Intelligent Systems with Applications 17 (2023) 200171

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

A systematic survey on multimodal emotion recognition using learning algorithms

Naveed Ahmed *, Zaher Al Aghbari, Shini Girija

Department of Computer Science, University of Sharjah, Sharjah 27272, United Arab Emirates

ARTICLE INFO

Keywords:
Multimodal emotion recognition
Virtual reality
Deep learning
Machine learning
Data fusion methods
Fine grained emotions

ABSTRACT

Emotion recognition is the process to detect, evaluate, interpret, and respond to people's emotional states and emotions, ranging from happiness to fear to humiliation. The COVID- 19 epidemic has provided new and essential impetus for emotion recognition research. The numerous feelings and thoughts shared and posted on social networking sites throughout the COVID-19 outbreak mirrored the general public's mental health. To better comprehend the existing ecology of applied emotion recognition, this work presents an overview of different emotion acquisition tools that are readily available and provide high recognition accuracy. It also compares the most widely used emotion recognition datasets. Finally, it discusses various machine and deep learning classifiers that can be employed to acquire high level features for classification. Different data fusion methods are also explained in detail highlighting their benefits and limitations.

1. Introduction

Emotion is crucial in interpersonal relationships, knowledge insight, perception, and other aspects of life (Baltrušaitis et al., 2018; Abdullah et al., 2021). Understanding emotions has become essential for humans' day-to-day functioning since emotion acquisition and experiences are essential for inter-communication in the social settings. Emotion recognition has emerged as a critical area of study in human-computer interactions, with an increasing demand for automated emotion identification systems (Xie & Guan, 2013; Alswaidan & Menai, 2020). Speech, text, facial cues, and electroencephalogram (EEG)-based brain waves have all been used to study emotion recognition (Dadebayev et al., 2021). Given the rising and diverse use of human-computer interaction, emotion recognition technologies offer a way to foster agreeable or intuitive user interactions. Emotion detection and recognition (EDR) market value was USD 19.87 million in 2020, and it's expected to increase to USD 52.86 million by 2026, with a CAGR of 18.01 percent over the forthcoming years (2021–2026) (Mordorintelligence, 2021).

Emotional signals from many modalities can be utilized to anticipate a subject's emotional state. The single modal model, on the other hand, struggles to assess the consumer's emotional state. We can't tell if someone is emotional only by looking at a specific object or event in front of us (Xie & Guan, 2013; Tang et al., 2017). This is one of the key reasons why emotion recognition must be considered a multimodal problem. The process of creating emotion-specific features is complicated by multiple factors that arise from non-linear interactions between different modalities, multi-dimensional data, and the evidence that individuals express emotions in different ways. Deep networks and machine learning have proven to bypass such limitations by detecting complex non-linear feature correlations in multimodal data (Zhang et al., 2020).

The two most critical procedures in emotion recognition are feature extraction and classification. The main feature classification algorithms that are now employed for greater recognition accuracy are deep learning, artificial neural networks, and machine learning techniques (Georgescu & Ionescu, 2019; Imani & Montazer, 2019). Traditional feature engineering and machine learning techniques may struggle to extract complicated and nonlinear patterns from multivariate time series information. Moreover, picking the most important characteristics from a huge feature set is critical, and dimensionality reduction techniques will be required. Furthermore, calculating feature extraction and selection takes a lengthy time. For example, when feature dimensionality increases, the computing overhead of feature selection may grow dramatically (Tang et al., 2017). Deep Learning (DL) technologies such as convolutional neural network (CNN), recurrent neural network (RNN), and autoencoder have resulted in a marked improvement in algorithms in almost all of the fields of computing, especially, natural

* Corresponding author.
E-mail address: nahmed@sharjah.ac.ae (N. Ahmed).

https://doi.org/10.1016/j.iswa.2022.200171
Received 27 September 2022; Received in revised form 27 November 2022; Accepted 24 December 2022
Available online 3 January 2023

language processing (NLP), computer vision, machine translation, audio recognition etc. (Liu et al., 2021). Because of its capacity to provide high-level data abstraction, DL has recently been used to build customizable structures for emotion recognition. In DL techniques, deep neural networks are employed to collect distinguishing traits from high-level data representation.

Most emotion acquisition systems evoke emotional states using 2D non-immersive visual tools, e.g., images or videos. Immersive virtual reality, on the other hand, is gaining popularity in emotion research because it aids researchers to replicate situations in regulated laboratory circumstances with a strong sensation of presence and engagement. Furthermore, it can be integrated with machine-learning techniques while utilizing its implicit measurements to explore solutions that can have a broad influence across a wide range of study fields, providing new opportunities for scientists (Marín-Morales et al., 2020). Sensor data, such as accelerometer and gyroscope data, has a wealth of information that can be utilized to assess the user's social habits, such as their physical activity, social communications, and location (Piskioulis et al., 2021).

The emotions being studied are usually straightforward: they are either positive, negative, or neutral. Positive emotions can boost subjective well-being and enhance emotional well-being, whereas chronic negative emotions can harm people's health and well-being, as well as their work prospects. Neutral feelings are comparatively bland, with neither a distinguishing pain nor a distinct pleasure (Long et al., 2021). Fine-grained emotion recognition essentially allows to figure out intensity of emotion (for example, the difference between a slightly positive experience and a very positive experience).

This paper presents a systematic literature review on the various emotion acquisition methods, emotion classification algorithms and fusion techniques produced from 2013 to 2021. We incorporate the analysis of fine-grained emotions in detail. The current available datasets for emotion recognition are discussed in depth. Finally, we reveal the findings of the existing research papers and identify the research gaps and opportunities for the future research scope. This study focuses mostly on machine learning and deep learning algorithms because they have proven to be an efficient way to deal with the time-consuming preprocessing and feature extraction processes in multimodal emotion recognition. They perform well in terms of classification, and they are more accurate at detecting emotions than conventional methods. To evaluate human emotion states, new technical techniques such as artificial intelligence, federated learning, and transfer learning can be used. However, these studies are still in the early stages and require more research to address a number of issues, including privacy and error rates.

The contributions of this paper are:

The rest of the paper is organized as follows: The background of emotion recognition and its applications are discussed in Section 2. In Section 3, the related work and survey papers in this field are presented. Section 4 explains the review method used for this research paper. Section 5 discusses emotion and its fine-grained forms that are relevant to emotion recognition, Section 6 discusses different tools and technologies that are employed in data acquisition for emotion recognition, and presents different datasets for emotion recognition in detail. Section 7 reviews the different classifiers for emotion recognition. Section 8 illustrates the various data fusion methods for emotion recognition. Section 9 presents discussion and the future directions of research. Finally, Section 10 summarizes the outcome ad the conclusion of this research.

2. Background

This section presents the authors' overview on the present state of emotion recognition, multimodal emotion recognition and broad review of the variety of its most important applications.

2.1. Emotion recognition

Emotion detection is a technique for identifying and recognizing human emotions that employs technical skills such as facial recognition, speech recognition, voice recognition, biosensors, deep learning, and pattern recognition (Mordorintelligence, 2021; Sahoo & Routray, 2016). Initial emotion interpretation based on self-assessment questionnaire, tone of voices, or facial expression is arbitrary. Moreover, it degrades accuracy of recognition due to biases in language, gestures, or expressions. Individuals could employ purposeful activities to mask or simulate their true feelings. Such human-based emotion identification could be prone to human error. As a result, physiological signal measures such as heart response, electromyogram, and galvanic skin response are increasingly used in emotion recognition techniques (Tang et al., 2021).

Modalities reflect a variety of information sources that can propose different types of information and different points of view. Emotion recognition methods can be classified as either unimodal or multimodal. Unimodal emotion recognition identifies human emotions using a single modality e.g., face, text, EEG, speech or image. The main limitation of the unimodal is that the chosen modality can have weaknesses in specific contexts. Every modality has benefits and drawbacks: In low-light settings, audio is preferable to video, while text is occasionally preferable to audio for forecasting valence dimension (Salazar et al., 2021).

To get a global picture, the system needs all of the accessible data. Distinct modalities can predict different emotional states in multimodal emotion identification. The aim behind the multimodal emotion recognition is to use the combined knowledge to strengthen the shortcomings of each modality, resulting in a more resilient system (Tashu et al., 2021).

2.2. Multimodal emotion recognition

Multimodal emotion recognition mimics the predictive power of humans, which combines the biological and behavioral features. Unlike unimodal emotion recognition, multimodal emotion recognition requires many data sources to be processed simultaneously. There is a lot of variety in prediction accuracy even within multimodal techniques, necessitating the development of robust approaches (Cimtay et al., 2020).

Song & Kim (2021) identified hidden emotions in which the facial expression retains a neutral emotion while the bio-signal is active. In addition, their method used CNN to analyze feature maps of video and EEG inputs in order to effectively discern hidden emotions. Initially, the subject's internal and outer emotional features were investigated utilizing multimodal signals when he or she intentionally hides a feeling. Then, using CNNs, which have strong cognitive abilities, a method is devised for detecting hidden emotions based on the mismatch between internal and external emotions. As a result of this research, the technology for profoundly comprehending inner emotions has advanced (Song & Kim, 2021). The different steps in multimodal emotion recognition are illustrated in Fig. 1.

Povolny et al. (2016) suggested a multimodal emotion identification system that used voice and video features. Bottle-neck (BN) characteristics created from a restricted hidden layer of a neural network trained toward phonetic targets supplemented the available material in audio. They used a convolutional neural network (CNN) trained to locate facial landmarks to supplement the baseline data in video. Geometrical information is blended with appearance information in these activations. Experimentation using text-based features generated from an automatic speech recognition system is also carried out. The main advantage is that BN features (as well as generic feature extraction approaches based on deep neural networks) have been found to be particularly effective in various areas of speech processing. The method's drawback is that it does not contain physiological features, hence recognition accuracy will be less than ideal (Povolny et al., 2016).

Poria et al. (2016) performed sentiment analysis and multimodal

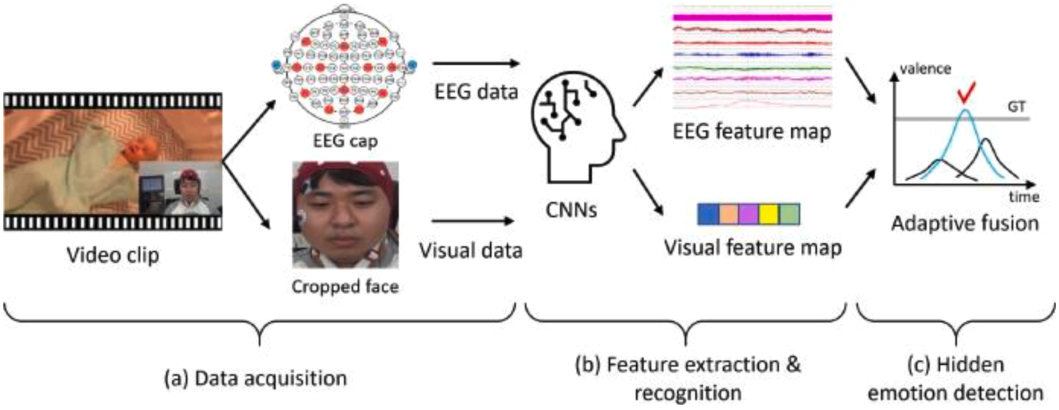

Fig. 1. Steps in multimodal emotion recognition (Song & Kim, 2021).

emotion recognition that combined voice, text, and facial gestures. They extracted visual and textural features using a novel temporal deep CNN. They combined extracted multi-modal heterogeneous data (video, audio, and text) using multiple kernel learning (MKL). RNNs are used to capture the spatio-temporal features embedded in the video sequences. They combine features from multiple modalities using MKL to create the CRMKL model, which combines RNN, CNN, and MKL. The main benefit is that merging RNN and deep CNN models increased speed and efficiency significantly. The limitation is that it does not contain physiological features, hence recognition accuracy will not be optimal (Poria et al., 2016). To boost emotion recognition, Lu et al. (2015) used eye movements and electroencephalogram (EEG) modalities in emotion recognition. The inherent patterns of sixteen eye movements related to emotions are identified for negative, positive, and neutral emotions. According to several modality fusion approaches, both EEG and eye movements complement emotion recognition. The key benefit of this technology is that it may be used to properly represent users' subconscious behaviors and cognitive processes when they are stimulated with different emotions by combining eye movement and EEG data. The shortcoming of this method is that they didn't test the model's applicability in real-world scenarios (Lu et al., 2015).

Cimtay et al. (2020) proposed the use of a variety of modalities, such as galvanic skin response (GSR), facial expressions, and electroencephalogram (EEG), to recognize emotions. Feature fusion on EEG and GSR was used to determine the state of arousal. A late merger of EEG, GSR, and facial modalities was the final stage. When the genuine emotional state is dominating, or when it is disguised due to normal misleading face actions, the proposed model can detect it. The model is subject-independent and as it does not require any feature extraction, it is an excellent real-time system. As a result, there is more flexibility, and the overall processing burden is reduced. The proposed methodology is notably useful in recognizing the precise emotional state when natural misleading facial expressions are present (Cimtay et al., 2020).

2.3. Applications of emotion recognition

Emotion recognition approaches can be used in video gaming, business intelligence and retailers, smart devices, vehicle automation, e-learning, social robots, and virtual assistants to create and automate individualized interfaces or sub-component technologies for bigger systems. The main application areas are discussed below.

- Online learning: During Covid pandemic, online platforms facilitate learning without any disruption and emotion recognition technology helps to identify the emotion states of the students in real time. This information can be used to plan the lesson content according to the differentiated learning abilities of children (Cen et al., 2016).
- Retail Market: Emotient, a company that specializes in emotion recognition, has been aiming to break into the retail market with the help of Google Glass. Retailers will be able to see what their consumers are thinking and feel, with the goal of leveraging this information to improve the in-store experience. Its goal is to compare data from standard satisfaction assessments with data from emotion recognition tools to see if emotion recognition may provide a more detailed picture or even replace satisfaction metrics (Forsberg, 2017).
- Medical applications: Automated IoMT systems are used in surveillance settings to collect multimodal information about patients. The latest information and data from respective portal and publications are verified for additional study of different emotion analysis methods. The suggested unique IoMT quality service provides real time monitoring and emotion-aware decision-making throughout the COVID-19 pandemic, greatly simplifying the procedure and providing continuous emotion-aware medical services (Zhang et al., 2020).
- Gaming: The study of emotional recognition in games can help to keep users engaged and improve their gaming experience. Automatic emotion recognition for game users is required for this purpose in order to sustain their interest without interfering with their gaming process (Du et al., 2020). Virtual reality consoles give their users the experience as if they are physically within the game, despite display resolution and delay issues. Emotion detection is an attractive addition to VR since it might allow developers to create games that react to a subscriber's sentiments instantaneously (Techxplore, 2021).
- Social robots: Emotion recognition garnered an interest in the larger disciplines of human–machine interaction and sentiment analysis. Social robots can infer and interpret human emotions, making them more effective in human interactions. They should be able to interpret human emotions and change their behavior accordingly, resulting in an acceptable response to those feelings (Spezialetti et al., 2020).

3. Related work

In this section, the authors explore existing surveys in emotion recognition field, by presenting the taxonomy followed, methods used and its limitations.

Several works have investigated methods for achieving high accuracy of emotion recognition. Older survey papers published in 2014: Wu et al. (2014) and Kołakowska et al. (2014) reviewed recent research on emotion recognition but did not adopt a comprehensive approach. The most relevant current emotion recognition research survey studies, to our knowledge, are Sharma & Dhall (2021), Imani & Montazer (2019) and Jam et al. (2021). Sharma & Dhall (2021) clearly pointed out the benefits of emotion recognition on visual, speech, text and EEG. The two fusion methods are thoroughly explained, and many datasets are investigated. But only unimodal methods are considered in the survey. In Imani & Montazer (2019), the authors evaluated numerous emotion recognition technologies used in e-learning systems and listed their benefits and drawbacks. Yet tools used for emotion acquisition are not mentioned. Jam et al. (2021) mainly focused on identifying social and emotional feature of human-robot interactions using deep learning methods. More research into the correlation of detailed descriptions of emotions to emojis is needed in the survey.

Alswaidan & Menai (2020) presented a detailed taxonomy of text emotion recognition methods, which also provided an in-depth explanation of main features and different available approaches. Various explicit and implicit emotion recognition techniques in text are explained and the different approaches found in the literature, their main features, their advantages and limitations, and comparisons are illustrated in their survey. On the other hand, the survey does not discuss tools and datasets. Abdullah et al. (2021) presented different multimodal emotion recognition methods that combine modalities, such as images and text, and facial expression combined with body's physiological responses. The findings in this survey indicated that combining multiple modalities in emotion recognition improves the recognition accuracy of emotions. Baltrušaitis et al. (2018) provided a five-tiered taxonomy based on the challenges that are encountered in every multimodal emotion recognition method. These five challenges include: fusion, alignment, representation, co-learning, and translation. Their paper identified co-learning as an area to be explored. However, this overview report mostly concentrated on multimodal research from the last ten years.

Lim et al. (2020) mainly focused on eye tracking methods and classifiers used for this data. This study underlines the necessity for combining physiological signals with eye-tracking modality. However, the results of this survey were not properly discussed or compared between inter-subject and intra-subject classification performance rates. Malla et al. (2020) specified a taxonomy for speech emotion recognition methods. The three main components of their work: speech data, feature extraction, and DFC taxonomy (classification), were classified using a CNN-based speech emotion recognition. This increases the likelihood of correctly identifying the emotion in the speech. The survey is exhaustive in its domain but is also limited as it only focuses on speech emotion resulting in a very low number of modalities.

The main shortcoming in existing surveys found through literature analysis is that neither of the studies mentioned above highlighted the value of emotion acquisition tools, data fusion techniques, or how to assess fine-grained emotions. The goal of our research is to undertake a thorough, systematic evaluation of the literature on the various emotion acquisition tools, multimodal emotion recognition classification models, in-depth analysis of fine-grained emotions, datasets, and fusion methodologies. Finally, we provide the results of the previous studies and point out any remaining research gaps as well as potential areas for further study. Table 1 presents a summary of all survey papers that we reviewed during our research.

Table 1
Comparison of existing surveys on emotion recognition.

Authors	Taxonomy followed	Method
Alswaidan & Menai (2020)	Text emotion recognition methods	Rule-based, classical learning, deep learning, and hybrid.
Abdullah et al. (2021)	Multimodal emotion recognition methods	Neural network architecture and deep learning technique.
Wu et al. (2014)	audiovisual emotion recognition techniques	Deep learning, datasets (GEMEP,RML and VAM) are explained.
Baltrušaitis et al. (2018)	Technical challenges faced by multimodal researchers	Different co-learning methods.
Jam et al. (2021)	Social signals and emotional expressions in real-world human-robot interactions.	Deep learning methods.
Lim et al. (2020)	Taxonomy involves Eye tracking methods	machine learning algorithms.
Malla et al. (2020)	Speech emotion recognition methods	MFCC, STFT, ECC and Classification is done by combining CNN and LSTM.
Kołakowska et al. (2014)	sensory data detection available on modern smartphones.	machine learning approaches implemented to recognize emotional states.
Imani & Montazer (2019)	Various emotions recognition methods have been represented for online learning platforms.	Classification based on feature and machine or deep learning methods.
Sharma & Dhall (2021)	emotion recognition on visual, speech, text, EEG.	Deep learning classifiers and feature level and decision level fusion.

4. Review method

This section summarizes academic work that has been done in the field of emotion recognition by outlining current research areas and providing a systematic review of the most recent findings.

4.1. Search strategy

Based on the requirements of this Systematic Literature Review (SLR), the following search terms were used: emotion recognition, multimodal emotion recognition, machine learning, deep learning. We started with an automated search for conference and journal papers in Google Scholar, followed by a manual search in digital databases such as the ACM, the IEEE, Elsevier and Springer. The following search queries were used:

- "Emotion Recognition".
- "Multimodal Emotion Recognition".
- "Emotion Recognition" AND ("Deep Learning" OR "Machine Learning").
- "Multimodal Emotion Recognition" AND ("Deep Learning" OR "Machine Learning").

4.2. Inclusion and exclusion criteria

In order to determine which of the Systematic Literature Review's (SLR) primary research works are relevant and related to this survey, a number of inclusion and exclusion criteria were used. We obtained data from the scientific texts written in English and published in digital sources such as the IEEE, the ACM, Sciencedirect and Springer between 2013 and 2021. Documents were included or excluded based on the following criteria:

Inclusion criteria:

1. The entire document is available for download.
2. The text of the document is in English.
3. The document was published between 2013 and 2021.
4. The document is relevant to the research questions.
5. The document mainly covers emotion recognition methods, mainly emphasis on multimodal emotion detection using deep learning.
6. The document was made available in high-resolution digital databases.

Exclusion criteria:

1 The complete text of the document is not available.
2 The document is not written in English.
3 The document does not fall inside the specified data range.

4.3. Study selection

By applying the above search strategy, inclusion and exclusion criteria, approximately 14,000 research papers were deemed relevant to this study and were given a priority focus. We drew a total of 100 publications and conference articles after reviewing them against inclusion and exclusion criteria. The number of studies on multimodal emotion recognition from 2013 to 2021 is depicted in Fig. 2. Since 2013, security research for the multimodal emotion recognition has been gradually expanding, with a sharp uptick in work in 2020. The main cause of this growth is the Covid-19′s appearance, which revealed the true significance of the mask and gave rise to fresh methods of emotion identification. The surge in activity in this area persisted in 2021, showing that the research community still places a high value on multimodal emotion recognition.

4.4. Research questions

This study looked into the following research questions (RQs):

RQ1: How different emotions can be analyzed and categorized?
RQ2: What are the different tools and technologies that can be used to acquire data for emotion recognition? RQ3: What are the different datasets available for emotion recognition and what are their main features?
RQ4: What are the different emotion classification methods available for obtaining a better recognition accuracy? RQ5: What are the data fusion techniques that guarantee a better emotion recognition accuracy?
RQ6: What are the challenges and future research areas that can advance the use of emotion recognition techniques?

Section 5 addresses our first research question (RQ1) and identifies fine-grained details of emotion and its various forms. Section 6 addresses the second research question (RQ2) and provides a comprehensive analysis and details of different tools and technologies that are employed for data acquisition to facilitate emotion recognition. Different datasets for emotion recognition are also presented and analyzed in Section 6 to address the third research question (RQ3). Section 7 outlines our fourth research question (RQ4) and different emotion classifiers are demonstrated. Section 8 lay out our fifth research question (RQ5) in which different data fusion approaches utilized for emotion recognition are explained. Section 9 addresses challenges and future scope in the emotion recognition to answer the final research question (RQ6).

5. Emotions

This section explains our first research question (RQ1) and outlines the characteristics of emotion and its numerous aspects.

Emotion is a conscious, subjective feeling that humans have when confronted with particular stimuli, and it is a fundamental part of natural social interaction. An individual is bound to experience various types and levels of emotions throughout the day. Emotions can also be described as a person's response to physical and psychological changes in their environment. These feelings can include everything from joy to rage to grief to exhilaration, and so on. The emotions of a person are complex to the point where diverse variables determine different emotions for each individual (Shirke et al., 2020). It is critical to be able to recognize people's emotional states automatically with the advancement of multimedia and human-computer interaction technologies (Lu et al., 2015). When playing a computer game, for example, the game's complexity can be changed by recognizing the player's emotional state. When a player is in a bad mood, the system will show him an easy and enjoyable game to help him relax and pass the time. When players are in a good mood, the game difficulty increases, making it more difficult for them to play. Music and video suggestions can also be used to alter emotions. When watching films or listening to music, positive multimedia content may help to decrease the impact of negative emotions. Different emotions are usually linked in some way; for instance, sorrowful emotions frequently involve a small bit of rage. As a result, it's worth looking at how to build a more effective emotion recognition model by merging emotional correlations (Zhang, 2020).

Emotions can be classified as either neutral or non-neutral. Neutral emotions are generally inexpressive feelings to any situations. Positive and negative emotions are two types of non-neutral emotions. Negative emotions have been linked to negative traits, e.g., anxiety, failure, and despair. Experiencing these bad emotions on a frequent basis can only be harmful to one's health. It was also discovered that it has a direct link to a person's attention span. These unpleasant emotions may even cause an individual's cardiovascular system to malfunction. Positive feelings, on the other hand, are perceived in an entirely different manner. Happiness and joy are two examples of good emotions that are considered to be representation of an optimal well-being (Shirke et al., 2020). The above discussed taxonomy of emotion is illustrated in Fig. 3.

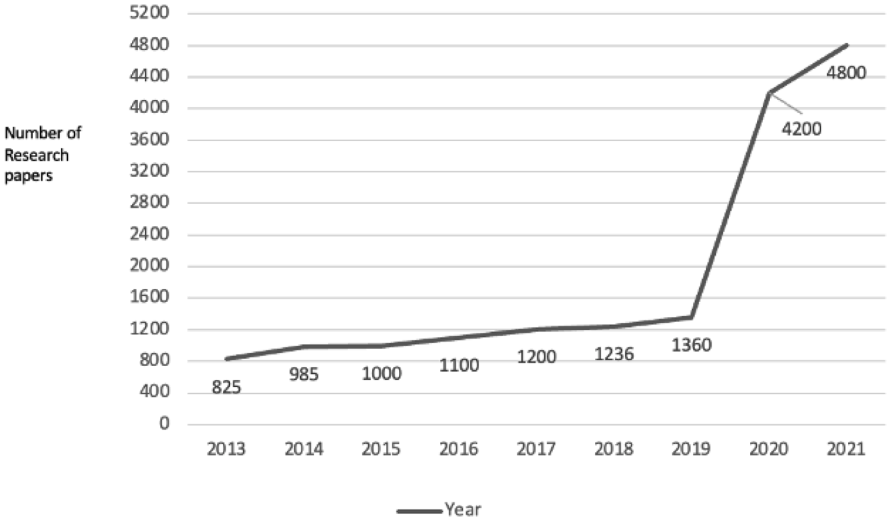

Fig. 2. Number of the multimodal emotion recognition related research papers published from 2013 to 2021.

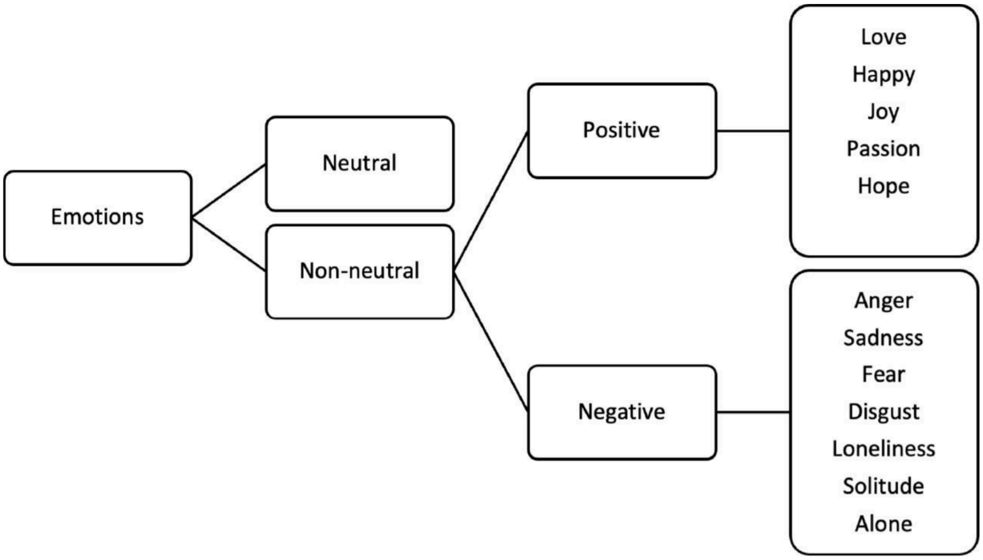

Fig. 3. Taxonomy of different emotions for multimodal emotion recognition.

Shirke et al. (2020) clearly stated the correlation between positive and negative emotions. Using Brain IoT, a multimodal emotion recognition method is developed that detects negative emotions and plays a film to generate positive emotions. The method is not accurate compared to other multimodal emotion recognition approaches. Hipson et al. (2021) clearly distinguished between solitude loneliness and being alone. The solitude is regarded as a positive emotion, compared to the negative emotion of loneliness, whereas being alone is considered a neutral state. Analysis of emotions in tweets is done using natural language processing. The findings only apply to the sentiment of words that appear in the same sentence as isolation phrases, not to the overall mood of a tweet (Hipson et al., 2021). In Subasi et al. (2021), the authors categorized the emotion into positive, negative and neutral using a neuroscan system. Signal processing hiccups can be easily addressed with the proposed method. This strategy, on the other hand, may be evaluated on larger and more varied datasets. Bosch et al. (2021) focused on frustration emotion, which is a form of a negative emotion. When comparing multilevel models, a system that includes individual expression estimates expression frequency far better than a model that only includes the expression. Suja & Tripathi (2016) used facial images to analyze emotions including anger, contempt, happiness, surprise, and neutral. This approach is used to recognize emotions in real time from a facial image. Their method has a lower accuracy compared to similar approaches (Bosch et al., 2021). In Lu et al. (2015), the authors distinguished positive, negative and neutral emotions from EEG and eye gaze. EEG and eye movements can be used in conjunction to boost performance in distinguishing different emotional states. The difficulty lies in obtaining the "ground truth" of an emotion due to the fuzzy boundaries of an emotion. Table 2 illustrates different fine-grained emotions recognized by different emotion recognition methods.

Recently, researchers have taken a purposeful step away from the six basic emotions to record a wider range of emotional experiences. Marsella et al. (2010) proposed computational emotions that show emotional behavior to be high dimensional, involving up to twenty five different types of emotion. The depiction of emotion in experience, expression, and neural processing is driven by specific categories of emotion, more so than by valence and arousal. The majority of emotional responses are discovered to be consistently blended rather than distinct. Computational models of emotion are software applications that use computational principles to analyze emotional stimuli, elicit emotional responses, and generate emotional actions in an effort to explain the phenomenon of emotions. Psychological theories of emotion are frequently used as the foundation for computational models of emotion (Marsella & Gratch, 2014). The idea at its most basic level seeks to identify what constitutes an emotion, such as what makes up an emotional state such as anger. The study of computational models of emotion is becoming a key part of efforts to better understand human behavior and enable human-machine interaction. Modern study on

Table 2
Fine grained emotion analysis.

Authors	Modalities	Dataset	emotions	tools
Lu et al. (2015)	Eye movements and EEG	Own dataset(twenty subjects)	Positive, Negative and neutral	SMI ETG eye tracking glasses, EEG signals collected by ESI NeuroScan System and a 62-channel electrode cap.
Suja & Tripathi (2016) et	Facial	CMU MultiPIE (337 subjects)	Negative (anger, disgust), Positive (hap- piness, surprise) and neutral	Raspberry Pi.
Bosch et al. (Bosch et al. (2021))	Facial and bodily expressions	Own dataset (70 subjects)	Frustration (Negative)	Camera.
B.Shirke et al. (Shirke et al., 2020)	EEG, heart rate	ThingSpeak Database	Fear, sadness, and anger (negative), happiness (positive).	EEG Headset, HRV, GSR sensors
Hipson et al. (Hipson et al., 2021)	Text	SOLO [19,277,359 Tweets])	Neutral(alone, Positive(solitude) and negative (lonely)	NRC Emotion Lexicon.
A.Subasi et al. (Subasi et al., 2021)	EEG	SEED (15 subjects)	Positive, Negative and Neutral	ESI NeuroScan System with active AgCl electrode.

human emotion has thoroughly proven the significant impact emotion plays in human behavior, which has occasionally been overlooked in work in cognitive science and AI. As a result, scientists are already using computer models of emotion as tools for studying human emotion and using it in practical applications (Osuna et al., 2020). Virtual humans, autonomous embodied figures capable of face-to-face interaction with real people through verbal and nonverbal behavior, are designed using computational models of emotion.

6. Data acquisition, and datasets for emotion recognition

The second research question (RQ2) is addressed in this section, which also gives a thorough analysis and information on the various tools and technologies used to collect data in order to assist emotion recognition. This also section addresses the third research question (RQ3) by presenting and analyzing several datasets for emotion recognition.

6.1. Tools and technologies for data acquisition

Emotion recognition systems are more economically viable for usage in special purposes, such as intelligent household robots to enable realistic and friendly human contact or employing emotion recognition for visual-audio surveillance. This is achievable because low-cost equipment can easily capture data required for emotion recognition (Xie & Guan, 2013). We devised a taxonomy of emotion recognition data acquisition tools and technologies based on two categories of classification (see Fig. 4). First classification is associated with sensors since they play a key role as emotion acquisition tools. Sensors are categorized into self-generating and modulating sensors based on how sensors handle physiological signals to recognize the emotions. Self-generating sensors capture physiological signals in which bodily processes are triggered automatically in emotional situations. They provide output voltages or currents that are proportional to the quantity being measured and their output bandwidth equals that of the quantity being measured. For example, EEG sensors are self-generating sensors that do not need additional trigger for emotion recognition. Modulating sensors require an external source for their functioning to handle emotions. GSR is an automatically controlled modulating sensor that measures skin conductance. However, it is controlled by sympathetic activity, which is responsible for a wide range of human behavior as well as mental and affective states. Commonly used self-generating, modulating sensors and VR tools are presented in Fig. 5.

The tools employed for the acquisition of physiological signals provide the second main classification category for our taxonomy. Smart wearable tools are attached to body directly to detect emotions quickly and accurately. For example, electrode caps with wet or dry electrodes are wearable to capture EEG signals for emotion detection. Mounted tools are attached to body with the aid of an external component. Gjoreski et al. (2021) developed a multi-sensor mask that can measure the user's facial physiological reactions, facial muscle signals, and motions. Head-mounted displays (HMDs) for virtual reality are becoming increasingly popular, as they allow completely immersive systems that insulate the user from external world stimuli (Marín-Morales et al., 2018). External tools are devices that are not linked to the human body and have been used to induce emotion using non-immersive stimuli. Camera, CCTV, virtual cave comes under this category. Different types of sensors used for emotion recognition is presented in Table 3.

6.1.1. . Physiological signal processing sensors

Self-generating sensors capture physiological signals in which bodily processes are triggered automatically in emotional situations. They provide output voltages or currents that are proportional to the quantity being measured and their output bandwidth equals that of the quantity being measured. The most widely used self-generating sensors are: electroencephalogram (EEG), electromyography (EMG) and electrodermal activity (EDA). EEG sensors monitor the brain impulses non-invasively, and the signals are used to pick the best combination of features for emotion recognition (Majid Mehmood et al., 2017; Quiroz et al., 2018). Nakisa et al. (2018) classified an emotional response to positive and negative states using EEG signal characteristics. Majid Mehmood et al. (2017) used 14-channel EEG machine to record EEG signals while the subjects watched images with elicitation stimuli, such as sad, happy, and scared. The EEG headcap used to record EEG data is presented in Fig. 5. The main advantages of EEG (Tang et al., 2017) are (1) Facial expressions and words can be used to hide emotions and EEG is independent of both stimuli. (2) Physiological signals contain reliable information as compared to facial expressions and speeches. (3) As wearable computing techniques advance, it becomes even more easier to capture users' physiological inputs and thus EEG signals are particularly

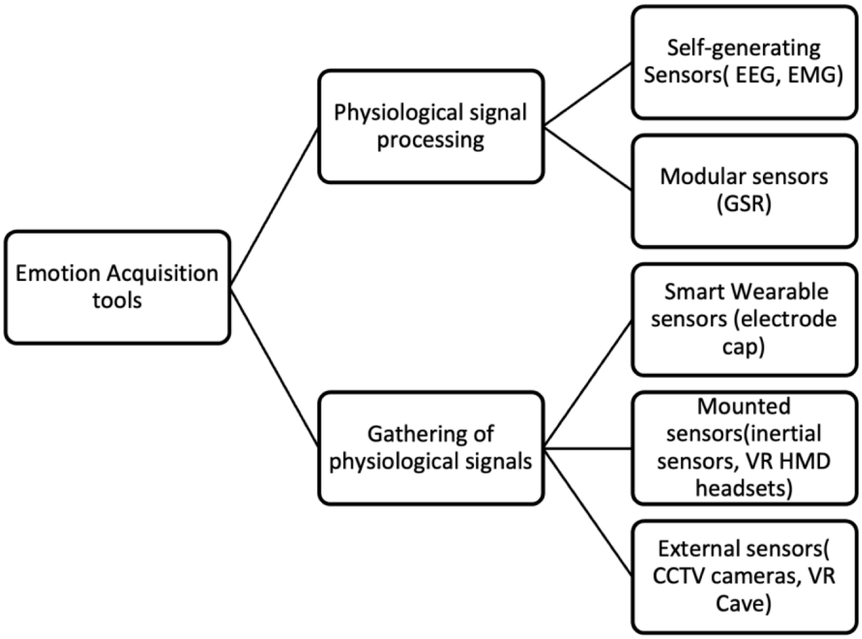
Emotion Acquisition tools
Physiological signal processing
Gathering of physiological signals
Self-generating Sensors(EEG, EMG)
Modular sensors (GSR)
Smart Wearable sensors (electrode cap)
Mounted sensors(inertial sensors, VR HMD headsets)
External sensors(CCTV cameras, VR Cave)

Fig. 4. Taxonomy of emotion acquisition tools (Techxplore, 2021).

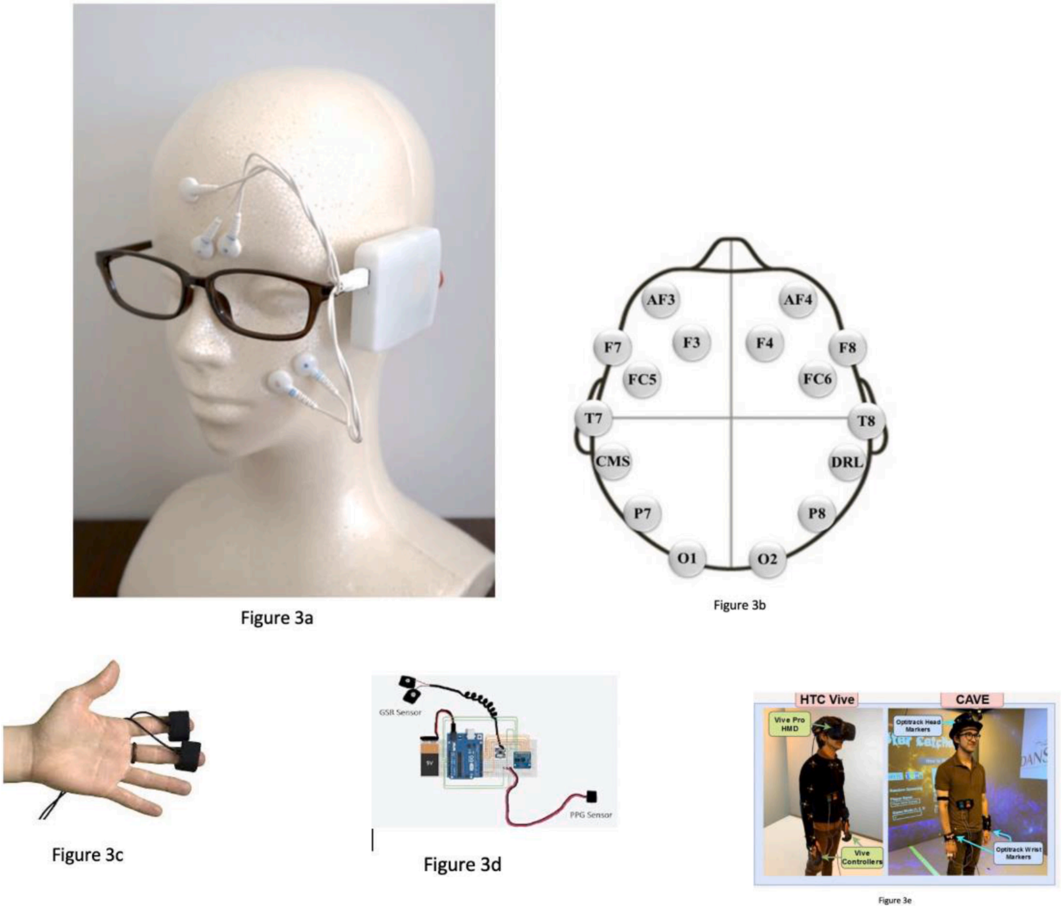

Fig. 5. a. Wearable device to record Electromyography (EMG) data (Sato et al., 2021) b. EEG headcap with 10/20 electrode placement (Majid Mehmood et al., 2017), c. EDA sensor (Ragot et al., 2017), d. GSR and PPG sensors (Competition, 2016), and e. Virtual HMD headset and virtual Cave environment (Elor et al., 2020).

useful. (4)EEG signals can be employed in multimodal retrieval to reflect a person's feelings.

Electromyography (EMG) is a technique for measuring muscle activity that has been linked to different forms of emotion. M.M. Hassan et al. (Hassan et al., 2019) used Zygomaticus Electromyography (zEMG), which encompasses the prevailing aspects of induced emotion, used to measure the facial muscle contraction through its electrical activity.

Cognition and emotion in humans are psychological processes that can be measured through the electrical conductivity of the facial skin, which is typically measured in palmar sites. Electrodermal activity (EDA) comprises of tonic and phasic components. Skin Conductance level (SCL) is a tonic component of EDA (Shukla et al., 2019).

In contrast to the self-generating sensors, modulating sensors, require an external source. Galvanic Skin Response (GSR) sensor is an unconsciously controlled modulating sensor that measures skin conductance. It is regulated independently by sympathetic activity, which affects aspects of human behavior along with mental and emotional states. GSR is a highly sensitive indicator of emotional arousal. It's usually taken from the fingers or palms. It regulates the amount of perspiration produced by skin pores. GSR is not under the conscious control of humans. Human emotion is indirectly regulated by sweating through the sympathetic nervous system (Competition, 2016). Another modulating sensor, Photoplethysmogram (PPG) is a low-cost, non-invasive optical technology that uses infrared light to assess epidermal blood flow. Heart rate, which is used in the emotion recognition process, is measured by PPG (Udovičić et al., 2017).

6.1.2. . Emotion acquisition tools to gather physiological signals

Smart Wearable tools: Emotions are detected swiftly and correctly using smart wearable tools that are physically attached to the body. Electrode caps with wet or dry electrodes, for example, can be worn to record EEG signals for emotion detection. Ragot et al. (2017) used Empatica wrist band to record EDA signals and cardiac activities. The authors pointed out that Wearable sensors for emotion identification were as efficient as laboratory sensors. The main drawback of this strategy is that it extracts a small set of features for emotion recognition. Hassan et al. (2019) used EDA, PPG and zEMG sensors for measuring physiological modalities. It aids in the extraction of actual feelings and the circumvention of the limits of emotion recognition algorithms based on functional neuroimaging. The suggested model had trouble distinguishing neutral feelings from other emotional states: Happy, sad, relaxed and disgust.

Quiroz et al. (2018) recorded physiological signals pressure and heart rate using smart wearable watch embedded with triaxial accelerometer and gyroscope. The use of physiological features in this method increased detection accuracy. The integrity of sensor data is one of the cited concerns. Yang et al. (2020) used FPGA controller as smart wearable tool to record physiological signals such as EEG, ECG and PPG. This tool can be used for real-time applications, but it is not suitable for complex neural networks.

Mounted emotion acquisition tools: Mounted tools are attached to body with the aid of an external component. Head-mounted displays (HMDs) are the most popular mounted devices used for emotion recognition. using these devices, a virtual environment is created that may be explored directly with our senses. Through virtual experiences, it generates a sense of being in the real world. As a result, virtual reality (VR) enables for the simulation and evaluation of spatial settings under controlled experimental conditions, providing for the time and cost-

Table 3
Different sensors used for emotion recognition.

Authors	Method	Tools	Modalities	Datasets
V.J Aiswaryadevi et al. (Aiswaryadevi et al., 2021)	CNN,RNN and autoencoder	Smart IoT sensors (External)	Text, image, and acoustic data	Own dataset.
M. Ragot et al. (Ragot et al., 2017)	SVM	the Biopac BioNomad ix MP150 and Empatica E4 Wristband (Smart Wearable).	EDA and cardiac activities	own dataset.
M Hassan et al. (Hassan et al., 2019)	DBN,FGSVM	EDA, PPG and zEMG sensor (Smart Wearable)	Physiological signals from EDA, PPG and zEMG	DEAP.
B. Nakisa et al. (Nakisa et al., 2018)	Ant Colony Optimization and EC algorithms	Mobile EEG sensors (Mounted)	EEG	DEAP and MAHNOB
J. Quiroz et al. (Quiroz et al., 2018)	RF	smart watch included a triaxial accelerometer and a triaxial gyroscope (Smart wearable)	accelerometer, heart rate, and gyroscope features.	own dataset.
Matsuda et al. (Matsuda et al., 2018)	min,max	android smartphone, mobile eye tracking headset, senstick (mounted)	audiovisua data and behavioral cues (eye and head/ body movements)	RECOLA, SEMAINE
C. Yang et al. (Yang et al., 2020)	CNN	Spartan6 FPGA controller and a Blue- tooth slave module. (Smart Wearable)	EEG, ECG, and PPG.	own dataset

efficient exclusion and manipulation of factors that would be unachievable in actual space. Depending on the visualization and interaction devices utilized, Vergara et al. (2019) divided VR applications into two major classes.

- Non-immersive (the world's most well-known window), in which the user views the environment through a computer's flat screen acting as a "window".
- Immersive, in which the user views the environment two small screens in front of their eyes, that is the user is immersed in the virtual environment by wearing glasses.
- According to the virtual world's visualization technology, VR immersive applications are separated into two classes (Marín-Morales et al., 2018).
- The head-mounted display consists of a display device that is positioned in front of both eyes, or a dedicated display device positioned in front of each eye separately.
- The virtual CAVE (cave automated virtual environment), in which stereoscopic projectors project the virtual world onto a room's walls, ceiling, and floor. In this instance, the user will need to wear passive stereo glasses to see the virtual world in 3D.

There are a variety of VR systems on the market presently, with a wide range of budgets and levels of customer immersion. Thus, current commercial devices are divided into three categories: lower immersion and lower cost, e.g.,

Google Cardboard and Samsung GearVR; average but acceptable immersion and average cost, e.g., Oculus Quest and HTC Vive Focus; and finally, a high-level immersion and a higher cost, e.g., Valve Index, Oculus Rift, and HTC Vive. A number of new VR platforms, e.g., Faceteq, allow to capture both facial expressions and biological responses while using an HMD (Mavridou et al., 2017). More recently, HP Reverb G2 provides a complete hardware and software VR solution that combines an immersive VR experience with eye tracking, pupillometry, heart rate detection, and a dedicated face camera to record facial expressions. Different VR tools used for emotion recognition are illustrated in Table 4.

Georgescu & Ionescu (2019) used VR headset to collect emotion features from facial expressions and the method utilized is less expensive and may be used with any VR headset. The VR headset enveloped the top portion of the face without considering realistic occlusion into consideration. Hence, extraction of all facial features is not possible (Georgescu & Ionescu, 2019). Hickson et al. (2019) acquired emotion features using VR headset with internal commercial HMD sensors. Without the need of any permanent external camera, facial expressions can be extracted from images captured by a tracking camera within a VR headset. The main limitation is that it is non-trivial to use own dataset or non-occluded data for the emotion recognition (Hickson et al., 2019). Houshmand & Khan (2020) used a VR headset with an internal camera to extract features from facial expression. The transfer learning method used for the feature fusion gives a special consideration to facial occlusions arising from VR headsets. However, an external camera was also needed to increase the detection accuracy. Hinkle et al. (2019) used wearable sensors along with a VR headset to gather physiological data from subjects. The information gathered is utilized to identify the respondents' emotional responses to stimuli. The shortcoming of this method is the usage of cumbersome equipment sensors with cables that are difficult to handle (Hinkle et al., 2019).

Nakisa et al. (2018) used mounted mobile EEG sensors to record EEG signals. The employed EC algorithms solved the high dimensionality

Table 4
Different VR tools used for emotion recognition.

Authors	Method	Tools	Modalities	Datasets
M.Georgescu et al. (Georgescu & Ionescu, 2019)	CNN	Virtual reality (VR) headset used an ex- ternal camera	Facial expression	FER+ and AffectNet
Hickson et al. (Hickson et al., 2019)	CNN	VR headsets with internal commercial HMD sensor	Facial expression	Own dataset
B.Houshmand et al. (Houshmand & Khan, 2020)	Transfer learning	Samsung Gear VR headset with internal camera	facial expression	FER+, RAFDB, AffectNet.
L. Hinkle et al. (Hinkle et al., 2019)	SVM, KNN	Wearable sensors, e.g., respiration straps, finger electrodes along with Oculus DK2 VR headset	Facial, voice, temperature, heart beat.	Own dataset
G.Lorenzo et al. (Lorenzo et al., 2016)	SVM	Cave Automated Virtual Environment (CAVE),head-mounted display (HMD)	facial expression	Own dataset.
J. Marin et al. (Marín-Morales et al., 2018)	SVM	Cave Automated Virtual Environment (CAVE)	EEG, ECG	own dataset.

problem of EEG features. However, a possibility of premature convergence problem in EC algorithms is always present (Nakisa et al., 2018). Matsuda et al. (2018) used a mounted Senstick to record audiovisual and behavior cues. The real-world experiments are conducted using subjects as tourists to generate own dataset, so all the issues were addressed. Optimization techniques are not used to reduce external interference. Hence, extraction of features will be less accurate compared to those from standard datasets (Matsuda et al., 2018).

External emotion acquisition tools: External tools are devices that are not linked to the human body and have been used to induce emotion using non-immersive stimuli. This category includes cameras, CCTV, and virtual caves. Stereoscopic projectors project the virtual world onto a room's walls, ceiling, and floor in the virtual CAVE (cave automated virtual environment). To see the virtual environment in 3D, the user will need to wear passive stereo glasses (Marín-Morales et al., 2018). Virtual CAVE is an innovative tool that can help stimulate the development of unique immersive affective elicitation. However, their accuracy is less compared to other emotion elicitation methods. Lorenzo et al. (2016) utilized both Virtual HMD and CAVE technology for feature extraction in emotion recognition. The work showed that an immersive virtual reality system is a better option to create social situations. In addition, a Virtual Cave is an expensive solution compared to an HMD (Lorenzo et al., 2016). Different Virtual cave tools used for emotion recognition are mentioned in Table 4.

Built in cameras, eye trackers and microphones are externals tools used for emotion recognition in applications such as online learning and traffic monitoring systems. They are non-intrusive devices but can annoy pupils by distracting them and making them feel uneasy about their behaviors being recorded (Petrovica et al., 2017). Aiswaryadevi et al. (2021) used smart IoT sensors as external emotion acquisition tools to obtain acoustic data. The usage of IoT sensors have revolutionized numerous emotion recognition applications in terms of speed, resilience, and pervasiveness. As a result, the volume of data that has to be processed has grown exponentially. Although Big Data processing can help with data refinement, any emotion identification system's accuracy will always be a challenge (Shamim Hossain & Muhammad, 2019).

6.2. Datasets for multimodal emotion recognition

Large data sets are required for researchers to delve deeper into human emotional experience and expression, as well as to develop benchmark methodologies for automatic emotion recognition. A number of datasets for emotion recognition are acquired with one or more modalities, which are available publicly. In this section the nine commonly used datasets for emotion recognition are presented.

emoFBVPDatabase: Face expressions, body expressions, vocal expressions, and physical signals were all used by emoFBVPD to record actors' reactions to affective emotion labels. In addition to the multimodal recordings, this provides data on facial feature tracking and skeletal tracking. After each excerpt of portraying emotions, the performers fill out an assessment form to rate the recordings of all the data. The recordings in this database are synced to allow researchers to examine multiple emotional responses at once using all of the available modalities. emoFBVP is the first emotion dataset to incorporate recordings of several modalities with varying intensities of emotional displays. This collection includes 1380 audio samples, face and body video sequences, and physiological data pertaining to various emotions (Ranganathan et al., 2016).

DEAP: DEAP (Database for Emotion Analysis using Physiological Signals) is a multimodal dataset for analyzing human affective states that is freely available to the public. While watching 40 one-minute-long samples of music videos, the electroencephalogram (EEG) and peripheral physiological data of 32 subjects were monitored. Every film was scored on dominance, valence, arousal, like/dislike, and familiarity by the participants. Frontal face footage was captured for 22 of the 32 individuals (Koelstra et al., 2011).

SEED: The SEED (SJTU Emotion EEG) dataset comprises subjects' EEG signals recorded when their emotion were elicited by showing film clips. The film snippets have been carefully chosen to elicit a range of emotions, including negative, happy, and neutral. From the pool of materials, fifteen Chinese film clips with emotion states were chosen as stimuli for the studies. Film clips are selected on basis of (a) the duration of the experiment should not be excessively long, as this may produce fatigue in the volunteers; (b) the videos should be simple to understand without explanation; and (c) the movies should evoke only one desirable target emotion. Participants were asked to fill out a form immediately after watching each video clip to reflect their emotional experiences (Zheng & Lu, 2015).

FER2013: FER2013 was constructed by searching for photographs of faces that matched a collection of 184 emotion-related keywords such as "blissful," "enraged," and so on, using the Google image search API. Nearly 600 strings were created by combining these keywords with phrases related to ethnicity, gender, and age to create facial picture search queries. For each query, the first 1000 photos retrieved were saved for the next round of processing. The dataset contains 35,887 facial photos, the majority of which were taken in natural settings. 6077 images depict sadness, 4002 images depict surprise, 4953 photos depict anger, 8989 images depict happiness, 547 images depict contempt, 6198 images depict emotional neutrality, and 5121 images depict fear. Faces vary widely in age, position, and occlusion circumstances, making a difficult dataset. Furthermore, human recognition accuracy is estimated to be around 65 percent (Goodfellow et al., 2013).

K-EmoCon: K-EmoCon is a multimodal dataset that includes extensive descriptions of constant emotion through- out realistic talks, making it excellent for investigating emotions in social contexts. The dataset includes video recordings, EEG, and peripheral physiological signals, as well as multimodal measures acquired with off-the-shelf tools during 16 sessions of roughly 10-minute-long period debates on a social subject. It varies from previous datasets in that it includes emotion evaluations from all different approaches: self, argument partner, and observers from the outside. Raters marked emotional displays every 5 s while watching the discussion footage in respect of arousal- valence and 18 other group emotions. K-EmoCon dataset enables for multiperspective emotional evaluation during interpersonal relationships (Park et al., 2020).

PMEmo: PMEmo is a research tool for public music retrieval and emotion recognition that operates across several modalities. PMEmo contains song-level elements, manually picked chorus fragments, and their shared MER attributes. PMEmo contains a large amount of music with contemporaneous physiological signals, standard valence-arousal annotations, as well as the components of annotators recruited. 794 music clips were annotated by 457 individuals in this dataset (Zhang et al., 2018).

MAHNOB-HCI: MAHNOB-HCI is a multimodal database that collects responses to emotional states in order to identify emotions and investigate implicit labeling. In a multimodal setup, eye gaze data, voice signals, face videos, and physiological signals were all synced. There were 27 male and female participants from varied cultural backgrounds involved in two experiments. They viewed 20 emotional movies in the first test and self-reported their feelings using emotional terms such valence, arousal, predictability, and dominance. In the second trial, brief videos and photographs were first shown without tags, then with correct or incorrect tags. The participants were asked to rate how much they agreed or disagreed with the tags that were displayed (Soleymani et al., 2011).

VREED: VREED (VR Eyes: Emotions dataset) is a multimodal affective dataset in which emotions were elicited by immersive 360° Video-Based Virtual Environments (360-VEs) given through a Virtual Reality headset. ECG data, eye tracking data, self-reported surveys, and GSR data are all part of the VREED. The data was acquired from 34 healthy subjects for a total of 408 trials. Focus groups and a pilot trial were used to choose the 360-VEs (with 12 additional volunteers). The dataset is documented and includes the instructions protocol and questionnaire samples. A preliminary machine learning analysis was performed, demonstrating state-of-the-art performance employing non- immersive modalities as documented in affective computing literature. VREED is one of the first multimodal VR datasets to use behavioral and physiological cues to recognize emotions (Tabbaa et al., 2021).

MEmoR: MEmoR is a unique dataset for video-based multimodal emotion reasoning. MEmoR is a technology that allows both speakers and non-speakers to receive fine-grained emotion annotations. MEmoR is based on 5502 video clips and 8536 data samples from the popular television show The Big Bang Theory. MEmoR is annotated at the individual level, including 14 emotions from Plutchik's wheel (Cimtay et al., 2020). In these movies, non-speakers are always empty of audio-textual signals and are frequently visually indistinguishable, whereas speakers may be missing in visual detail. MEmoR is a more realistic yet challenging testbed. The suggested MEmoR dataset incorporates both short-term settings and external data sources to suit various reasoning styles (Shen et al., 2020).

The mostly used datasets for emotion recognition are illustrated in Table 5. When a subject is asked to perform a specific feeling, this is known as posed emotion. The instinctive reaction of subject viewing the elicitation is called spontaneous emotion.

7. Classifiers for emotion recognition

This section outlines our fourth research question (RQ4) in which different learning algorithms (ML/DL) for multimodal emotion classification are demonstrated.

Pattern recognition methods or a classifier classifies the features applied on different set of emotional inputs. Classifier of emotions can be divided into two categories: basic machine learning and deep learning. The taxonomy of emotion recognition classifiers is presented in Fig. 6 which represents three domains in emotion recognition- classifier types, learning algorithms used and its applications.

7.1. Machine learning classifiers

Emotion recognition activities have long been aided by computational intelligence, particularly machine learning. Before being used to forecast emotions, a machine learning classifier is constructed on a training dataset. Particularly, supervised machine learning tools are commonly used to recognize emotions (Torres et al., 2018). To deal with today's increasingly demanding emotion recognition applications, more complex machine learning models have been investigated by researchers. Moreover, various machine learning tools are combined to handle multimodal emotion recognition incorporating many modalities such as image, text, and physiological signals. When dealing with a large number of diverse emotional characteristics, machine learning techniques should increase their performance. Most of these experiments, on the other hand, use standard machine learning approaches with a minimal number of samples, resulting in methods that either don't generalize well or don't have enough data for recognize emotions in realistic situations. Different machine learning classifiers used for emotion recognition is summarized in Table 6.

7.2. Deep learning classifiers

The main shortcoming of traditional feature engineering and machine learning techniques are the extraction of nonlinear patterns from multivariate time series information (Fabricio & Liang, 2013). Furthermore, picking the most important characteristics from a huge set of data is crucial and will necessitate dimensionality reduction strategies. Machine learning models perform with high accuracy when fed with carefully selected features that are extracted using statistical valuables, such as variance, entropy, and mean (Aiswaryadevi et al., 2021). As features' dimensionality increases, the computing complexity of feature extraction, for instance, may rise significantly. Generally, finding the optimum set of features is challenging. To address the difficulty of extracting usable time series data features, several academics have concentrated on deep learning techniques. The time-consuming process of generating handcrafted features for machine learning models is taken care of by the deep learning (DL).

Deep learning technologies such as CNN, RNN, and autoencoder have resulted in a remarkable improvement in almost all of the fields of computing, especially, computer vision, machine translation, audio recognition etc. Because of its capacity to provide high-level data abstraction, deep learning models have recently been used to build reconfigurable structures for emotion recognition. Deep neural networks are employed to collect prominent variables from high-level data representation (Liu et al., 2021).

The capacity to work effectively with raw data and to extract effective features, makes DL approaches desirable. Standard statistical samples are presented to the network, and each nonlinear transformation results in the construction a hierarchical data representation structure. In this structure, each hidden layer is an abstract feature representation of the features in the previous hidden layer. The comparison of machine learning and deep learning classifiers for emotion recognition is shown in Table 7. For emotion recognition, Table 8 presents the deep learning

Table 5
Different datasets for emotion recognition.

Dataset	Stimulus	Modalities	Annotation Method	Spontaneous/ posed	Purpose
emoFBVPD	1380 samples	Audio, video, physiological data	Per stimuli	Spontaneous	Intensities of emotions in multiple modalities recorded simultaneously
DEAP	40 music	Face videos,	Per stimuli	Spontaneous	Affective moods of humans are
SEED	videos	physiological signals	Per stimuli	Posed	studied.
	15 film clips	EEG signals			EEG standard dataset.
FER2013	35,887 facial	Images	Continuous	Posed	Search for images of faces that
K-	images	Face, gesture, speech, audio, physiological	Per stimuli	Spontaneous	match a set of 184 emotion. Emotions in the social context are
Emocon	16 sessions	signals			studied.
PMEmo	457 subjects	audio	Continuous	Posed	Music retrieval and emotion recognition research in multiple modali- ties.
MAHNOB-HCI	20 film excerpts	EEG, ECG, GSR, RESP and TEMP	Per stimuli	Posed	Affective stimuli for emotion recognition and implicit tagging research.
VREED	34 subjects	Physiological features:- ECG and GSR ,Behavioural features (eye gaze)	Per stimuli	Posed	Emotions were triggered using immersive 360° Video-Based Virtual Environments.
MEmoR	5502 video clips	video, audio, text	Per stimuli	Spontaneous	Tool that allows both speakers and non-speakers to have fine grained emotion annotations.

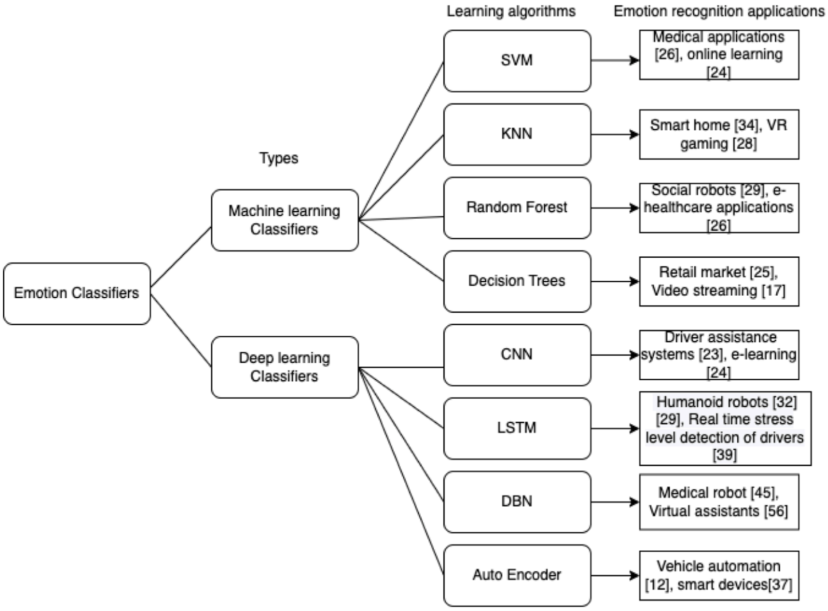
Learning algorithms
Emotion recognition applications
Types
Emotion Classifiers
Machine learning Classifiers
Deep learning Classifiers
SVM
KNN
Random Forest
Decision Trees
CNN
LSTM
DBN
Auto Encoder
Medical applications [26], online learning [24]
Smart home [34], VR gaming [28]
Social robots [29], e-healthcare applications [26]
Retail market [25], Video streaming [17]
Driver assistance systems [23], e-learning [24]
Humanoid robots [32] [29], Real time stress level detection of drivers [39]
Medical robot [45], Virtual assistants [56]
Vehicle automation [12], smart devices[37]

Fig. 6. Taxonomy of feature classifiers for emotion recognition.

Table 6
Machine learning classifiers.

Authors	Method	Tools	Modalities	Datasets	Recognition accuracy
T.D. Kusumaningrum et al. (Kusumaningrum et al., 2020)	Random Forest	Sensors	EMG and EOG signals	DEAP	62.58 percent.
C. Qing, et al. (Qing et al., 2019)	Decision Tree, KNN and Random Forest.	EEG head cap	EEG signals	DEAP, SEED	Accuracy of DEAP is 62.63percent and that of SEED is 74.85percent.
O.Bazgir et al. (Bazgir et al., 2018)	Support vector machine (SVM)	Mobile EEG sensors	EEG	DEAP	91.3percent.
J.Singh et al. (Singh & Gill, 2022)	SVM, KNN	Eye tracker, EEG sensors	eye features, EEG signals	SEED	79.63percent.
Q.Tomas et al. (JohnPaul Quilingking Tomas et al., 2020)	k-Nearest Neighbors, Naive Bayes, and Support Vector Machines.	Senstick mounted on ear	audio and lyric features	Own dataset	61 percent.
Y.dai et al. (Dai et al., 2015)	Reputation-driven Support Vector Machine (RSVM) classification algorithm	Wearable headset with dry electrode.	EEG, heart rate, ECG, blood pressure,	DEAP	Accuracy is 75.19percent and standard deviation 10.85percent.
Prasanth et al. (Prasanth et al., 2021)	MLPClassifier, which trains using Back propagation	Senstick	Speech: (MFCC, MEL and Chroma)	RAVDESS	85.12percent.

Table 7
Comparison of ML classifiers and DL classifiers for emotion recognition.

Machine Learning Classifiers	Deep learning Classifiers
Need extracted features for classification.	Deal with unprocessed data.
Data pre-processing and handcrafted feature construction are needed.	The process of extracted features and selection should be automated.
Ineffective in eliciting the complicated and nonlinear patterns in multivariate time se- ries datasets.	Deep learning extracts effective and consistent characteristics from time series data.
With growing feature dimensionality, the computing cost of feature selection may rise significantly.	The computational cost is less since deep learning can automate the feature selection process.
Frequently lack sufficient information for emotion detection in realistic scenarios.	Best for real time emotional recognition.

classifiers.

7.3. Machine learning vs deep learning classifiers

We compared the detection accuracy of six machine learning and deep learning emotion classifiers from various research papers which are using DEAP dataset for emotion recognition. The DEAP dataset was used for the comparison of detection accuracy since it was the standard dataset for multimodal emotion recognition in more than half of the research papers reviewed. By analyzing the detection accuracy of different classifiers as shown in Table 9, deep learning models give better detection accuracy than machine learning models. Among the machine learning models we selected for study, SVM is the best machine learning model which gives a higher detection accuracy and for deep learning, the best model is the combination of three deep learning algorithms: CNN, RNN, AE. Fig. 7 indicates the usage frequency of ML/DL models in the study. The most commonly utilized models were CNN (*n*

Table 8
Deep learning classifiers.

Authors	Method	Tools	Modalities	Datasets	Recognition accuracy
K.Yang et al. (Yang et al., 2021)	LSTM	sensors from a smartphone and EDA and thermopile sensor incorporated in a wristband	visual, acoustic, typing, and physiological modality	own dataset	89.2 percent.
Hassan M.M et al. (Hassan et al., 2019)	DBN and Fine Gaussian Support Vector Machine (FGSVM)	EDA, PPG and zEMG sensors signals	physiological data : EEG signals, ECG	DEAP	89.53 percent.
W. Liu et al. (Liu et al., 2016)	Bimodal Deep AutoEncoder (BDAE)	ECG sensors and eye tracker	EEG and eye features	SEED, DEAP	mean accuracies of 91.01 percent for SEED and 83.25 percent for DEAP.
H.Tang et al. (Tang et al., 2017)	Bimodal-LSTM model	EDA,PPG and zEMG sensors	EEG signals and peripheral physiological signals	DEAP, SEED	mean accuracy of 93.97 percent for SEED and 83.53 percent for DEAP.
J. Ma et al. (Ma et al., 2019)	plain deep, residual and MM- ResLSTM network	EEG sensors	EEG and peripheral phys- iological signals (PPS)	DEAP	Arousal and valence with an accuracy of 92.87percent and 92.30percent
Aiswary devi, V. J et al. (Aiswaryadevi et al., 2021)	a CNN, RNN and autoencoder	Smart IoT sensors	Text, image and acoustic data	DEAP	95 percent .

Table 9
Detection accuracy of different machine and deep learning classifiers using DEAP dataset.

Machine learning		Deep learning	
Method	Recognition accuracy	Method	Recognition accuracy
O. Bazgir et al. (Bazgir et al., 2018) **SVM**	91.3	W. Liu et al. (Liu et al., 2014) **BDAE**	83.25
C. Qing et al. (Qing et al., 2019) **KNN-Decision tree-RF**	62.63	Aiswaryadevi et al (Aiswaryadevi et al., 2021) **CNN-RNN-AE**	95
Y.Dai et al. (Dai et al., 2015) **RSVM**	75.19	Hassan M et al. (Hassan et al., 2019) **DBN**	89.53
R.W. Picard et al. (Picard et al., 2015) **KNN**	88.3	H.Tang et al. (Tang et al., 2017) **BLSTM**	83.53
P.Chen et al. (Chen & Zhang, 2017) **RF, NN**, EC	86.7	J.Maet et al. (Ma et al., 2019) **LSTM**	92.3
T.D. Kusumaningrum et al. (Kusumaningrum et al., 2020) **RF**	62.58	H.Zhang et al. (Zhang, 2020) **BDAE, LIBSVM**	85.71

= 21), SVM ($n = 18$), and autoencoder ($n = 16$).

8. Data Fusion methods

This section presents our fifth research question (RQ5), which explains several data fusion techniques used for emotion recognition.

Multimodal fusion is one of the most active areas of research due to its potential for a wide range of applications, including image segmentation, emotion recognition, video classification and event detection. To build effective emotion identification models, it's critical to combine different features with fusion technology. The two most important challenges in blending diverse modalities are (i) to determine at what abstraction level the modalities need to integrate and (ii) what are the procedures for integrating modalities (Yang et al., 2021). A unique way for using the complementarity of distinct types of features has been the mixing of features such as different physiological signals, EEG, and exterior behaviors, eye gaze (Qiu et al., 2018). There are three sorts of traditional fusion techniques based on their level of fusion: feature-level fusion, decision-level fusion, and hybrid multimodal fusion. Different modalities are assumed to be independent in decision level fusion. When the modalities have significant differences in temporal properties, feature level fusion implies precise time synchrony between them and does not generalize well. As a result, synchronization between various aspects becomes critical.

Other significant data fusion methods used for multimodal emotion recognition based on modality dimensions and optimization are deep learning-based fusion and attention-based fusion. With deep learning's rapid advancement, recent research methods are using deep learning models to aid multimodal fusion. Depending on the context, the attention model allows the network to emphasize features from specific temporal or spatial regions. Taxonomy of data fusion methods for multimodal emotion recognition is presented in Fig. 8.

8.1. *Feature level fusion*

Blending different modalities at the feature level, also called early fusion, is a common and simple strategy. Before being sent to the models as a whole, the extracted multimodal features are combined into a high-dimensional feature vector. Before fusion, most multimodal emotion identification techniques extract significant characteristics from unimodal data. Feature fusion can be done in either a unimodal or a cross-modal manner (Chaparro et al., 2018). Unimodal features are unable to convey all relevant data, however multimodal features can better capture the characteristics of emotions. Chaparro et al. (2018) reported a 97% accuracy rate from a feature level fusion of EEG and facial expression modalities using basic machine learning classifiers.

Human emotions are reflected in more ways than EEG and facial expression, although this method only used two physiological markers (Chaparro et al., 2018). With audio, text, and visual as trimodal inputs, Zhang et al. (2021) investigated unimodal and multimodal features. To develop multimodal integrated variables and maximize their correlations, Deep Canonical Correlation Analysis is used along with an encoder-decoder system. The main limitation is that users must set the fusion parameters based on their prior expertise.

There are two advantages to feature-level fusion: (1) It can take advantage of the relationship among several modalities immediately on, making task completion easier; and (2) the merged results contain additional data than a single modality, implying an increase in

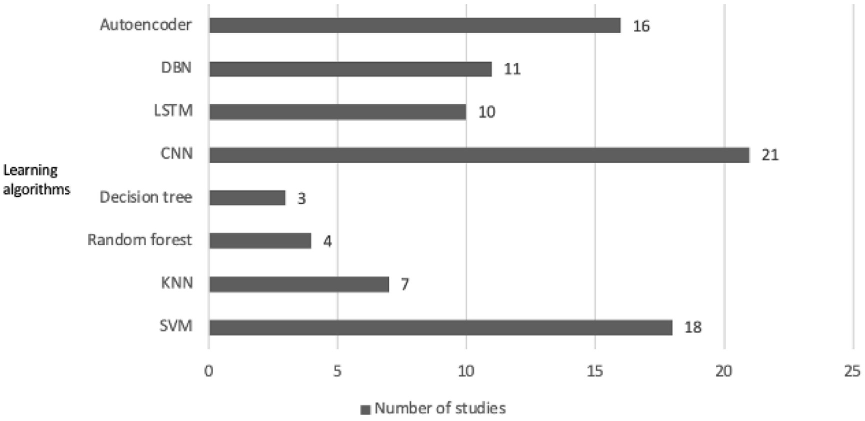

Fig. 7. Usage frequency of ML/DL models in multimodal emotion recognition.

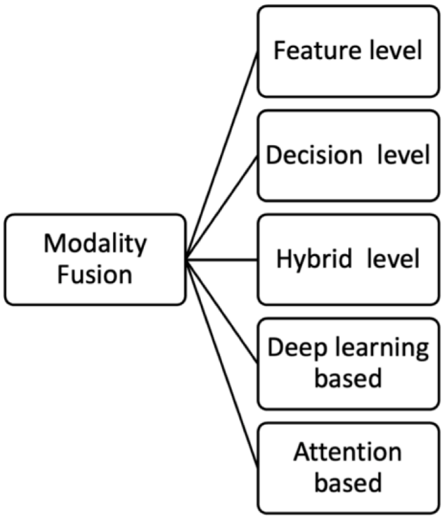

Fig. 8. Taxonomy of data fusion methods for multimodal emotion recognition.

performance.

8.2. Decision-level fusion

When using multiple, possibly different, classifiers in the multimodal emotion recognition, integrating the results of the individual classifiers is known as a decision-level fusion, or late fusion. In machine learning, combining classifiers and their results into a single decision is known as ensemble learning. The term "decision-level fusion" refers to a set of techniques for combining and ensemble the results into a single decision. To achieve decision-level fusion, three principles must be followed: the sum strategy, the maximum strategy, and the logistic regression technique. The sum approach predicts the emotion result based on the combination of multiple classifiers' probabilities. In the maximum strategy, the prediction of emotion is based on the maximum output probability of the individual classifiers. A logistic regression technique predicts the emotion based on the weighted sum of output probabilities of the individual classifiers. Decision-level fusion has an advantage that it allows each modality to choose the optimal classifier for the task (Liu et al., 2021).

Sahoo & Routray (2016) developed an audiovisual emotion identification system that employs decision level fusion of two modalities. Separate emotion identification systems based on facial expressions and speech were initially created and tested separately. Two standard speech emotion datasets were used to test the speech emotion recognition system: the Berlin EMODB database and the Assamese database (Sahoo & Routray, 2016). The eNTREFACE'05 database was used to assess the effectiveness of a visual emotion recognition system (Picard et al., 2015). Then, at the decision level, the speech and visual input were combined based on decision rules (Sahoo & Routray, 2016). Verification of the system was carried out in this study using emotion databases acquired in a lab setting with no background noise. The shortcoming of this methods is when applied to different data, the same decision rule with the same threshold values may not operate.

8.3. Hybrid level fusion

Hybrid fusion combines feature-level and decision-level fusions. For example, one classifier might predict the emotion result of two modalities, e.g., speech and facial expression, using the feature-level fusion, while the result of other modalities, e.g., say physiological signal, is predicted by decision-level fusion. The final prediction of emotion is the integration of the prediction results of feature-level and decision-level results (Cimtay et al., 2020). Tan et al. (2021) merged feature- level and decision-level fusion results of a hierarchical classifier that is created to recognize emotion. To categorize emotional states with compressed sensory representation, Georgescu & Ionescu (2019) created a hybrid classifier to integrate the results of support vector machines and fuzzy cognitive map. To fuse EEG and eye movement features, Lu et al. (2015) used hybrid level fusion of several modalities to improve recognition accuracy. It was shown that EEG has complimentary emotion prediction effect to eye movements. The usage of non-standard dataset can lead to inconsistency in results (Lu et al., 2015). Cimtay et al. (2020) used biometric and soft modalities that showed an improved recognition accuracy. For enhanced accuracy, physiological modalities evaluated at different time intervals are combined. The model was also shown to perform with high accuracy with natural deceptive facial expressions (Cimtay et al., 2020).

Mittal et al. (2020) proposed M3ER, a learning-based approach for emotion identification from multiple input modalities. M3ER represents a unique, data-driven hybrid fusion algorithm that learns to boost more trustworthy cues while suppressing others on a per-sample basis in order to integrate the modalities. Because M3ER contains a check step that uses Canonical Correlational Analysis to discern between ineffective and effective modalities, it is immune to sensor noise. M3ER develops proxy features in the absence of ineffective modalities. For each class, the authors now employ binary categorization; nevertheless, human perception is highly subjective, so a probability distribution over these unique emotions would be more equivalent to a probability distribution over these distinct emotions. This method, which incorporates data from a variety of co-occurring modalities (such as face, text, and speech), is more resistant to sensor noise in any of the modalities than previous systems (Mittal et al., 2020).

8.4. Deep learning based fusion

Different multimodal fusion methods have been developed for deep learning models. Qiu et al. (2018) suggested the use of deep canonical correlation analysis (DCCA) to extract strongly associated high-level features of two modalities. It was found that DCCA when used with deep learning, improved the emotion recognition. DCCA intro- duced coordinated representation into multimodal emotion identification, as well as a new technique of representing multimodal information in high-level fusion features (Qiu et al., 2018). The fundamental limitation of DCCA is that it only allows two input modalities. In addition, fusion parameters must be set based on their prior expertise.

Gogate et al. (2017) developed a deep learning based two-level fusion strategy. In this model, the first level uses CNN-based compression to extract unimodal features. The CNN model is pre-trained on individual unimodal features to improve its emotion classification. In the second step, the optimized unimodal features are concatenated and fed as input to the second level of the classifier. This results in an improved emotion recognition into four classes: Happiness, Sadness, Anger, and Neutral (Gogate et al., 2017). However, as compared to other optimizations techniques, its recognition accuracy is lower.

8.5. Attention-based fusion

The attention mechanism enables a neural network to acquire adaptive fusion weights for multiple modalities, resulting in an improved multimodal fusion and emotion identification. To instruct the classifier regarding which bits of the input are more important to the output class, a self-attention method particular to different classifiers has been provided (Lan et al., 2020). To enhance the original DCCA model, Liu et al. (2021) proposed two multimodal fusion methods: attention-based, and weighted-sum. In the attention-based fusion approach, the weights are computed adaptively. However, in the weighted-sum fusion approach, weights of different modalities are determined by users.

The multimodal emotion recognition performance was evaluated based on the following five datasets (see Table 5): SEED, SEED-IV, SEED-V, DEAP, and DREAMER. The findings of the experiments on these five datasets show that DCCA and BDAE outperform existing multimodal fusion approaches (Liu et al., 2021). However, the DCCA metric employed in this study can only fuse two modalities at a time, which limits the DCCA method’s usage in real-world situations where more than two modalities are fused at the same time.

Lan et al. (2020) combined deep generalized canonical correlation analysis with an attention mechanism (DGCCA-AM) to improve multimodal emotion recognition. This model is shown to outperform other multimodal emotion recognition systems. DGCCA-AM captures emotion-related information from various modalities while filtering out noise by modifying the weights of a deep learning classifier to maximize the correlation between different modalities (Hori et al., 2017), The attention model responds selectively to various modalities of input, such as picture, motion, and auditory elements, rather than only to specific times. Recognition accuracy is less compared to other methods.

By analyzing the capabilities of different fusion methods, the best data fusion methods suitable for multimodal emotion recognition are feature level, decision level and attention-based fusion. Table 10 presents different data fusion methods for emotion recognition. Table 11 presents the comparison of prominent data fusion methods: Feature, Decision and Attention based.

9. Challenges and future directions

This section addresses the final research question (RQ6) in which various challenges and future scope in the emotion recognition are discussed. We formulated our survey using six research questions. Our analysis showed that sensors and virtual reality headsets are the most

Table 10
Different fusion methods for emotion recognition.

Authors	Model	Dataset	Fusion method	Modalities
Y.T Lan et al. (Lan et al., 2020)	Deep generalized canonical correlation analysis with an attention mechanism (DGCCA-AM)	SEED V	Attention based	EEG, eye image (EIG), and eye movement (EYE).
J.L Qiu et al. (Qiu et al., 2018)	Deep Canonical Correlation Analysis (DCCA)	SEED, DEAP	Deep learning	EEG and eye movements.
Y.Lu et al. (Lu et al., 2015)	SVM with fuzzy integral fusion	Own dataset	Hybrid fusion	eye movements and EEG.
V.Chaparro et al. (Chaparro et al., 2018)	Neural network and random forest	MAHNOB - HCI	Feature level	EEG and facial expression.
Y.Cimtay et al. (Cimtay et al., 2020)	CNN, Decision tree	LUMED-2	Hybrid level	facial expressions, galvanic skin response (GSR) and electroencephalogram (EEG).
K.Zhang et al. (Zhang et al., 2021)	MERDCCA	IEMOCAP, CMU-MOSI	Feature level	text, audio and visual.
W. Liu et al. (Liu et al., 2021)	DCCA, BDAE	SEED,DEAP, DREAMER	Attention based, Decision level	EEG, ECG, eye movements.
T.Mittal et al. (Mittal et al., 2020)	M3ER	IEMOCAP, CMUMOSEI	Hybrid level	face, speech, and text
S.Sahoo et al. (Sahoo & Routray, 2016)	MFCC, LBP	Berlin EMODB and eNTREFACE’05	Decision level	speech and facial expression.
M. Gogate et al. (Gogate et al., 2017)	Compression-based Optimized Multimodal Fusion Approach using CNN, MLP	IEMOCAP	Deep learning based	Voice, text, image.
Hori et al. (2017)	RNN	YouTube2Text, MSRVTT	Attention based	audio, image, motion features.

Table 11
Comparison of feature level, decision level and attention based fusion methods.

Feature level	Decision level	Attention based
It takes advantage of the early correlation between several features, which often result in better job com- pletion.	The feature-level correlation between modalities is underutilized.	The local attention mechanism can discern between the emotion of dif- ferent images and can make full use of the emotional features of differ- *ent* images.
Fusion of modalities of different sizes/dimensions are difficult	Late fusion local decisions have the same representation, making fusion easier.	The global attention mechanism can discriminate between distinct modes' emotional differences and make full use of the emotional characteristics of each mode.
Uneven feature dimensions result in non-uniform neural network weight distribution, resulting in poor col- lective learning.	Good collective learning is possible by decision level fusion	The network can detect interference such as noise and factors triggering uncertainty in each modality and dynamically downweight the less confident modalities, allowing for good collective learning.
Simple techniques are incapable of extracting redundant information from many modalities such as audio, video, and text, or contextual information between video utterances	Simple techniques are incapable of extracting redundant information from several modalities such as au- dio, video, and text, or contextual information between video utterances.	Extract the contextual information among the utterances before fusion.

effective tools for the data acquisition required for emotion recognition. It also showed that SVM is the most efficient machine classifier, whereas CNN is the most effective deep learning classifier. In terms of data fusion, our analysis identified Attention-based fusion as the most effective technique compared to feature and decision level-based fusion. Our analysis also identified a number of challenges that are needed to be overcome in this area that once solved will have a direct impact on the recognition accuracy and time taken for recognition. The most relevant challenges that we figured out in different emotion modalities during our review are explained in following sections. Table 12 illustrates the percentage of research papers reviewed to address different challenges.

9.1. Speech /audio emotion recognition challenges

The majority of the utterances in natural speech data sets come from talk shows, contact center recordings, and other circumstances where both parties are aware that they are being recorded. The speech datasets collected may not fully reflect how people feel because they do not capture all emotions. Furthermore, there are difficulties with the utterances' tagging. Human annotators label the speech data once the utterances are captured. It's possible that the speaker's true emotion differs from the feelings observed by human annotators. Human annotator identification rates aren't more than 90 percent (Mehmet Berkehan & Oğuz, 2020; Malla et al., 2020).

Table 12
Different challenges in emotion recognition.

Challenges	Percentage of articles reviewed that mention a particular challenge
Speech/ audio emotion recognition challenges	12 percent
Image emotion recognition challenges	12 percent
Video emotion recognition challenges	11 percent
Facial emotion recognition challenges	18 percent
Multimodal emotion recognition challenges	47 percent

9.2. Image/video/facial emotion recognition challenges

Emotion subjectivity: Images are subjective in terms of human perception and comprehension. Different people experience different feelings while viewing comparable visuals, and even the same person experiences different emotions when viewing the same image at different times. A generalized model is needed to learn more objective image emotions with less human bias (Zhou et al., 2021).

Generalized classifier: Deep learning and machine learning classifiers that are trained well on one particular dataset usually do not perform well on another dataset especially when data distributions are extremely different. Developing a generic classifier that can perform across different datasets is a challenge especially faced by image recognition techniques (Zhou et al., 2021).

Compound emotions: Most of the ongoing research focuses detection of simple emotions such as joy, sorrow, sad, happy, anger, fear and surprise. But fine-grained compound emotions such as pain, happily surprised, angrily disgusted are difficult to detect (Kuruvayil & Palaniswamy, 2021).

9.3. Multimodal emotion recognition challenges

Dataset availability: Although most research suggest that the recognition rate increases with the usage of multiple modalities (image, physiological signals, speech, video), the best combination of modalities has yet to be found. Furthermore, very few multimodal emotion datasets are available that address most of the conceivable modalities in unrestricted real life scenarios. (Bota et al., 2019).

Emotion deviations: While all multimodal cues are present, the emotions elicited by a combination of signals may be vastly different. As a result, future research will concentrate on using CNN to simulate and analyze various mismatch scenarios (Song & Kim, 2021).

Emotional misclassification: In the classification of distinct emotions, there are several underlying challenges. Some emotions are more difficult to recognize compared to others because their expressions are more subtle, or their patterns are less obvious. The misclassification of emotion features makes emotion recognition and accuracy more prone to error (Küntzler et al., 2021).

9.4. Future directions

Multi label emotion reasoning: To construct data samples, at least 2 different subjects must agree on the exact label. It is, nevertheless, conceivable to be sad or furious at the same time. As a result, using MemoR to perform multi-label emotion reasoning is a promising future direction. Emotion detection based on physiological signals, audiovisual group emotion recognition, and driver gaze prediction due to the importance of the job in online learning, engagement prediction in the wild is also an important future direction (Shen et al., 2020).

Genetic or swarm based optimization: Most of the emotion research has paid less attention to the optimization of elements that affect multimodal fusion performance, such as dimensionality reduction and optimal fusion procedures. The most appropriate fusion or expansion level for the best possible performance and accuracy can be determined by integrating reasoning through modern genetic or swarm-based optimization (Gogate et al., 2017).

Human robot interaction (HRI): Multimodal techniques will be critical in improving emotion identification performance over single-modality techniques, necessitating the development of ML methods and DL structures to deal with heterogeneous data. Data used to train and test emotion recognition requires special attention: HRI has some critical and problematic characteristics that could render data obtained in controlled conditions or from various contexts inappropriate for real-world HRI applications. A dataset from an actual HRI can be developed in future (Spezialetti et al., 2020).

Enhancing facial emotion recognition: The recognition of non-standardized facial emotions, which may be presented in more authentic situations, is a major shortcoming of facial emotion recognition. As a result, more research is needed to improve face recognition as a study aid in classifying non-prototypical and delicate facial gestures. Although the system appears to be a suitable fit for non-contact evaluation of emotional facial expressions, researchers should be cautious when interpreting data from non-prototypical emotions in non-restrictive situations such as significant head movement or partial occlusion (Küntzler et al., 2021).

Brain-IoT: The usage of a Brain-IoT-based system is to detect negative emotions such as fear, loneliness, sadness and anger and then plays a movie to stimulate happy emotions. More tests should be created in order to create a system that can recognize and monitor emotions in real time. This unique method can be used specially in online learning particularly in scenarios such as boredom, or the lack of concentration among students (Shirke et al., 2020).

Active deep learning (ADL): ADL is similar to machine learning-based active learning in terms of concept. The sole distinction between ADL and machine learning-based active learning is that ADL uses a deep neural network-based methodology. Because of the large amount of data, the deep neural network-based ADL may achieve two significant benefits: scalability and resilience. As a result, the ADL-based method will be suitable and effective for a wide range of training scenarios. Furthermore, this strategy can be useful for training a deep neural network with a limited number of labeled training instances, but not all of them. Using ADL to train a deep neural network for speech emotion recognition can drastically minimize the amount of time it takes a human to manually categorize all of the speaker utterances (Jahangir et al., 2022).

10. Conclusion

Emotion acquisition tools and feature classifiers plays a major role in accuracy of emotion recognition process. We performed a fine-grained emotion analysis to show different inner states of human emotions. We have enlisted reliable list of emotion acquisition tools that focus on sensing technology and virtual reality and provided a comprehensive detail of publicly available dataset for multimodal emotion recognition. We explained different deep learning and machine learning classifiers with enhanced recognition accuracy. The commonly used data fusion methods are analyzed in detail since handling high dimensional features of multimodal emotion recognition is a challenge. The major research gap identified during the study was the absence of policies to address privacy issues and the biased nature of ML/DL models. More research should be done to avoid misclassifying emotion traits, which could produce inaccurate results. An ideal multimodal dataset that takes into account real world scenarios should be developed. Further investigation in optimization techniques is prescribed to obtain highly efficient emotion recognition techniques.

Funding

This work is completely funded internally by the University of Sharjah.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

Abdullah, S. M. S. A., Ameen, S. Y. A., Sadeeq, M. A. M., & Zeebaree, S. (2021). Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends, 2*(02), 52–58.

Aiswaryadevi, V. J., Priyanka, G., Sathya Bama, S., Kiruthika, S., Soundarya, S., Sruthi, M., et al. (2021). Smart iot multimodal emotion recognition system using deep learning networks. *Artificial Intelligence and IoT Smart Convergence for Eco Friendly Topography*, 3–19.

Alswaidan, N., & El Bachir Menai, M. (2020). A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge & Information Systems, 62*(8).

Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(2), 423–443.

Bazgir, O., Mohammadi, Z., & Habibi, S. A. H. (2018). Emotion recognition with machine learning using eeg signals. In *Proceedings of the 25th national and 3rd international iranian conference on biomedical engineering (ICBME)* (pp. 1–5). IEEE.

Bosch, E. J., Käthner, D., & Drewitz, U. (2021). Evidence for individual-specific expressions of frustration. *Proceedings of the European Society for Cognitive and Affective Neuroscience*. https://elib.dlr.de/143073/.

Bota, P. J., Wang, C., Fred, A. L. N., & Silva, HPDa (2019). A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals. *IEEE Access Practical Innovations Open Solutions, 7*, 140990–141020.

Cen, L., Wu, F., Yu, Z. L., & Hu, F. (2016). A real-time speech emotion recognition system and its application in online learning. *Emotions, technology, design, and learning* (pp. 27–46). Elsevier.

Chaparro, V., Gomez, A., Salgado, A., Lucia Quintero, O., Lopez, N., & Villa, L. F. (2018). Emotion recognition from eeg and facial expressions: A multimodal approach. In *Proceedings of the 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 530–533). IEEE.

Chen, P., & Zhang, J. (2017). Performance comparison of machine learning algorithms for eeg-signal-based emotion recognition. In *Proceedings of the international conference on artificial neural networks* (pp. 208–216). Springer.

Cimtay, Y., Ekmekcioglu, E., & Caglar-Ozhan, S. (2020). Cross-subject multimodal emotion recognition based on hybrid fusion. *IEEE Access Practical Innovations Open Solutions, 8*, 168865–168878.

Dadebayev, D., Goh, W. W., & Tan, E. (2021). *EEG-based emotion recognition: Review of commercial EEG devices and machine learning techniques*. Journal of King Saud University-Computer and Information Sciences.

Dai, Y., Wang, X., Li, X., & Zhang, P. (2015). Reputation-driven multimodal emotion recognition in wearable biosensor network. In *Proceedings of the IEEE international instrumentation and measurement technology conference (I2MTC) proceedings* (pp. 1747–1752). IEEE.

Du, G., Long, S., & Yuan, H. (2020). Non-contact emotion recognition combining heart rate and facial expression for interactive gaming environments. *IEEE Access Practical Innovations Open Solutions, 8*, 11896–11906.

Elor, A., Powell, M., Mahmoodi, E., Hawthorne, N., Teodorescu, M., & Kurniawan, S. (2020). On shooting stars: Comparing cave and hmd immersive virtual reality exergaming for adults with mixed ability. *ACM Transactions on Computing for Healthcare, 1*(4), 1–22.

Fabricio, B., & Liang, Z. (2013). Fuzzy community structure detection by particle competition and cooperation. *Soft Computing, 17*, 659–673.

A. Forsberg, Shopping for emotion-evaluating the usefulness of emotion recognition data from a retail perspective, MSc Thesis, Umea University, Faculty of Science and Technology, Department of Computing Science, 2017.

Georgescu, M. I., & Ionescu, R. T. (2019). Recognizing facial expressions of occluded faces using convolutional neural networks. In *Proceedings of the international conference on neural information processing* (pp. 645–653). Springer.

Gjoreski, H., Mavridou, II., Fatoorechi, M., Kiprijanovska, I., Gjoreski, M., Cox, G., et al. (2021). emteqpro: Face-mounted mask for emotion recognition and affective computing. In *Proceedings of the Adjunct ACM international joint conference on pervasive and ubiquitous computing and proceedings of the ACM international symposium on wearable computers* (pp. 23–25).

Gogate, M., Adeel, A., & Hussain, A. (2017). A novel brain-inspired compression-based optimised multimodal fusion for emotion recognition. In *Proceedings of the IEEE symposium series on computational intelligence (SSCI)* (pp. 1–7). IEEE.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., et al. (2013). Challenges in representation learning: A report on three machine learning contests. In *Proceedings of the international conference on neural information processing* (pp. 117–124). Springer.

Hassan, M. M., Alam, MdGR, Uddin, MdZ, Huda, S., Almogren, A., & Fortino, G. (2019). Human emotion recognition using deep belief network architecture. *Information Fusion, 51*, 10–18.

Hickson, S., Dufour, N., Sud, A., Kwatra, V., & Essa, I. (2019). Eyemotion: Classifying facial expressions in vr using eye-tracking cameras. In *Proceedings of the IEEE winter conference on applications of computer vision (WACV)* (pp. 1626–1635). IEEE.

Hinkle, L., Khoshhal, K., & Metsis, V. (2019). Physiological measurement for emotion recognition in virtual reality. In *Proceedings of the 2nd international conference on data intelligence and security (ICDIS)* (pp. 136–143). IEEE.

Hipson, W. E., Kiritchenko, S., Mohammad, S. M., & Coplan, R. J. (2021). Examining the language of solitude versus loneliness in tweets. *Journal of Social and Personal Relationships, 38*(5), 1596–1610.

Hori, C., Hori, T., Lee, T. Y., Zhang, Z., Harsham, B., Hershey, J. R., et al. (2017). Attention-based multimodal fusion for video description. In *Proceedings of the IEEE international conference on computer vision* (pp. 4193–4202).

Houshmand, B., & Khan, N. M. (2020). Facial expression recognition under partial occlusion from virtual reality headsets based on transfer learning. In *Proceedings of the IEEE 6th international conference on multimedia big data (BigMM)* (pp. 70–75). IEEE.

Imani, M., & Montazer, G. A. (2019). A survey of emotion recognition methods with emphasis on e-learning environments. *Journal of Network and Computer Applications, 147*, Article 102423.

Jahangir, R., Teh, Y. W., Hanif, F., & Mujtaba, G. (2022). Deep learning approaches for speech emotion recognition: State of the art and research challenges. *Multimedia Tools and Applications, 80*(16), 23745–23812.

Jam, G. S., Rhim, J., & Lim, A. (2021). Developing a data-driven categorical taxonomy of emotional expressions in real world human robot interactions. In *Proceedings of the companion of the ACM/IEEE international conference on human-robot interaction* (pp. 479–483).

JohnPaul Quilingking Tomas, R., Jamilla, AS., Lopo, KS., & Camba, CE. (2020). Multimodal emotion detection model implementing late fusion of audio and lyrics in filipino music. In *Proceedings of the the 3rd international conference on computing and big data* (pp. 78–84).

Küntzler, T., Höfling, T. T. A., & Alpers, G. W. (2021). Automatic facial expression recognition in standardized and non-standardized emotional expressions. *Frontiers in psychology, 5*(12), 1086. https://doi.org/10.3389/fpsyg.2021.627561

Kołakowska, A., Landowska, A., Szwoch, M., Szwoch, W., & Wrobel, M. R. (2014). Emotion recognition and its applications. In *Human-Computer systems interaction: Backgrounds and applications, 3* pp. 51–62). Springer.

Koelstra, S., Muhl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., et al. (2011). Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing, 3*(1), 18–31.

Kuruvayil, S., & Palaniswamy, S. (2021). Emotion recognition from facial images with simultaneous occlusion, pose and illumination variations using meta-learning. *Journal of King Saud University Computer and Information Sciences,, 34*(9), 7271–7282.

Kusumaningrum, T. D., Faqih, A., & Kusumoputro, B. (2020). Emotion recognition based on deap database using eeg time-frequency features and machine learning methods. *Journal of Physics Conference Series, 1501*, Article 012020. IOP Publishing.

Lan, Y. T., Liu, W., & Lu, B. L. (2020). Multimodal emotion recognition using deep generalized canonical correlation analysis with an attention mechanism. In *, 1–6. Proceedings of the international joint conferences on neural networks (IJCNN)*. IEEE.

Lim, J. Z., Mountstephens, J., & Teo, J. (2020). Emotion recognition using eye-tracking: Taxonomy, review and current challenges. *Sensors, 20*(8), 2384.

Liu, W., Pellegrini, M., & Wang, X. (2014). Detecting communities based on network topology. *Scientific Reports, 4*, 5739.

Liu, W., Zheng, W. L., & Lu, B. L. (2016). Emotion recognition using multimodal deep learning. In *Proceedings of the international conference on neural information processing* (pp. 521–529). Springer.

Liu, W., Qiu, J. L., Zheng, W. L., & Lu, B. L. (2021). Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems, 14*(2), 715–729. https://doi.org/10.1109/TCDS.2021.3071170

Long, F., Zhao, S., Wei, X., Ng, S. C., Ni, X., Chi, A., et al. (2021). Positive and negative emotion classification based on multi-channel. *Frontiers in Behavioral Neuroscience, 15*, 71–81.

Lorenzo, G., Lledó, A., Pomares, J., & Roig, R. (2016). Design and application of an immersive virtual reality system to enhance emotional skills for children with autism spectrum disorders. *Computers & Education, 98*, 192–205.

Lu, Y., Zheng, W. L., Li, B., & Lu, B. L. (2015). Combining eye movements and EEG to enhance emotion recognition. In *Proceedings of the 24th international joint conference on artificial intelligence*.

Ma, J., Tang, H., Zheng, W. L., & Lu, B. L. (2019). Emotion recognition using multimodal residual LSTM network. In *Proceedings of the 27th ACM international conference on multimedia* (pp. 176–183).

Majid Mehmood, R., Du, R., & Lee, H. J. (2017). Optimal feature selection and deep learning ensembles method for emotion recognition from human brain EEG sensors. *IEEE Access Practical Innovations Open Solutions, 5*, 14797–14806.

Malla, S., Alsadoon, A., & Bajaj, S. K. (2020). A dfc taxonomy of speech emotion recognition based on convolutional neural network from speech signal. In *Proceedings of the 5th international conference on innovative technologies in intelligent systems and industrial applications (CITISIA)* (pp. 1–10). IEEE.

Marín-Morales, J., Higuera-Trujillo, J. L., Greco, A., Guixeres, J., Llinares, C., Scilingo, E. P., et al. (2018). Affective computing in virtual reality: Emotion recognition from brain and heartbeat dynamics using wearable sensors. *Scientific Reports, 8*(1), 1–15.

Marín-Morales, J., Llinares, C., Guixeres, J., & Alcañiz, M. (2020). Emotion recognition in immersive virtual reality: From statistics to affective computing. *Sensors, 20*(18), 5163.

Marsella, S., & Gratch, J. (2014). Computationally modeling human emotion. *Communications of the ACM, 57*(12), 56–67.

Marsella, S., Gratch, J., Petta, P., et al. (2010). Computational models of emotion. *A Blueprint for Affective Computing A Sourcebook and Manual, 11*(1), 21–46.

Matsuda, Y., Fedotov, D., Takahashi, Y., Arakawa, Y., Yasumoto, K., & Minker, W. (2018). Emotour: Multimodal emotion recognition using physiological and audio-visual features. In *Proceedings of the ACM international joint conference and 2018 international symposium on pervasive and ubiquitous computing and wearable computers* (pp. 946–951).

Mavridou, I., McGhee, J. T., Hamedi, M., Fatoorechi, M., Cleal, A., Ballaguer-Balester, E., et al. (2017). Faceteq interface demo for emotion expression in vr. *2017 IEEE virtual reality (VR)* (pp. 441–442). IEEE.

Mehmet Berkehan, A., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication, 116*, 56–76.

Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020). M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*, 1359–1367.

Monitoring of arduino-based PPG and GSR signals through an android device. *Proceedings of IEEE Engineering in Medicine and Biology Society (EMBS) International Student Conference*, (2016). https://site.ieee.org/embs-isc-2016/files/2016/05/IEEE-EMBS-ISC-2016-Top-15-Design-Competition.pdf.

Mordorintelligence https://www.mordorintelligence.com/industry-reports/emotion-detection-and-recognition-edr-market. 2021.

Nakisa, B., Rastgoo, M. N., Tjondronegoro, D., & Chandran, V. (2018). Evolutionary computation algorithms for feature selection of eeg-based emotion recognition using mobile sensors. *Expert Systems with Applications, 93*, 143–155.

Osuna, E., Rodríguez, L. F., Octavio Gutierrez-Garcia, J., & Castro, L. A. (2020). Development of computational models of emotions: A software engineering perspective. *Cognitive Systems Research, 60*, 1–19.

Park, C. Y., Cha, N., Kang, S., Kim, A., Habib Khandoker, A., Hadjileontiadis, L., et al. (2020). K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data, 7*(1), 1–16.

Petrovica, S., Anohina-Naumeca, A., & Ekenel, H. K. (2017). Emotion recognition in affective tutoring systems: Collection of ground- truth data. *Procedia Computer Science, 104*, 437–444.

Picard, R. W., Vyzas, E., & Healey, J. (2015). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*(10), 1175–1191.

Piskioulis, O., Tzafilkou, K., & Economides, A. (2021). Emotion detection through smartphone's accelerometer and gyroscope sensors. In *Proceedings of the 29th ACM conference on user modeling, adaptation and personalization* (pp. 130–137).

Poria, S., Chaturvedi, I., Cambria, E., & Hussain, A. (2016). Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *Proceedings of the IEEE 16th international conference on data mining (ICDM)* (pp. 439–448). IEEE.

Povolny, F., Matejka, P., Hradis, M., Popková, A., Otrusina, L., Smrz, P., et al. (2016). Multimodal emotion recognition for avec 2016 challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge* (pp. 75–82).

Prasanth, S., Roshni Thanka, M., Bijolin Edwin, E., & Nagaraj, V. (2021). Speech emotion recognition based on machine learning tactics and algorithms. *Proceedings of the International Conference on Emerging Trends in Materials Science, Technology and Engineering,*, 10–16. https://doi.org/10.1016/j.matpr.2020.12.207

Qing, C., Qiao, R., Xu, X., & Cheng, Y. (2019). *Interpretable emotion recognition using EEG signals, 7* pp. 94160–94170). IEEE Access.

Qiu, J. L., Liu, W., & Lu, B. L. (2018). Multi-view emotion recognition using deep canonical correlation analysis. In *Proceedings of the international conference on neural information processing* (pp. 221–231). Springer.

Quiroz, J. C., Geangu, E., & Yong, M. H. (2018). Emotion recognition using smart watch sensor data: Mixed-design study. *JMIR Mental Health, 5*(3), E10153.

Ragot, M., Martin, N., Em, S., Pallamin, N., & Diverrez, J. M. (2017). Emotion recognition using physiological signals: Laboratory vs. wearable sensors. In *Proceedings of the international conference on applied human factors and ergonomics* (pp. 15–22). Springer.

Ranganathan, H., Chakraborty, S., & Panchanathan, S. (2016). Multimodal emotion recognition using deep learning architectures. In *Proceedings of the IEEE winter conference on applications of computer vision (WACV)* (pp. 1–9). IEEE.

Sahoo, S., & Routray, A. (2016). Emotion recognition from audio-visual data using rule based decision level fusion. In *Proceedings of the IEEE students' technology symposium (TechSym)* (pp. 7–12).

Salazar, C., Montoya-Múnera, E., & Aguilar, J. (2021). Analysis of different affective state multimodal recognition approaches with missing data-oriented to virtual learning environments. *Heliyon, 7*(6), Article e07253. https://doi.org/10.1016/j.heliyon.2021.e07253

Sato, W., Murata, K., Uraoka, Y., Shibata, K., Yoshikawa, S., & Furuta, M. (2021). Emotional valence sensing using a wearable facial emg device. *Scientific Reports, 11* (1), 1–11.

Shamim Hossain, M., & Muhammad, G. (2019). Emotion recognition using secure edge and cloud computing. *Information Sciences, 504*, 589–601.

Sharma, G., & Dhall, A. (2021). A survey on automatic multimodal emotion recognition in the wild. *Advances in data science: Methodologies and applications* (pp. 35–64). Springer.

Shen, G., Wang, X., Duan, X., Li, H., & Zhu, W. (2020). Memor: A dataset for multimodal emotion reasoning in videos. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 493–502).

Shirke, B., Wong, J., Libut, J. C., George, K., & Oh, S. J. (2020). Brain-iot based emotion recognition system. In *Proceedings of the 10th annual computing and communication workshop and conference (CCWC)* (pp. 0991–0995). IEEE.

Shukla, J., Barreda-Angeles, M., Oliver, J., Nandi, G. C., & Puig, D. (2019). Feature extraction and selection for emotion recognition from electrodermal activity. *IEEE Transactions on Affective Computing, 12*(4), 857–869.

Singh, J., Gill, R., et al. (2022). Multimodal emotion recognition system using machine learning and psychological signals: A review. *Soft Computing Theories and Applications*, 657–666.

Soleymani, M., Lichtenauer, J., Pun, T., & Pantic, M. (2011). A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing, 3* (1), 42–55.

Song, B. C., & Kim, DHa (2021). Hidden emotion detection using multi-modal signals. In *Proceedings of the extended abstracts of the 2021 CHI conference on human factors in computing systems* (pp. 1–7).

Spezialetti, M., Placidi, G., & Rossi, S. (2020). Emotion recognition for human-robot interaction: Recent advances and future perspectives. *Frontiers in Robotics and AI, 7*.

Subasi, A., Tuncer, T., Dogan, S., Tanko, D., & Sakoglu, U. (2021). Eeg-based emotion recognition using tunable q wavelet transform and rotation forest ensemble classifier. *Biomedical Signal Processing and Control, 68*, Article 102648.

Suja, P., Tripathi, S., et al. (2016). Real-time emotion recognition from facial images using raspberry pi ii. In *Proceedings of the 3rd international conference on signal processing and integrated networks (SPIN)* (pp. 666–670). IEEE.

Tabbaa, L., Searle, R., Bafti, S. M., Hossain, MdM, Intarasisrisawat, J., Glancy, M., et al. (2021). Vreed: Virtual reality emotion recognition dataset using eye tracking & physiological measures. In *, 5*. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* (pp. 1–20).

Tan, Y., Sun, Z., Duan, F., Solé-Casals, J., & Caiafa, C. F. (2021). A multimodal emotion recognition method based on facial expressions and electroencephalography. *Biomedical Signal Processing and Control, 70*, Article 103029.

Tang, H., Liu, W., Zheng, W. L., & Lu, B. L. (2017). Multimodal emotion recognition using deep neural networks. In *Proceedings of the international conference on neural information processing* (pp. 811–819). Springer.

Tang, T. B., Chong, J. S., Kiguchi, M., Funane, T., & Lu, C. K. (2021). Detection of emotional sensitivity using fnirs based dynamic functional connectivity. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 29*, 894–904.

Tashu, T. M., Hajiyeva, S., & Horvath, T. (2021). Multimodal emotion recognition from art using sequential co-attention. *Journal of Imaging, 7*(8), 157.

Techxplore https://techxplore.com/news/2019-09-method-emotion-recognition-gaming.html. 2021.

Torres, A. D., Yan, H., Haj Aboutalebi, A., Das, A., Duan, L., & Rad, P. (2018). Patient facial emotion recognition and sentiment analysis using secure cloud with hardware acceleration. *In computational intelligence for multimedia big data on the cloud with engineering applications* (pp. 61–89). Elsevier.

Udovičić, G., Ðerek, J., Russo, M., & Sikora, M. (2017). Wearable emotion recognition system based on gsr and ppg signals. In *Proceedings of the 2nd international workshop on multimedia for personal health and health care* (pp. 53–59).

Vergara, D., Rubio, M. P., Lorenzo, M., & Rodríguez, S. (2019). On the importance of the design of virtual reality learning environments. In *Proceedings of the international conference in methodologies and intelligent systems for techhnology enhanced learning* (pp. 146–152). Springer.

Wu, C. H., Lin, J. C., & Wei, W. (2014). Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies. *APSIPA Transactions on Signal and Information Processing, 3*.

Xie, Z., & Guan, L. (2013). Multimodal information fusion of audiovisual emotion recognition using novel information theoretic tools. In *Proceedings of the IEEE international conference on multimedia and expo (ICME)* (pp. 1–6). IEEE.

Yang, C. J., Fahier, N., He, C. Y., Li, W. C., & Fang, W. C. (2020). An ai-edge platform with multimodal wearable physiological signals monitoring sensors for affective computing applications. In *Proceedings of the IEEE international symposium on circuits and systems (ISCAS)* (pp. 1–5). IEEE.

Yang, K., Wang, C., Gu, Y., Sarsenbayeva, Z., Tag, B., Dingler, T., et al. (2021). Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition. *IEEE Transactions on Affective Computing,, 13*(4), 1–15.

Zhang, K., Zhang, H., Li, S., Yang, C., & Sun, L. (2018). The pmemo dataset for music emotion recognition. In *Proceedings of the 2018 ACM on international conference on multimedia retrieval* (pp. 135–142).

Zhang, X., Wang, M. J., & Guo, X-Da (2020a). Multi-modal emotion recognition based on deep learning in speech, video and text. In *Proceedings of the IEEE 5th international conference on signal and image processing (ICSIP)* (pp. 328–333).

Zhang, T., Liu, M., Yuan, T., & Al-Nabhan, N. (2020b). Emotion-aware and intelligent internet of medical things towards emotion recognition during covid-19 pandemic. *IEEE Internet of Things Journal, 8*, 16002–16013. https://doi.org/10.1109/JIOT.2020.3038631. no. 21, 1 Nov.1, 2021.

Zhang, Ke, Li, Y., Wang, J., Wang, Z., & Li, X. (2021). Feature fusion for multimodal emotion recognition based on deep canonical correlation analysis. *IEEE Signal Processing Letters, 28*, 1898–1902.

Zhang, H. (2020). Expression-eeg based collaborative multimodal emotion recognition using deep autoencoder. *IEEE access Practical Innovations Open Solutions, 8*, 164130–164143.

Zheng, W. L., & Lu, B. L. (2015). Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development, 7*(3), 162–175.

Zhou, F., Cao, C., Zhong, T., & Geng, J. (2021). Learning meta-knowledge for few-shot image emotion recognition. *Expert Systems with Applications, 168*, Article 114274.