additionally, the observation that some participants exhibited high decision probabilities from the classifier despite low variability in self-assessments suggests that these individuals may not have comprehended the assessment instructions properly or may have been inconsistent in their self-reporting. Findings from van der Linden support this concern [54], who discussed the validity of self-reports in dynamic and multifaceted tasks, highlighting the potential for inaccuracies when participants misunderstand or inconsistently apply the assessment criteria.

Moreover, the group of subjects displaying high variability in self-assessments but low classifier performance underscores the potential mismatch between subjective experiences and external labels. This discrepancy might arise because external labels do not fully capture the nuanced emotional states of individual participants, as noted by Russell [19] in his circumplex model of affect, which emphasizes the complexity and variability of emotional experiences. The challenges in aligning subjective self-assessments with objective measures like EEG data highlight the need for improved methodologies and the consideration of individual differences in emotional processing and reporting.

Our study supports the notion that while self-assessment questionnaires provide valuable insights, various factors can compromise their reliability, necessitating a cautious interpretation of the data. Future research should focus on refining self-assessment tools and integrating multimodal approaches to capture individuals' emotional states more accurately.

VIII. CONCLUSION

The primary objective of this work was to present the EEG NeuroSense dataset and demonstrate its potential as a tool for advancing emotion recognition tasks. By making the dataset publicly available, we aim to foster further exploration and innovation within the research community.

Looking forward, the development of regression models to predict continuous values of valence and arousal will be a natural progression from the current discrete classification approach. This shift toward continuous prediction is aligned with current trends in affective computing, which emphasize the importance of capturing the full spectrum of emotional states. In addition, further exploration of the relationships between different EEG frequency bands and specific emotional states will provide deeper insights into the neural underpinnings of emotions, thereby enhancing model interpretability and accuracy.

A key advantage of this study is the use of the Muse2 EEG device, which, with only four electrodes, provides a non-invasive and accessible method for acquiring EEG data. Despite the simplicity of the device, it demonstrated reliable accuracy during signal acquisition, offering a viable alternative to more complex and costly EEG systems. This accessibility is critical for expanding the application of emotion recognition technologies, particularly in real-world contexts where portability and ease of use are essential.

The results obtained with the NeuroSense dataset illustrate the potential of sparse electrode setups for emotion recognition. Even with a limited number of electrodes, our findings suggest that it is possible to achieve competitive performance when compared to more elaborate systems. This highlights the practical value of portable EEG devices, particularly in applications where simplicity and accessibility are crucial.

While this study focuses on the introduction and demonstration of the NeuroSense dataset, we recognize the importance of testing the generalizability of our findings across other public EEG datasets. Future research could also explore more advanced ML techniques to fully exploit the dataset's potential and uncover new insights. In this regard, NeuroSense provides an open platform for the community to explore novel methods and contribute to the evolving field of affective computing.

ACKNOWLEDGMENT

This work is carried out while Tommaso Colafiglio is enrolled in Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome in collaboration with Politecnico di Bari.

REFERENCES

[1] K. H. Kim, S. W. Bang, and S. R. Kim, "Emotion recognition system using short-term monitoring of physiological signals," *Med. Biol. Eng. Comput.*, vol. 42, no. 3, pp. 419–427, May 2004.

[2] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auto. Mental Develop.*, vol. 7, no. 3, pp. 162–175, Sep. 2015.

[3] M. Egger, M. Ley, and S. Hanke, "Emotion recognition from physiological signal analysis: A review," *Electron. Notes Theor. Comput. Sci.*, vol. 343, pp. 35–55, May 2019.

[4] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for continuous emotion detection," *IEEE Trans. Affect. Comput.*, vol. 7, no. 1, pp. 17–28, Jan. 2016.

[5] Y. Liu, O. Sourina, and M. K. Nguyen, "Real-time EEG-based human emotion recognition and visualization," in *Proc. Int. Conf. Cyberworlds*, Oct. 2010, pp. 262–269.

[6] S. Katsigiannis and N. Ramzan, "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 1, pp. 98–107, Jan. 2018.

[7] N. S. Suhaimi, J. Mountstephens, and J. Teo, "EEG-based emotion recognition: A state-of-the-art review of current trends and opportunities," *Comput. Intell. Neurosci.*, vol. 2020, pp. 1–19, Sep. 2020.

[8] C. Ardito, I. Bortone, T. Colafiglio, T. D. Noia, E. D. Sciascio, D. Lofù, F. Narducci, R. Sardone, and P. Sorino, "Brain computer interface: Deep learning approach to predict human emotion recognition," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2022, pp. 2689–2694.

[9] A. Lombardi, S. Marzo, T. Di Noia, E. Di Sciascio, and C. Ardito, "Exploring the usability and trustworthiness of AI-driven user interfaces for neurological diagnosis," in *Adjunct Proc. 32nd ACM Conf. User Model., Adaptation Personalization*, Jun. 2024, pp. 627–634.

[10] O. Bălan, G. Moise, L. Petrescu, A. Moldoveanu, M. Leordeanu, and F. Moldoveanu, "Emotion classification based on biophysical signals and machine learning techniques," *Symmetry*, vol. 12, no. 1, p. 21, Dec. 2019.

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

[12] T. Colafiglio, P. Sorino, D. Lofu, A. Lombardi, F. Narducci, and T. Di Noia, "Combining mental states recognition and machine learning for neurorehabilitation," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, vol. 5, Oct. 2023, pp. 3848–3853.

[13] A. T. Sohaib, S. Qureshi, J. Hagelbäck, O. Hilborn, and P. Jerčić, "Evaluating classifiers for emotion recognition using EEG," in *Proc. 7th Int. Conf. Found. Augmented Cognition*, Las Vegas, NV, USA. Berlin, Germany: Springer, Jul. 2013, pp. 492–501.

[14] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 327–339, Jul. 2014.

[15] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," 2015, *arXiv:1511.06448*.

[16] E. Gibson, N. J. Lobaugh, S. Joordens, and A. R. McIntosh, "EEG variability: Task-driven or subject-driven signal of interest?" *NeuroImage*, vol. 252, May 2022, Art. no. 119034.

[17] W.-L. Zheng and B.-L. Lu, "Personalizing EEG-based affective models with transfer learning," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, Jul. 2016, pp. 2732–2738.

[18] G. Chanel, J. Kronegg, D. Grandjean, and T. Pun, "Emotion assessment: Arousal evaluation using EEG's and peripheral physiological signals," in *Proc. Int. Workshop Multimedia Content Represent., Classification Secur.* Berlin, Germany: Springer, 2006, pp. 530–537.

[19] J. A. Russell, "A circumplex model of affect.," *J. Personality Social Psychol.*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980.

[20] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012.

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

[22] T. Song, W. Zheng, C. Lu, Y. Zong, X. Zhang, and Z. Cui, "MPED: A multi-modal physiological emotion database for discrete emotion recognition," *IEEE Access*, vol. 7, pp. 12177–12191, 2019.

[23] W. G. Parrott, *Emotions in Social Psychology: Essential Readings*. London, U.K.: Psychology Press, 2001.

[24] M. Soleymani, J. J. M. Kierkels, G. Chanel, and T. Pun, "A Bayesian framework for video affective representation," in *Proc. 3rd Int. Conf. Affect. Comput. Intell. Interact. Workshops*, Sep. 2009, pp. 1–7.

[25] P. D. Symes, *Video Compression*. New York, NY, USA: McGraw-Hill, 1998.

[26] P. Kelm, S. Schmiedeke, and T. Sikora, "Feature-based video key frame extraction for low quality video sequences," in *Proc. 10th Workshop Image Anal. Multimedia Interact. Services*, May 2009, pp. 25–28.

[27] K. Cantrell, M. M. Erenas, I. de Orbe-Payá, and L. F. Capitán-Vallvey, "Use of the hue parameter of the hue, saturation, value color space as a quantitative analytical parameter for bitonal optical sensors," *Anal. Chem.*, vol. 82, no. 2, pp. 531–542, Jan. 2010.

[28] M. Mahy, L. Van Eycken, and A. Oosterlinck, "Evaluation of uniform color spaces developed after the adoption of CIELAB and CIELUV," *Color Res. Appl.*, vol. 19, no. 2, pp. 105–121, Apr. 1994.

[29] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 143–154, Feb. 2005.

[30] I. E. Richardson, *Video Codec Design: Developing Image and Video Compression Systems*. Hoboken, NJ, USA: Wiley, 2002.

[31] B. Gorjian, A. Hayati, and P. Pourkhoni, "Using praat software in teaching prosodic features to EFL learners," *Proc. Social Behav. Sci.*, vol. 84, pp. 34–40, Jul. 2013.

[32] M. Tipping, "The relevance vector machine," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 12, 1999, pp. 1–7.

[33] R. W. Homan, J. Herman, and P. Purdy, "Cerebral location of international 10–20 system electrode placement," *Electroencephalogr. Clin. Neurophysiology*, vol. 66, no. 4, pp. 376–382, Apr. 1987.

[34] A. Barachant, D. Morrison, H. Banville, J. Kowaleski, U. Shaked, S. Chevallier, and J. J. T. Tresols, "muse-lsl," May 2019, doi: 10.5281/zenodo.3228861.

[35] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. Behav. Therapy Experim. Psychiatry*, vol. 25, no. 1, pp. 49–59, Mar. 1994.

[36] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. S. Hämäläinen, "MEG and EEG data analysis with MNE-Python," *Frontiers Neurosci.*, vol. 7, no. 267, pp. 1–13, 2013.

[37] S. A. Engel, D. E. Rumelhart, B. A. Wandell, A. T. Lee, G. H. Glover, E.-J. Chichilnisky, and M. N. Shadlen, "fMRI of human visual cortex," *Nature*, vol. 369, p. 525, Jun. 1994.
[38] A. Dempster, D. F. Schmidt, and G. I. Webb, "MiniRocket: A very fast (Almost) deterministic transform for time series classification," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2021, pp. 248–257.
[39] M. Löning, A. Bagnall, S. Ganesh, V. Kazakov, J. Lines, and F. J. Király, "Sktime: A unified interface for machine learning with time series," 2019, *arXiv:1909.07872*.
[40] F. Pedregosa, S. Varoquaux, A. Gramfort, V. Michel, and B. Thirion, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Dec. 2011.
[41] D. L. Paulhus and S. Vazire, "The self-report method," in *Handbook of Research Methods in Personality Psychology*, vol. 1. New York, NY, USA: Guilford Press, 2007, pp. 224–239.
[42] P. M. Podsakoff, S. B. MacKenzie, J.-Y. Lee, and N. P. Podsakoff, "Common method biases in behavioral research: A critical review of the literature and recommended remedies.," *J. Appl. Psychol.*, vol. 88, no. 5, pp. 879–903, 2003.
[43] A. A. Stone, S. S. Shiffman, and M. W. DeVries, "Ecological momentary assessment," *Ann. Behav. Med.*, vol. 16, pp. 199–202, 1994.
[44] N. Schwarz, "Retrospective and concurrent self-reports: The rationale for real-time data capture," in *The Science of Real-Time Data Capture: Self-Reports in Health Research*, vol. 11. New York, NY, USA: Oxford Univ. Press, 2007, p. 26.
[45] K. Singh, M. K. Ahirwal, and M. Pandey, "Quaternary classification of emotions based on electroencephalogram signals using hybrid deep learning model," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 3, pp. 2429–2441, Mar. 2023.
[46] G. Cui, X. Li, and H. Touyama, "Emotion recognition based on group phase locking value using convolutional neural network," *Sci. Rep.*, vol. 13, no. 1, p. 3769, Mar. 2023.
[47] M. Meng, Y. Zhang, Y. Ma, Y. Gao, and W. Kong, "EEG-based emotion recognition with cascaded convolutional recurrent neural networks," *Pattern Anal. Appl.*, vol. 26, no. 2, pp. 783–795, May 2023.
[48] O. Saha, Md. S. Mahmud, S. A. Fattah, and M. Saquib, "Automatic emotion recognition from multi-band EEG data based on a deep learning scheme with effective channel attention," *IEEE Access*, vol. 11, pp. 2342–2350, 2023.
[49] W. Ye, X. Li, H. Zhang, Z. Zhu, and D. Li, "Deep spatio-temporal mutual learning for EEG emotion recognition," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2022, pp. 1–8.
[50] Y. Zhang, Y. Zhang, and S. Wang, "An attention-based hybrid deep learning model for EEG emotion recognition," *Signal, Image Video Process.*, vol. 17, no. 5, pp. 2305–2313, Jul. 2023.
[51] V. Gupta, M. D. Chopda, and R. B. Pachori, "Cross-subject emotion recognition using flexible analytic wavelet transform from EEG signals," *IEEE Sensors J.*, vol. 19, no. 6, pp. 2266–2274, Mar. 2019.
[52] G. Zhang, M. Yu, Y.-J. Liu, G. Zhao, D. Zhang, and W. Zheng, "SparseDGCNN: Recognizing emotion from multichannel EEG signals," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 537–548, Jan. 2023.
[53] A. Criscuolo and E. Brattico, "Fundamentals of electroencephalography and magnetoencephalography," in *Language Electrified: Principles, Methods, and Future Perspectives of Investigation*. Berlin, Germany: Springer, 2023, pp. 163–194.
[54] W. J. Van der Linden, *Handbook of Item Response Theory: Three Volume Set*. Boca Raton, FL, USA: CRC Press, 2018.

TOMMASO COLAFIGLIO is currently pursuing the Ph.D. degree with Sapienza University of Rome, Italy. His research interest includes user personalization, using data from brain–computer interfaces.

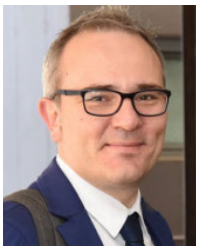
ANGELA LOMBARDI received the degree (Hons.) in telecommunications engineering from the Polytechnic University of Bari, in 2014, and the Ph.D. degree (cum laude) in signal processing and computational neuroscience from the Department of Electrical and Information Engineering, Polytechnic University of Bari, in January 2018, with a thesis on multidimensional dynamic analysis of physiological brain signals. She was a Postdoctoral Researcher in computational neuroscience with the National Institute for Nuclear Physics, from July 2018 to December 2020. From December 2020 to August 2022, she was an Assistant Professor with the Physics Department, University of Bari. She is currently an Assistant Professor (tenure-track position) with the Polytechnic University of Bari, Italy. Her research focused on novel, efficient algorithms for early diagnosis and personalized staging of neurodegenerative diseases, and aging processes with advanced artificial intelligence methods in distributed computing environments. Her primary research interests include brain–computer interfaces, human-centered artificial intelligence, and quantitative methods for explainable artificial intelligence (XAI).

PAOLO SORINO is currently pursuing the Ph.D. degree. He is a Research Assistant with Politecnico di Bari. He is a Researcher with IRCCS Saverio de Bellis, Castellana Grotte, where he worked on the analysis of epidemiological data using AI techniques.

ELVIRA BRATTICO is currently a Full Professor of neuroscience with Aarhus University, the Principal Investigator of learning of the Center for Music in the Brain, Denmark, and a Full Professor of general psychology with the University of Bari, Italy. She has published more than 150 peer-reviewed articles and is a board member of several institutions and publishers. She periodically acts as a panelist for European Commission. Her research interests include brain changes after sensorimotor learning, neural mechanisms for evaluative judgments and emotions, and sound-derived neuroplasticity. For these, she integrates methodologies from neuroimaging, experimental psychology, genetics, and engineering.

DOMENICO LOFÙ (Member, IEEE) received the master's degree (Hons.) in computer engineering from Politecnico di Bari, Italy, and the Ph.D. degree in electrical and information engineering from the Department of Electrical and Information Engineering (DEI), Polytechnic University of Bari, in December 2022. From 2019 to 2023, he was a Security&AI Researcher with Innovation Laboratory (ILAB), Exprivia S.p.A., Italy, where he is currently involved in the EU H2020 ECHO Project. He has been an Assistant Professor with the Polytechnic University of Bari, since December 2024. His major research interests include artificial intelligence for intelligent systems for healthcare and information security. He is serving on the TPC of several conferences.

DANILO DANESE is currently pursuing the Ph.D. degree with Politecnico di Bari. His main research interests include data augmentation in high-dimensional and low-sample-size domains.



TOMMASO DI NOIA (Member, IEEE) is currently a Full Professor of information processing systems with the Polytechnic University of Bari. His primary research interests include artificial intelligence and data management, with a particular emphasis on machine learning techniques, applications, and recommender systems. In recent years, he has focused on leveraging information encoded within Big data datasets, such as those available through the Linking Open Data Initiative, to develop content-based or hybrid recommendation engines. Additionally, his research has concentrated on trustworthy AI, particularly in adversarial machine learning, explainability, fairness, and privacy protection in recommendation models. He has numerous papers published in international journals, conference proceedings, book chapters, and several best paper awards.

EUGENIO DI SCIASCIO He is currently a Full Professor with the Polytechnic University of Bari, Italy, where he is also the Scientific Coordinator of SisInfLab. Over the years, he has held numerous academic and administrative positions, contributing significantly to the university and the broader academic community. As the Scientific Coordinator of SisInfLab, he leads a research group that has produced more than 250 publications. He has led or participated in numerous international and national research projects (FP 6/7, Horizon 2020, Interreg, PON, POR, and PRIN). His research interests include structured knowledge representation and artificial intelligence, particularly in the areas of intelligent pervasive systems, the web of data, and applications related to smart cities, recommender systems, cyber-physical systems, and advanced information systems. He has received numerous awards and recognitions for his research. He has been an invited speaker at more than 100 international conferences, including the International Joint Conference on Artificial Intelligence (IJCAI), the International Semantic Web Conference (ISWC), the ACM International Conference on Electronic Commerce, American Association for Artificial Intelligence (AAAI), and the WWW Conference.

FEDELUCIO NARDUCCI received the Ph.D. degree in computer science, defending his thesis titled ''Knowledge-Enriched Representations for Content-Based Recommender Systems.'' From April 2012 to July 2014, he was a Postdoctoral Researcher with the University of Milano-Bicocca. From August 2014 to March 2018, he was a Postdoctoral Researcher with the University of Bari Aldo Moro, working on conversational recommender systems, health assistant chatbots, multimedia, and emotion-aware recommender systems, with a focus on explanation and sentiment analysis. Currently, he is an Associate Professor with the Polytechnic University of Bari. He is an Associate Professor with the Polytechnic University of Bari, Italy. Over the years, he has authored several papers in international journals and conferences in the artificial intelligence and recommender-system research areas. His research is focused on techniques for intelligent access to multilingual e-gov services. His current research interests include recommender systems, conversational agents, natural language processing, e-health, personalized information access, user modeling and personalization, and explanation and fairness for AI systems.

Open Access funding provided by 'Politecnico di Bari' within the CRUI CARE Agreement