Submitted: 12 January, 2024 Accepted: 08 April, 2024 Published: 08 December, 2024 DOI:10.22514/sv.2024.153

Open Access

Signa Vitae

ORIGINAL RESEARCH

AI-based cancer pain assessment through speech emotion recognition and video facial expressions classification

Marco Cascella[1,*], Francesco Cutugno[2], Fabio Mariani[2], Vincenzo Norman Vitale[2], Manuel Iuorio[2], Arturo Cuomo[3], Sabrina Bimonte[3], Valeria Conti[1], Francesco Sabbatino[1], Alfonso Maria Ponsiglione[2], Jonathan Montomoli[4], Valentina Bellini[5], Federico Semeraro[6], Alessandro Vittori[7], Elena Giovanna Bignami[5], Ornella Piazza[1]

[1] Department of Medicine, Surgery and Dentistry, University of Salerno, 84081 Baronissi, Italy
[2] DIETI, University of Naples "Federico II", 80125 Naples, Italy
[3] Department of Anesthesia and Pain Medicine, National Cancer Institute, 80131 Naples, Italy
[4] Department of Anesthesia and Intensive Care, Infermi Hospital, AUSL Romagna, 47923 Rimini, Italy
[5] Anesthesiology, Critical Care and Pain Medicine Division, Department of Medicine and Surgery, University of Parma, 43126 Parma, Italy
[6] Department of Anaesthesia, Intensive Care and Emergency Medical Services, Maggiore Hospital, 40133 Bologna, Italy
[7] Department of Anesthesia and Critical Care, ARCO Roma Ospedale Pediatrico Bambino Gesù IRCCS, 00165 Rome, Italy

***Correspondence**
mcascella@unisa.it
(Marco Cascella)

Abstract

The effective assessment of cancer pain requires a meticulous analysis of all the components that shape the painful experience collectively. Implementing Automatic Pain Assessment (APA) methods and computational analytical approaches, with a specific focus on emotional content, can facilitate a thorough characterization of pain. The proposed approach moves towards the use of automatic emotion recognition from speech recordings alongside a model we previously developed to examine facial expressions of pain. For training and validation, we adopted the EMOVO dataset, which simulates six emotional states (the Big Six). A Neural Network, consisting of a Multi-Layered Perceptron, was trained on 181 prosodic features to classify emotions. For testing, we used a dataset of interviews collected from cancer patients and selected two case studies. Speech annotation and continuous facial expression analysis (resulting in pain/no pain classifications) were carried out using Eudico Linguistic Annotator (ELAN) version 6.7. The model for emotion analysis achieved 84% accuracy, with encouraging precision, recall, and F1-score metrics across all classes. The preliminary results suggest the potential use of artificial intelligence (AI) strategies for continuous estimation of emotional states from video recordings, unveiling predominant emotional states, and providing the ability to corroborate the corresponding pain assessment. Despite limitations, the proposed AI framework exhibits potential for holistic and real-time pain assessment, paving the way for personalized pain management strategies in oncological settings. Clinical Trial registration: NCT04726228.

Keywords

Automatic pain assessment; Pain; Cancer pain; Artificial intelligence; Speech analysis; Computational language analysis; Speech emotion recognition

1. Introduction

Pain is a prevalent and debilitating symptom in cancer patients. Research indicates that approximately 55% of patients encounter pain during anticancer treatment, a percentage that escalates to 66% for individuals grappling with metastatic, advanced, or terminal diseases [1]. While accurate symptom assessment is crucial for effective pain management [2], conventional unidimensional pain rating scales, such as the 0–10 numeric rating scale (NRS) and the Visual Analog Scale (VAS), exhibit significant limitations. These scales are susceptible to reporting bias, influenced by psychosocial factors like tendencies to catastrophize or underreport pain [3]. Furthermore, although multidimensional assessment tools prove useful in evaluating various features of a patient's pain experience, they fail to provide an objective and exhaustive measurement of pain [4]. Additionally, the inherently subjective nature of pain, coupled with self-reported assessments relying on individual interpretation and communication, introduces another paramount bias [5].

In this complex landscape, the exploration of pain assessment can be enriched through multifaceted evaluations, harnessing the capabilities of artificial intelligence (AI) methodologies for comprehensive pain analysis [6]. Notably, AI has the potential to significantly impact pain assessment by offering innovative solutions for personalized treatment strategies. Therefore, efforts are directed toward advancing automatic pain assessment (APA) systems [7]. For instance, there have been successful applications of computer vision and machine learning (ML) techniques in assessing pain through facial expressions [8], alongside other advancements centered around neurophysiology-based pain detection methods [7].

Since pain encompasses both sensory and emotional dimensions, the careful assessment of the emotional component has a significant impact on the evaluation of the whole painful experience [9]. Significantly, speech analysis could be used for this purpose. Speech analysis, the systematic examination of spoken language, aims to extract meaningful information, recognize patterns, and derive valuable insights [10]. This comprehensive process encompasses various speech aspects, including acoustic features, linguistic content, intonation, rhythm and more. Therefore, speech analysis has numerous applications such as speech recognition, speaker identification, language identification, speech pathology, and emotion recognition (*i.e.*, speech emotion recognition, SER) [11]. In the context of vocal emotion recognition, the focus of analysis revolves around the acoustic features of the voice. These features, including pitch, tone, rhythm and intensity, are scrutinized to identify distinctive patterns associated with different emotional states [12].

By implementing AI strategies, automated systems can more efficiently process and interpret speech data. Currently, speech analysis is regarded as a specialized branch of Natural Language Processing (NLP), a subfield of AI that focuses on human-computer interaction through spoken and written natural language [13] both spoken and written. Indeed, some very specific speech analysis applications deal with the interpretation and understanding of natural spoken language and related nuances. Among other tasks, speech analysis contributes to emotion recognition, identifying and understanding the speaker's emotional state. This approach is commonly used for sentiment analysis and emotional well-being assessments [14]. In medicine, computational language analyses have been performed for multiple purposes such as functional evaluation after laryngeal surgery [15], respiratory disease detection from voice [16–18], automatic depression detection [19], and investigation in neurodegenerative diseases [20].

Previously, we developed and validated an APA model based on facial expression analysis. By analyzing videos obtained from patients with oncological pain, the binary classifier demonstrated excellent performance in recognizing the presence or absence of pain. The classifier achieved 94% accuracy, and the area under the receiver operating characteristic Curve (AUROC) value was 0.98 [8].

Given these considerations, we introduce an AI framework for APA in cancer patients. In the first part of the study, we describe the computational language analysis for emotion recognition. Subsequently, the APA approach entails the concurrent use of two distinct models, one for automatic emotion recognition from speech recordings and another for analyzing facial expressions of pain. Two case studies are utilized to demonstrate the viability of the framework. The aim is to pave the way for a holistic pain assessment and, ultimately, personalized pain management strategies.

2. Materials and methods

This study is integral to a broader multimodal APA project, incorporating computer vision and natural language processing methodologies, alongside the analysis of physiological signals.

2.1 Study framework

Cancer patients participated in a brief interview and video recording process. Following a methodology previously utilized in another study on APA, a psychologist conducted the interviews, covering predetermined subjects such as personal information and pain-related aspects [8]. The gathered data was then uploaded to our repository for further processing, including normalization and cleaning procedures. Subsequently, we implemented a model, as detailed in [8], aimed at discerning pain from non-pain by examining non-volitional facial expressions associated with pain. Additionally, an audio signal processing model, trained on speech datasets, focused solely on analyzing prosodic features of the audio files, categorizing them based on different emotional states. Finally, continuous analysis was conducted on both video footage (facial expression) and language patterns throughout the video.

2.2 Dataset implementation for training and validation

The leading idea is that prosody and speaking features can significantly express pain as a set of complex emotions. Nevertheless, this approach faces challenges due to the scarce availability of speech data annotated according to any pain scale. To overcome this issue, we adopted the EMOVO dataset (https://paperswithcode.com/dataset/emovo), a repository created from the voices of 6 actors (3 males and 3 females) replicating 14 phrases to simulate 6 emotional states (disgust, fear, anger, joy, surprise and sadness), along with a neutral state (`http://voice.fub.it/EMOVO`) [21]. Overall, these emotions constitute the Big Six. According to Ekman's theories, since these emotions are widely recognized across cultures, they are commonly referred to as the basic or primary emotions and find extensive use in research [22]. The dataset contains a balanced distribution of all classes (emotions) [21].

2.3 Signal processing

We used the Python package Librosa (https://librosa.org/) for music and audio analysis to extract prosodic features directly from the speech signals. The features included those addressing mel-frequency cepstral coefficients that depict the short-term power spectrum of sound and are widely applied in speech and audio processing; chroma features, illustrating the energy distribution of pitch classes (musical notes) within an audio signal. They prove beneficial for music analysis and chord recognition. Moreover, the library includes spectral contrast metrics to gauge the amplitude difference between peaks and valleys in the audio spectrum, serving purposes such as instrument recognition and speech analysis; zero-crossing rate to quantify how frequently the audio signal crosses the zero axis, providing insights into the signal's noisiness, and root mean square energy to measure the overall energy of the audio signal, making it valuable for tasks like audio event detection and classification [23, 24]. Readings were taken every 32 milliseconds. All readings were aggregated by calculating the mean along the temporal axis.

Therefore, by implementing the Sklearn library, we realized the neural network, *i.e.*, a Multi-Layered Perceptron, which processes a total of 181 prosodic features to assign a score to each of the 6 + 1 considered states. Regarding the input signals referred to in the previous paragraph, they present a sampling frequency of 16 kHz.

2.4 Neural network

The developed neural network is a Multi-Layered Perceptron that uses a Rectified Linear Unit (ReLU) activation function and is trained with the Adam (Adaptive Moment Estimation) algorithm for weight optimization to minimize the Mean Square Error. Adam combines principles from other optimization algorithms such as Root Mean Square Propagation (RMSprop) and Momentum, rendering it especially well-suited for tasks of this nature. Specifically, this stochastic optimization method excels in addressing challenges posed by substantial data and/or parameter sizes. The approach is also fitting for objectives that evolve and problems featuring exceedingly noisy and/or sparse gradients [25].

The Neural Network is composed of an input layer with 181 prosodic features, followed by three hidden layers of 300, 200 and 100 neurons respectively, and an output layer that maps the resulting stimuli over six emotional states (disgust, joy, fear, anger, surprise, sadness) plus a neutral state. The resulting emotional state is the one with the highest probability (Fig. 1).

2.5 Model validation and performance assessment

For a more robust analysis, the dataset can be divided into train and validation through K-Fold Cross-Validation. At each iteration, different portions are used for training and validation. Consequently, each time it is tested on a different portion of the dataset. The final accuracy result is then the average of the accuracies from the various iterations.

In this study, the original dataset was partitioned into two distinct subsets. The larger subset, constituting 90% of the total data, was allocated for the training of our network, while the smaller subset, comprising the remaining 10% of the data, was reserved for testing purposes. In the initial phase, we implemented a stratified K-Fold Cross Validation on the training subset. The hyperparameter K was set to 15, a value that was

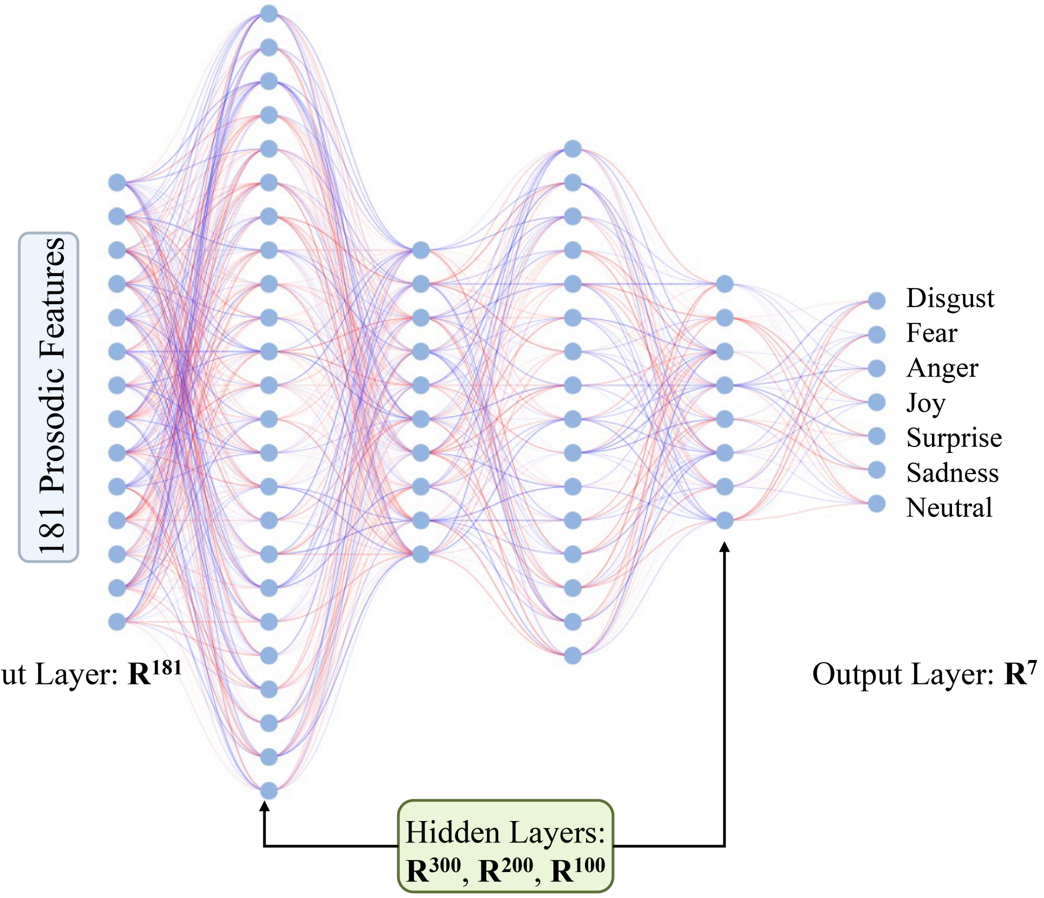
181 Prosodic Features
Input Layer: $\mathbf{R}^{181}$
Disgust
Fear
Anger
Joy
Surprise
Sadness
Neutral
Output Layer: $\mathbf{R}^{7}$
Hidden Layers:
$\mathbf{R}^{300}$, $\mathbf{R}^{200}$, $\mathbf{R}^{100}$

FIGURE 1. Multilayer perceptron classifier. The classifier is made up of five dense layers. The first layer (input) consists of 181 nodes associated with prosodic features extracted for the audio clip of each image. There are three hidden layers of 300, 200 and 100 nodes respectively. The output encompasses six emotional states (disgust, joy, fear, anger, surprise, sadness) plus a neutral state. We implemented the Rectified Linear Unit (ReLU) activation function and the Adaptive Moment Estimation (Adam) algorithm for weight optimization. Softmax function was applied to the final (output) layer to obtain the probability distribution of the output across the 7 possibilities. The training process was guided by the cross-entropy loss. Speech emotion analysis is performed through the Multilayer Perceptron Classifier.

confirmed through subsequent optimization processes. This resulted in the creation of 14 folds for training the network and a single fold for its validation. The partitioning of data into folds adhered to a specific criterion, which ensured that each fold maintained the same proportion of observations as the initial 90% partition. An examination of the training and validation loss curves of the model, yielded by the K-Fold Cross Validation procedure, revealed no signs of overfitting. The K-Fold model was subsequently evaluated using the dedicated test set. This set consisted of data that had not been encountered or assessed during the training phase, thus accounting for the remaining 10% of the total data.

In terms of metrics, the following have been adopted to assess the performance of the overall model and for each class:

- Accuracy is the ratio of correct predictions to the total number of predictions. It indicates the proportion of cases that were correctly predicted. The higher the accuracy value, the higher the percentage of samples correctly classified.

- Recall: defined as the ratio of true positives to the total number of true positive samples. It indicates the proportion of positive cases that were correctly predicted as positive.

- Precision: defined as the ratio of true positives to the total number of examples predicted as positive (TP + FP). It indicates the proportion of examples predicted as positive that are positive.

- F1-score: is a weighted average of precision and recall and is calculated as:

$$F1\text{-}score = 2 \times (Precision \times Recall)/(Precision + Recall)$$

2.6 Dataset for prediction

Audio files were obtained from brief interviews, approximately two minutes long, with patients experiencing cancer-related pain. Patients were asked for demographic information, clinical data related to the underlying pathology, and a description of the pain symptoms [8].

2.7 Speech annotation and video-audio correlation

For speech annotation, we used the Eudico Linguistic Annotator (ELAN), version 6.7 (Max Planck Institute for Psycholinguistics, Nijmegen, Gelderland, Netherlands) (https://archive.mpi.nl/tla/elan/download). ELAN stands out as a sophisticated software tool tailored for the creation, editing, and comprehensive analysis of multimedia data, encompassing video, audio, and text. Its time-aligned annotations feature proved invaluable for meticulously examining the temporal relationships between different events in multimedia content. To achieve continuous estimation, our methodology involved incorporating the entire patient interview into ELAN. This encompassed not only the spoken content but also its corresponding frame prediction output, encompassing emotions and neutral expressions.

The software was used to perform video analysis and language analysis continuously. In a prior research endeavor, we leveraged two datasets containing pain-related data [26, 27] to train a model focused on video recordings of oncology patients experiencing pain. For this aim, we utilized a set of 17 facial expressions (Action Units, AUs) (Table 1).

TABLE 1. The set of 17 action units adopted for automatic video analysis [8].

Action Unit (AU)	Description
AU1	Inner Brow Raiser
AU2	Outer Brow Raiser
AU4	Brow Lowerer
AU5	Upper Lid Raiser
AU6	Cheek Raiser
AU7	Lid Tightener
AU9	Nose Wrinkler
AU10	Upper Lip Raiser
AU12	Lip Corner Puller
AU14	Dimpler
AU15	Lip Corner Depressor
AU17	Chin Raiser
AU20	Lip Stretched
AU23	Lip Tightener
AU25	Lips Part
AU26	Jaw Drop
AU28	Lip Suck
AU45	Blink

The OpenFace toolkit was employed to extract these specific AUs for each image [28]. Our developed Neural Network classifier consists of two dense layers. The initial layer encompasses 17 nodes, each associated with the facial AUs extracted by OpenFace for every image. The output layer assigns a classification label of "pain" (1) or "no pain" (0) [8].

2.8 Application of continuous audio-video analysis to use cases

We used the previously created APA model [8] for facial expression analysis from video recordings as well as the emotion recognition model from audio recordings to combine the emotion recognition model and the cancer pain recognition model within a larger framework for the assessment of emotional states and pain in oncological patients. Therefore, the approach adopts a continuous audio-video analysis through the two models. Fig. 2 shows a screenshot of an ongoing analysis that captures both pain from facial expressions and emotions from audio traces within the same video acquisition.

For this study, two cases of patients referred to the Pain Medicine Unit (National Cancer Center, Napoli, Italy) were examined to demonstrate the application of the model in real-world scenarios. The first case involves an oncological patient without cancer pain, while the second case pertains to a patient experiencing cancer-related pain. Comprehensive clinical descriptions for each case are provided in the subsequent sections.

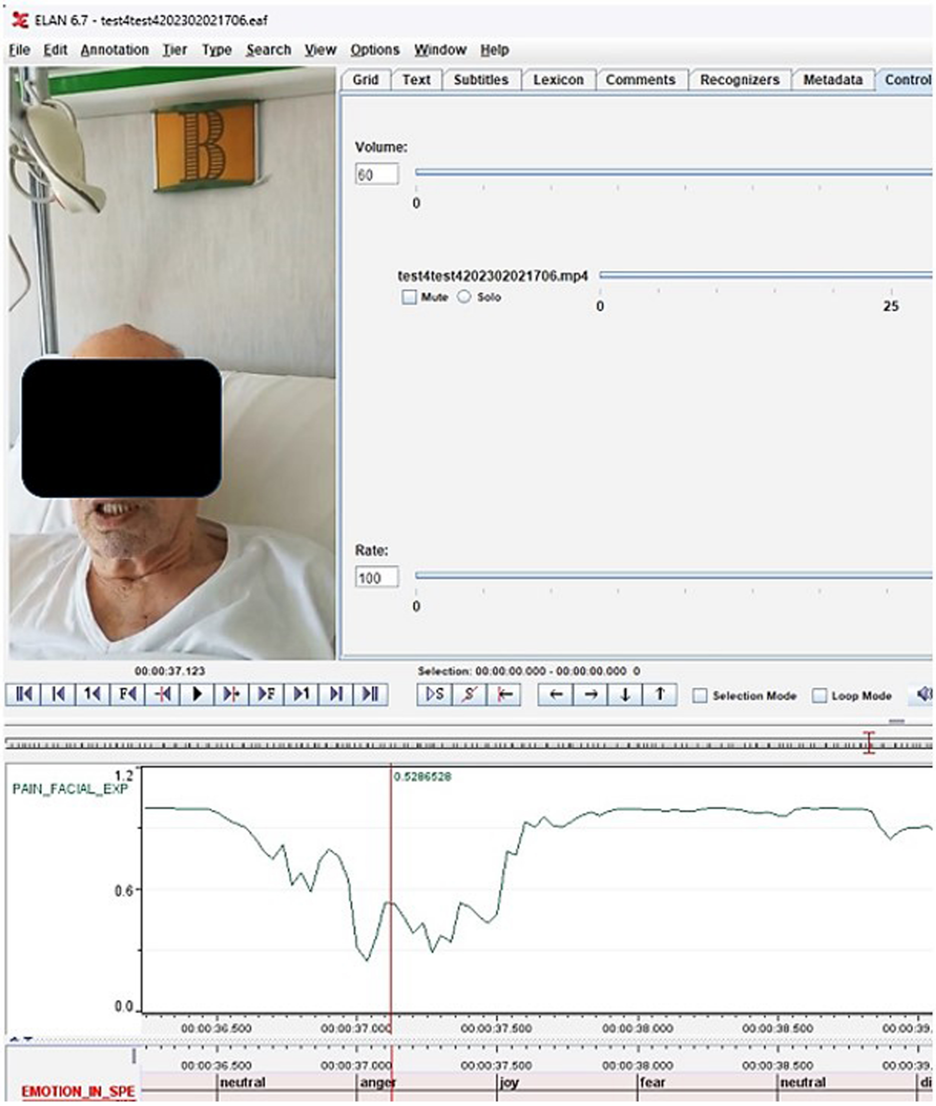
[redacted]
ELAN 6.7 - test4test4202302021706.eaf
File Edit Annotation Tier Type Search View Options Window Help
Grid Text Subtitles Lexicon Comments Recognizers Metadata Control
Volume:
60
0
test4test4202302021706.mp4
□ Mute ○ Solo
0
25
Rate:
100
0
00:00:37.123
Selection: 00:00:00.000 - 00:00:00.000 0
□ Selection Mode □ Loop Mode
PAIN_FACIAL_EXP
1.2
0.6
0.0
0.5286528
00:00:36.500
00:00:37.000
00:00:37.500
00:00:38.000
00:00:38.500
00:00:39.
EMOTION_IN_SPE
neutral
anger
joy
fear
neutral
di

FIGURE 2. The evaluation of emotional states related to automatic pain analysis performed through a model for facial expressions investigation [8]. The underlying Neural Network classifier for video analysis is based on the Shallow Network architecture. It consists of 17 input nodes (17 action units related to pain expressions) and a single output node that provides the probability of it being Pain versus No Pain. The Eudico Linguistic Annotator (ELAN) version 6.7 was implemented.

3. Results

3.1 Model performances

The model performed exceptionally well on the test set, effectively mitigating overfitting. The details of cross-validation accuracies are shown in Fig. 3.

In addition, the overall model performance provided 84% average accuracy across the folds of the cross-validation and demonstrated promising metrics values for all the classes (Table 2).

It is observed that neutral and sadness states achieved the highest metrics, with precision, recall, and F1-score ranging from 90% to 100%. Disgust and joy states reached the lowest values, probably due to the lower number of occurrences supporting the model. Finally, anger, fear and surprise states achieved the most stable and balanced metrics within 80%–87%, with surprise reaching a stable value of 84% for all the performance metrics considered.

3.2 Continuous audio-video analysis in cancer patients: use cases

3.2.1 Case study No. 1: clinical presentation of a patient without cancer pain

A 67-year-old female was referred to the Pain Medicine Unit due to chronic postoperative pain. The patient reported moderate pain, assessed using the Numeric Rating Scale (NRS 0–10) with ratings of 4–5. She also described a significant neuropathic component, including tactile allodynia, dysesthesias, and paresthesias, at the site of a surgical procedure performed approximately 2 years earlier for the removal of a lung lesion. A non-invasive method (percutaneous electrical nerve stimulation, PENS) was utilized to manage chronic post-thoracotomy pain. However, at the time of the video recording and interview, the patient reported being pain-free (NRS 0), with neuropathic symptoms nearly completely resolved, except for persistent paresthesias.

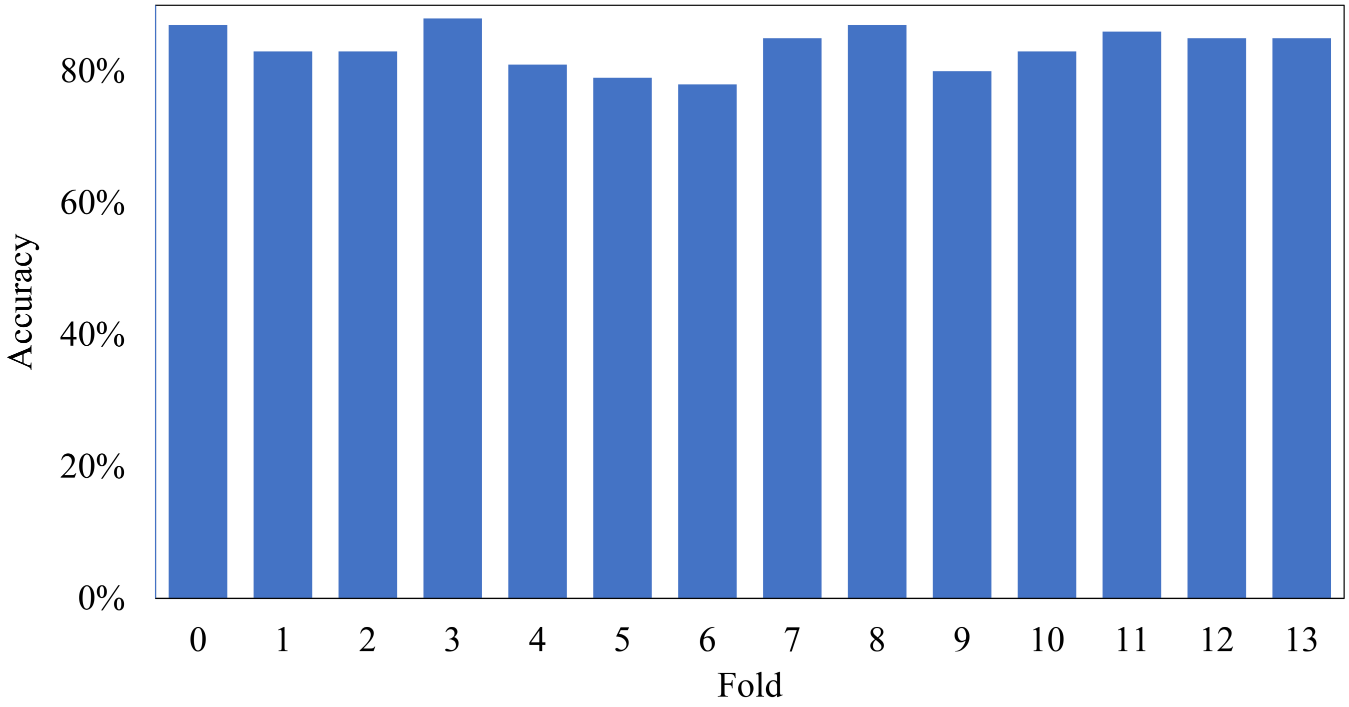

FIGURE 3. Cross-validation results. Accuracy obtained per each fold. The hyperparameter K was set to 15, a value that was confirmed through subsequent optimization processes. This resulted in the creation of 14 folds for training the network and a single fold for its validation.

TABLE 2. Overall model performances.

	Precision	Recall	F1-score	Support
Anger	0.87	0.80	0.83	25
Disgust	0.69	0.85	0.76	13
Fear	0.82	0.86	0.84	21
Joy	0.71	0.73	0.67	19
Sadness	0.90	1.00	0.95	26
Surprise	0.84	0.84	0.84	19
Neutral	1.00	0.88	0.93	24
Accuracy		0.84		147

3.2.2 Case study No. 2: clinical presentation of a patient with cancer pain

A 72-year-old male presented to the Pain Medicine Unit with severe pain (NRS 0–10, 8) attributed to multiple bone lesions resulting from prostate cancer. He underwent multimodal treatment involving pharmacological approaches (including opioids and adjuvant medications) and non-pharmacological interventions (such as palliative radiotherapy). Despite these interventions, the patient reported insufficient pain control, especially when triggered by movement (known as incident breakthrough cancer pain [29]).

3.2.3 Implementation of audio-video analysis and comparison of results

To implement the emotion recognition model in the framework of oncological pain assessment, we adopted both the previously developed APA model [8] for facial expression analysis from video recordings (binary classifier: pain/no pain) along the emotion recognition model from audio recordings. The models were applied together within the same framework of continuous audio-video analysis for both the use cases here described, with and without cancer pain respectively. As far as emotion recognition, we computed emotional states throughout the entire video, with each label assigned per 500 ms frame. In the analysis of the audio file, we excluded dialogue pauses and interviewer speech phases.

In addition to accurately identifying the presence or absence of pain in both patients through video analysis, the audio recording results also offered clear insights into the emotional characteristics of the two considered patients, as depicted in the following Fig. 4.

As from Fig. 4, the continuous analysis of the emotional status and pain expression in patients with and without cancer pain revealed marked differences in terms of emotion prevalence.

The neutral and surprise states were predominant for the patient who did not suffer from cancer pain, accounting for

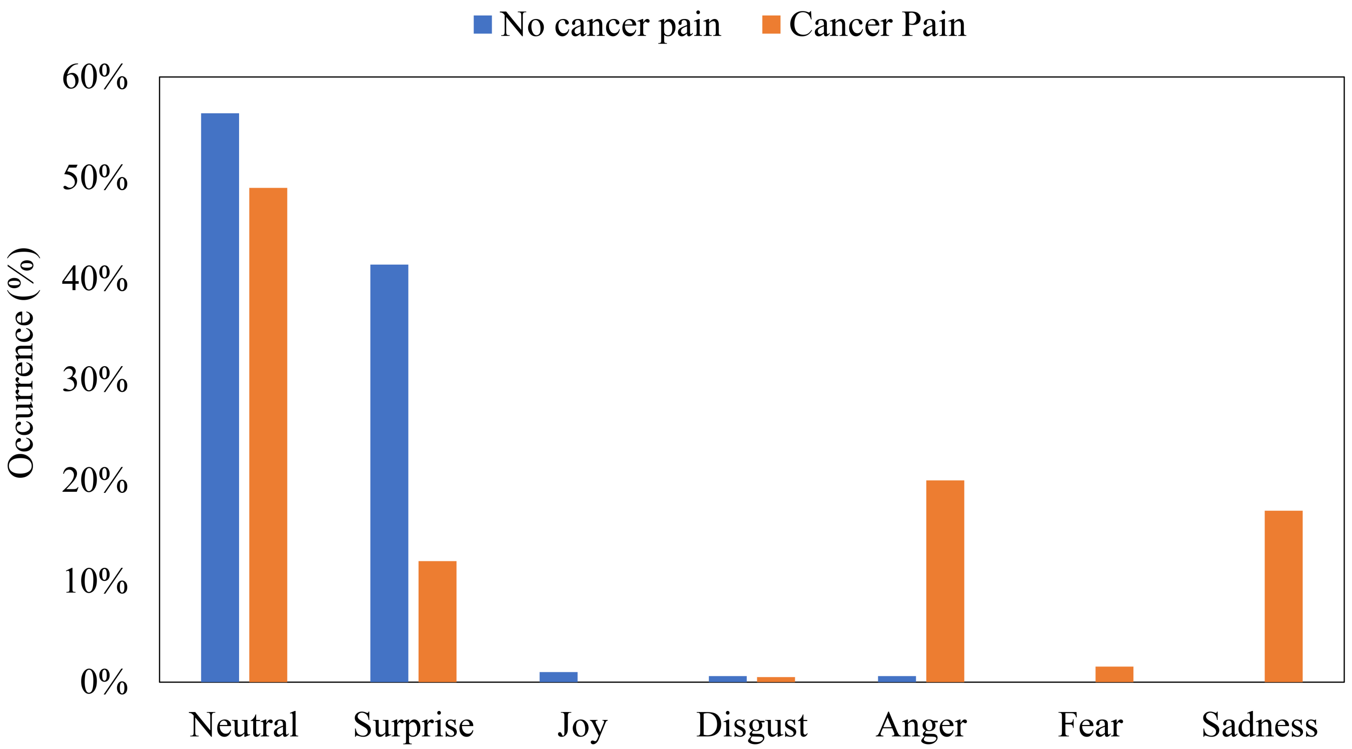

FIGURE 4. Comparison of emotional state analyses for the two examined use cases: with and without cancer pain.

56% and 41%, respectively. On the other hand, when the assessment framework was applied to the patient reporting cancer-related pain, the analysis highlighted a predominance of the neutral and anger states, accounting for 49% and 20%, respectively, followed by sadness (17%) and surprise (12%), which, in this case, was less represented (12%) compared to the case of non-cancer pain (41%).

The differences in terms of the percentage of emotions between the two presented clinical cases of non-cancer pain and cancer-related pain are reported in the following Table 3.

TABLE 3. Differences in the percentage of emotional state prevalence between the two presented clinical cases of non-cancer pain and cancer-related pain.

	Differences in emotional state prevalence
Neutral	+7%
Surprise	+29%
Joy	+1%
Disgust	0%
Anger	−19%
Fear	−2%
Sadness	−17%

Legend: Computes as Emotion in the "No cancer pain" Class—"Cancer pain" Class.

The proposed approach can detect the highest prevalence of negative emotions (anger, fear, sadness) in the case of cancer-related pain and the highest prevalence of neutral or positive emotions in the absence of cancer pain (neutral, surprise, joy). No differences were detected for the disgust state (Fig. 5).

4. Discussion

The implementation of AI methodologies in pain assessment, particularly in the context of cancer patients, holds significant promise for advancing our understanding and improving personalized pain management strategies [30]. AI and APA systems present innovative solutions for objective measurement [31]. Our study explored the application of AI in the form of computational language analysis for extracting emotional content from speech records. In particular, we focused on the simultaneous automatic emotion recognition and video analysis, aimed at an AI-based multifaceted approach to pain assessment. The implementation of APA processes aligns with the broader trend in the field directed toward leveraging AI strategies to enhance the efficiency and objectivity of pain assessment [7, 32, 33]. Speech analysis represents a valuable strategy for APA investigations. In response to advancements in signal processing and machine learning, researchers are now creating algorithms to automate the assessment of pain levels through speech. For example, Tsai *et al.* [34] embedded stacked bottleneck vocal features in a Long short-term memory (LSTM) architecture to detect pain, concentrating on prosodic signals within a subset of the triage dataset. Additionally, Li *et al.* [35] enhanced the variational acoustic model by introducing age and gender factors. Other projects focused on the analysis of bioacoustic pain markers related to the speech signal [36]. Therefore, we propose a step towards the fusion of different information channels aimed at pain assessment based on specific factors tied to the channel.

The central premise is that prosody and speaking features can significantly convey pain as a complex set of emotions, often referred to as the "Big Six" [9–11, 22]. Nevertheless, the

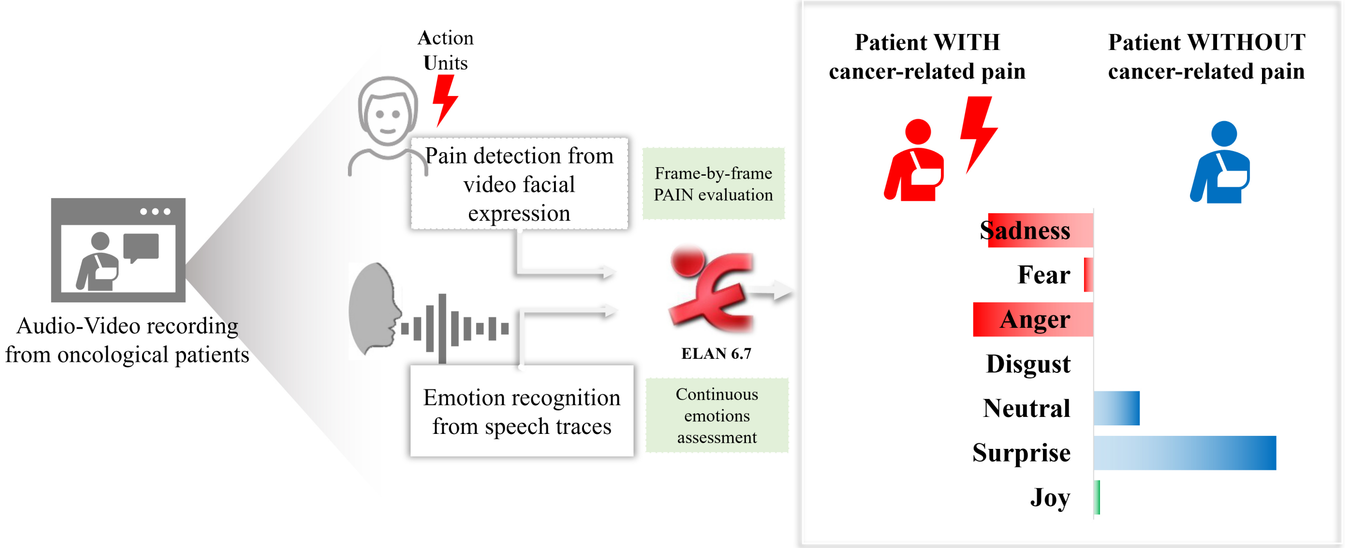
Action Units
Pain detection from video facial expression
Frame-by-frame PAIN evaluation
Audio-Video recording from oncological patients
Emotion recognition from speech traces
Continuous emotions assessment
ELAN 6.7
Patient WITH cancer-related pain
Patient WITHOUT cancer-related pain
Sadness
Fear
Anger
Disgust
Neutral
Surprise
Joy

FIGURE 5. Framework of the audio-video analyses in the two case studies. The video-audio analysis allows for simultaneous assessment of facial expressions of pain (action unit-based frame-by-frame assessment) and speech emotion recognition with the output of the full emotional status. For the assessment of audio and video tracks, we employed an Annotator (Eudico Linguistic Annotator, ELAN version 6.7). Although effective for ongoing evaluation, the tool itself is not an integrated platform.

difficulties hidden in this approach lie in the scarce availability of speech data annotated along any pain scale. Consequently, we are working towards the development of multimodal audio/video integrated models. The preliminary step involves defining the different frameworks to be implemented. Within our frameworks, the video analysis was performed by designing an AU-based classifier. The speech analysis was performed by designing a neural architecture. Therefore, by examining acoustic features, linguistic content and prosody, we aimed to capture emotional states that were subsequently associated with video-based APA analysis (binary classifier: Pain/No Pain) [8]. The developed neural network, trained on prosodic features extracted from speech signals, demonstrated promising results in emotion recognition, with accuracy reaching 0.84. Concerning the model's performances, we utilized accuracy as primary metric due to the balanced nature of the classes. This metric provides a holistic view of the s overall correctness in predicting emotions across all classes. Precision measures the proportion of correctly identified emotions among all instances where the model predicted a certain emotion. Recall, on the other hand, quantifies the ability of the model to correctly identify instances of a specific emotion out of all instances where that emotion truly occurred. The F1-score, which is the harmonic mean of precision and recall, gives a balanced assessment of the model's performance, particularly useful when dealing with imbalanced datasets or when precision and recall need to be considered together [37].

An important aspect is the definition of the set of prosodic features to include for the training and testing of the model. The number of prosodic features that can be extracted from speech is not fixed or standardized [38]. It can vary depending on the specific research or application, and different studies or systems may use different sets of features. Researchers might choose a subset of relevant features based on the goals of their study or the requirements of their application. For instance, in some studies focused on emotion recognition, researchers may extract features related to pitch, duration, and intensity variations associated with different emotional states. In contrast, applications like speaker identification might prioritize features related to pitch patterns, rhythm, and voice quality [39, 40]. For this reason, we have opted to incorporate a diverse array of features based on the Librosa library.

Given these preliminary results, the continuous evaluation revealed different emotional states, showcasing the potential of our approach for real-time assessment and personalized pain management pathways. The diversified emotional states and pain features unveiled through continuous evaluation underscore the versatility and potential applications of our method, suggesting its adaptability in dynamically responding to individuals' changing pain experiences and needs.

To demonstrate the practical applications of our SER framework, we have selected two exemplary cases for analysis. In the first case, involving a patient who did not report pain, our assessment revealed predominant emotional states of neutrality and surprise. Since the patient did not describe pain issues, feelings of neutrality might indicate a lack of emotional distress related to discomfort or suffering. The presence of surprise could be linked to unexpected events or interactions during the assessment process. Conversely, in the second case, our analysis identified a significant component of anger and sadness. These emotions may stem from the patient's experience of pain and associated distress. Chronic pain can evoke feelings of frustration, helplessness and sadness [41]. Additionally, the patient's ongoing struggle with poorly controlled breakthrough cancer pain despite multimodal treatment strategies, as mentioned in the case description, could contribute to heightened emotional responses such as anger and sadness [29]. These emotional states may also reflect the patient's psychological and emotional resilience or vulnerability, as well as their coping mechanisms and support systems. Factors such as past experiences, personality traits, and social support networks can influence how individuals perceive and respond to pain and distress [42]. These findings underscore the complex interplay between physical symptoms, psychological factors, and emotional well-being in cancer patients, highlighting the importance of holistic and patient-centered approaches to pain management [41–43].

Finally, from our initial findings emerges that a crucial aspect demanding in-depth exploration is the correlation between facial expressions and emotional responses through computational methodologies. Physiologically, emotions manifest through different facial expressions. For example, surprise may be expressed as widened eyes, raised eyebrows, and open mouth. These physical reactions are often involuntary and are part of the body's automatic response to unexpected stimuli. Despite the conducted experiments within the medical domain [44], particularly in the field of pain medicine, this realm remains largely untapped for comprehensive investigation and understanding. For instance, an intriguing avenue of research involves studying the correlations between emotion circuits and the neurobiology of chronic pain [45].

This investigation acknowledges several limitations that warrant consideration. An important limitation is the model's lack of verification across a large population sample. It is emphasized that the objective of the paper was not the generalization and validation of the model across a large sample; instead, it focused on the potential preliminary application of the proposed framework in a clinical setting of oncological patients. In defining the methods, we selected two case studies of oncological patients, analogously to what has been done in another explorative study on APA through biosignal analyses [46]. Although the obtained results may be potentially influenced by cultural and language biases and lack external validation, they show promising performance metrics of the emotion recognition classifier and possible applicability in clinical settings. This encourages further assessment to investigate the reliability and validity of the model on a larger scale, which will be addressed in future works.

Concerning model building, a key constraint revolves around the utilization of datasets not explicitly designed for the study of cancer pain. The predictor expresses outputs (*i.e.*, emotion) for audio traces that reflect what it has learned from the EMOVO dataset. This aspect could significantly impact the reliability of the outcomes, as the datasets might lack essential variables for a thorough analysis of cancer pain phenomena. Consequently, caution is advised in concluding the results, given that this main limitation may influence the generalizability and applicability of the findings. Future research endeavors should prioritize the collection of more targeted and comprehensive datasets to overcome this limitation, thereby enhancing the accuracy and reliability of the results. Furthermore, the model employed in the study has limitations in conducting temporal analysis as it focuses on a single time instant without considering contextual parameters.

In the context of facial expression for APA, the pain phenomenon was evaluated using a binary classification system, distinguishing between its presence and absence [8]. However, it is essential to acknowledge that this study did not delve into more granular analyses regarding the various types or gradations of pain. Another limitation of the analysis is the overall quantization of emotions, considering the entire video. For increased reliability, it would be advisable to segment the video into smaller sections instead of computing based on the overall duration. However, this limitation is also associated with the need for precise labeling within the context of audio and video content. These limitations suggest the necessity for a multimodal model to enhance the accuracy of the study's findings and address gaps in the analysis. Significant improvements could involve procedures to label text and integrate various behavioral analyses, such as facial emotion recognition [47–49] as well as to adopt further algorithms and models to cluster [50–52] and classify [53–55] the patients based on the facial, vocal patterns and additional multimodal patterns. Our forthcoming research in the field of APA will delve deeper into this aspect.

5. Conclusions

In APA research, exploring automatic emotion recognition offers valuable insights. These initial findings indicate the promising application of AI models to continuously estimate emotional states from video recordings. This strategy has the potential to complement other APA approaches as part of a multimodal analysis. Further research is warranted to refine this process, particularly in integrating this data with multiparameter analyses, including text analysis, and multimodal physiological parameters. In the subsequent stages, the model will need to anticipate the integration of individual approaches and ultimately its practical implementation.

AVAILABILITY OF DATA AND MATERIALS

The raw data can be obtained on request from the corresponding author. The dataset implemented for the binary classifier is available at: https://doi.org/10.5281/zenodo.7557362 [28].

AUTHOR CONTRIBUTIONS

MC, FC and VNV—designed the research study. MC, FM, MI and VNV—performed the research. AC, SB, FSa and VC—provided help and advice on the clinical evaluations. AV, MI, AMP, JM, VB and FSe—analyzed the data. MC, EGB and OP—wrote the manuscript. All authors contributed to editorial changes in the manuscript. All authors read and approved the final manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The research adhered to the principles of the Declaration of Helsinki, and written informed consent was obtained from all patients. The Medical Ethics Committee of the Istituto Nazionale Tumori, Fondazione Pascale (Napoli, Italy), approved the PASCALE study (Pain ASsessment in CAncer Patients by Machine Learning), under the protocol code 41/20 Oss (26 November 2020). (ClinicalTrials.gov Identifier: NCT04726228). Patient consent was acquired for the study and scientific divulgation of personal identifiable information.

ACKNOWLEDGMENT

The authors thank Cristina Romano and Mariacinzia Grizzuti from the National Cancer Institute of Naples for the provided administrative and technical support.

FUNDING

This research received no external funding.

CONFLICT OF INTEREST

JM is a co-founder and shareholder of Callisia srl University Spin-off at Università Politecnica delle Marche developing a smart bracelet collecting patient data intelligently for real-time visualization and data analysis. The rest of the authors have no relevant financial or non-financial interests to disclose. Marco Cascella is serving as one of the Editorial Board members of this journal. We declare that Marco Cascella had no involvement in the peer review of this article and has no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to MAM.

REFERENCES

[1] van den Beuken-van Everdingen MH, Hochstenbach LM, Joosten EA, Tjan-Heijnen VC, Janssen DJ. Update on prevalence of pain in patients with cancer: systematic review and meta-analysis. Journal of Pain and Symptom Management. 2016; 51: 1070–1090.e9.

[2] Cascella M, Vittori A, Petrucci E, Marinangeli F, Giarratano A, Cacciagrano C, *et al*. Strengths and weaknesses of cancer pain management in Italy: findings from a nationwide SIAARTI survey. Healthcare. 2022; 10: 441.

[3] Caraceni A, Shkodra M. Cancer pain assessment and classification. Cancers. 2019; 11: 510.

[4] Giordano V, Deindl P, Olischar M. The limitations of pain scales—reply. JAMA Pediatrics. 2020; 174: 623.

[5] Baamer RM, Iqbal A, Lobo DN, Knaggs RD, Levy NA, Toh LS. Utility of unidimensional and functional pain assessment tools in adult postoperative patients: a systematic review. British Journal of Anaesthesia. 2022; 128: 874–888.

[6] Aung MSH, Kaltwang S, Romera-Paredes B, Martinez B, Singh A, Cella M, *et al*. The automatic detection of chronic pain-related expression: requirements, challenges and the multimodal *EmoPain* dataset. IEEE Transactions on Affective Computing. 2016; 7: 435–451.

[7] Gkikas S, Tsiknakis M. Automatic assessment of pain based on deep learning methods: a systematic review. Computer Methods and Programs in Biomedicine. 2023; 231: 107365.

[8] Cascella M, Vitale VN, Mariani F, Iuorio M, Cutugno F. Development of a binary classifier model from extended facial codes toward video-based pain recognition in cancer patients. Scandinavian Journal of Pain. 2023; 23: 638–645.

[9] Gilam G, Gross JJ, Wager TD, Keefe FJ, Mackey SC. What is the relationship between pain and emotion? Bridging Constructs and Communities. Neuron. 2020; 107: 17–21.

[10] Asghar A, Sohaib S, Iftikhar S, Shafi M, Fatima K. An Urdu speech corpus for emotion recognition. PeerJ Computer Science. 2022; 8: e954.

[11] Atmaja BT, Sasou A. Sentiment analysis and emotion recognition from speech using universal speech representations. Sensors. 2022; 22: 6369.

[12] Deshpande G, Schuller BW, Deshpande P, Joshi AR. Automatic breathing pattern analysis from reading-speech signals, 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). Sydney, Australia 24th–27th July 2023. 2023.

[13] Chowdhary KR. Natural language processing. In Chowdhary KR (ed.) Fundamentals of Artificial Intelligence (pp. 603–649). Springer: New Delhi. 2020.

[14] Srinivasan R, Subalalitha CN. Sentimental analysis from imbalanced code-mixed data using machine learning approaches. Distributed and Parallel Databases. 2023; 41: 37–52.

[15] Ku PKM, Vlantis AC, Yeung ZWC, Ho OYM, Cho RHW, Lee AKF, *et al*. Perceptual voice and speech analysis after supraglottic laryngeal closure for chronic aspiration in head and neck cancer. Laryngoscope. 2021; 131: E1616–E1623.

[16] Husain M, Simpkin A, Gibbons C, Talkar T, Low D, Bonato P, *et al*. Artificial Intelligence for Detecting COVID-19 with the aid of human cough, breathing and speech signals: scoping review. IEEE Open Journal of Engineering in Medicine and Biology. 2022; 3: 235–241.

[17] Kerdvibulvech C, Chen L. The power of augmented reality and artificial intelligence during the COVID-19 outbreak, HCI International 2020—late breaking papers: multimodality and intelligence. Copenhagen, Denmark, July 19–24, 2020. Springer International Publishing. 2020.

[18] Xie X, Cai H, Li C, Wu Y, Ding F. A voice disease detection method based on MFCCs and shallow CNN. To be published in Journal of Voice. 2023. [Preprint].

[19] Koops S, Brederoo SG, de Boer JN, Nadema FG, Voppel AE, Sommer IE. Speech as a biomarker for depression. CNS & Neurological Disorders Drug Targets. 2023; 22: 152–160.

[20] Yokoi K, Iribe Y, Kitaoka N, Tsuboi T, Hiraga K, Satake Y, *et al*. Analysis of spontaneous speech in Parkinson's disease by natural language processing. Parkinsonism & Related Disorders. 2023; 113: 105411.

[21] Costantini G, Iaderola I, Paoloni A, Todisco M. EMOVO corpus: an Italian emotional speech database, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland. May 26–31, 2014. 2014.

[22] Cowie R, Cornelius RR. Describing the emotional states that are expressed in speech. Speech Communication. 2003; 40: 5–32.

[23] Khare Y. Hands-on-guide to Librosa for handling audio files. 2024. Available at: `https://www.analyticsvidhya.com/blog/2024/01/hands-on-guide-to-librosa-for-handling-audio-files/` (Accessed: 04 January 2023).

[24] Jeevitha M. Exploring Librosa: a comprehensive guide to audio feature extraction from WAV files. 2023. Available at: `https://www.linkedin.com/pulse/exploring-librosa-comprehensive-guide-audio-feature-extraction-m/` (Accessed: 04 January 2023).

[25] Kingma DP, Ba LJ. Adam: a method for stochastic optimization, International Conference on Learning Representations (ICLR). San Diego, May 7–9, 2015.

[26] Mende-Siedlecki P, Qu-Lee J, Lin J, Drain A, Goharzad A. The Delaware pain database: a set of painful expressions and corresponding norming data. PAIN Reports. 2020; 5: e853.

[27] Lucey P, Cohn JF, Prkachin KM, Solomon PE, Matthews I. Painful data: the UNBC-McMaster shoulder pain expression archive database, 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG), Santa Barbara, CA, USA. 2011.

[28] Cascella M. Dataset for binary classifier_Pain. 2023. Available at: `https://doi.org/10.5281/zenodo.7557362` (Accessed: 04 January 2023).

[29] Cuomo A, Cascella M, Forte CA, Bimonte S, Esposito G, De Santis S, *et al*. Careful breakthrough cancer pain treatment through rapid-onset transmucosal fentanyl improves the quality of life in cancer patients: results from the BEST multicenter study. Journal of Clinical Medicine. 2020; 9: 1003.

[30] Zhang M, Zhu L, Lin SY, Herr K, Chi CL, Demir I, *et al*. Using artificial

intelligence to improve pain assessment and pain management: a scoping review. Journal of the American Medical Informatics Association. 2023; 30: 570–587.

[31] Cascella M, Schiavo D, Cuomo A, Ottaiano A, Perri F, Patrone R, *et al*. Artificial intelligence for automatic pain assessment: research methods and perspectives. Pain Research and Management. 2023; 2023: 6018736.

[32] Nagireddi JN, Vyas AK, Sanapati MR, Soin A, Manchikanti L. The analysis of pain research through the lens of artificial intelligence and machine learning. Pain Physician. 2022; 25: E211–E243.

[33] Sankaran R, Kumar A, Parasuram H. Role of artificial intelligence and machine learning in the prediction of the pain: a scoping systematic review. Proceedings of the Institution of Mechanical Engineers. Part H, Journal of Engineering in Medicine. 2022; 236: 1478–1491.

[34] Tsai FS, Weng YM, Ng CJ, Lee CC. Embedding stacked bottleneck vocal features in a LSTM architecture for automatic pain level classification during emergency triage, Proceedings of the 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). San Antonio, TX, USA, 23–26 October 2017. 2017.

[35] Li JL, Weng YM, Ng CJ, Lee CC. Learning conditional acoustic latent representation with gender and age attributes for automatic pain level recognition, Proceedings of the Interspeech 2018. Hyderabad, India, 2–6 September 2018. 2018.

[36] Schneiders E, Williams J, Farahi A, Seabrooke T, Vigneswaran G, Bautista JR, *et al*. TAME pain: trustworthy assessment of pain from speech and audio for the empowerment of patients, Proceedings of the First International Symposium on Trustworthy Autonomous Systems. New York, July 2023. 2023.

[37] Bellini V, Cascella M, Cutugno F, Russo M, Lanza R, Compagnone C, *et al*. Understanding basic principles of artificial intelligence: a practical guide for intensivists. Acta Biomedica. 2022; 93: e2022297.

[38] Dahan D. Prosody and language comprehension. Wiley Interdisciplinary Reviews: Cognitive Science. 2015; 6: 441–452.

[39] Speer S, Blodgett A. Prosody. In Traxler M, Gernsbacher MA (eds.) Handbook of psycholinguistics (pp. 505–537). 2nd edn. Academic Press: San Diego, CA. 2006.

[40] Carlson K. How prosody influences sentence comprehension. Language and Linguistics Compass. 2009; 3: 1188–1200.

[41] Schiavo D, Cumo A, Nocerino D, Monaco F, Cascella M. The body of pain. The experience of pain in the cancer patient. Recenti Progressi in Medicina. 2023; 114: 410–413. (In Italian)

[42] Siler S, Borneman T, Ferrell B. Pain and suffering. Seminars in Oncology Nursing. 2019; 35: 310–314.

[43] Erol O, Unsar S, Yacan L, Pelin M, Kurt S, Erdogan B. Pain experiences of patients with advanced cancer: a qualitative descriptive study. European Journal of Oncology Nursing. 2018; 33: 28–34.

[44] Venkitakrishnan S, Wu YH. Facial Expressions as an index of listening difficulty and emotional response. Seminars in Hearing. 2023; 44: 166–187.

[45] Zhang H, Chen X, Chen S, Li Y, Chen C, Long Q, Yuan J. Facial expression enhances emotion perception compared to vocal prosody: behavioral and fMRI studies. Neuroscience Bulletin. 2018; 34: 801–815.

[46] Cascella M, Vitale VN, D'Antò M, Cuomo A, Amato F, Romano M, *et al*. Exploring biosignals for quantitative pain assessment in cancer patients: a proof of concept. Electronics. 2023; 12: 3716.

[47] De Carolis B, Macchiarulo N, Palestra G, De Matteis AP, Lippolis A. FERMOUTH: facial emotion recognition from the MOUTH region. In Foresti GL, Fusiello A, Hancock E (eds.) Image Analysis and Processing—ICIAP 2023. Springer: Cham. 2023.

[48] Castellano G, De Carolis B, Macchiarulo N. Automatic facial emotion recognition at the COVID-19 pandemic time. Multimedia Tools and Applications. 2023; 82: 12751–12769.

[49] Samadiani N, Huang G, Cai B, Luo W, Chi CH, Xiang Y, *et al*. A review on automatic facial expression recognition systems assisted by multimodal sensor data. Sensors. 2019; 19: 1863.

[50] Hajarolasvadi N, Demirel H. 3D CNN-based speech emotion recognition using K-means clustering and spectrograms. Entropy. 2019; 21: 479.

[51] Seyala N, Abdullah SN. Cluster analysis on longitudinal data of patients with kidney dialysis using a smoothing cubic B-spline model. International Journal of Mathematics, Statistics, and Computer Science. 2024; 2: 85–95.

[52] Kunz M, Lautenbacher S. The faces of pain: a cluster analysis of individual differences in facial activity patterns of pain. European Journal of Pain. 2014; 18: 813–823.

[53] Aung MS, Kaltwang S, Romera-Paredes B, Martinez B, Singh A, Cella M, *et al*. The automatic detection of chronic pain-related expression: requirements, challenges and the multimodal EmoPain dataset. IEEE transactions on affective computing. 2015; 7: 435–451.

[54] Cascella M, Semeraro F, Montomoli J, Bellini V, Piazza O, Bignami E. The breakthrough of large language models release for medical applications: 1-year timeline and perspectives. Journal of Medical Systems. 2024; 48: 22.

[55] Jang EH, Rak B, Kim SH, Sohn JH. Emotion classification by machine learning algorithm using physiological signals, 2012 IACSIT Hong Kong Conferences. Singapore, 26–27th October 2012. IACSIT Press. 2012.

How to cite this article: Marco Cascella, Francesco Cutugno, Fabio Mariani, Vincenzo Norman Vitale, Manuel Iuorio, Arturo Cuomo, *et al*. AI-based cancer pain assessment through speech emotion recognition and video facial expressions classification. Signa Vitae. 2024; 20(12): 28-38. doi: 10.22514/sv.2024.153.