IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 4 July 2024, accepted 15 July 2024, date of publication 18 July 2024, date of current version 6 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3430850

SURVEY

A Systematic Review on Multimodal Emotion Recognition: Building Blocks, Current State, Applications, and Challenges

SEPIDEH KALATEH, LUIS A. ESTRADA-JIMENEZ, SANAZ NIKGHADAM-HOJJATI, (Member, IEEE), AND JOSE BARATA, (Member, IEEE)

Centre of Technology and Systems (CTS-UNINOVA), 2829-516 Caparica, Portugal
Associated Laboratory on Intelligent Systems (LASI), 2829-516 Caparica, Portugal
Department of Electrical Engineering, NOVA School of Science and Technology, NOVA University of Lisbon, 1099-085 Lisbon, Portugal

Corresponding author: Sepideh Kalateh (sepideh.kalateh@uninova.pt)

This work was supported in part by Fundação para a Ciência e Tecnologia under Grant UIDB/00066/2020, and in part by the Center of Technology and Systems (CTS).

ABSTRACT Emotion recognition involves accurately interpreting human emotions from various sources and modalities, including questionnaires, verbal, and physiological signals. With its broad applications in affective computing, computational creativity, human-robot interactions, and market research, the field has seen a surge in interest in recent years. This paper presents a systematic review of multimodal emotion recognition (MER) techniques developed from 2014 to 2024, encompassing verbal, physiological signals, facial, body gesture, and speech as well as emerging methods like sketches emotion recognition. The review explores various emotion models, distinguishing between emotions, feelings, sentiments, and moods, along with human emotional expression, categorized in both artistic and non-verbal ways. It also discusses the background of automated emotion recognition systems and introduces seven criteria for evaluating modalities alongside a current state analysis of MER, drawn from the human-centric perspective of this field. By selecting the PRISMA guidelines and carefully analyzing 45 selected articles, this review provides comprehensive perspectives into existing studies, datasets, technical approaches, identified gaps, and future directions in MER. It also highlights existing challenges and current applications of the MER.

INDEX TERMS Multimodal emotion recognition, artificial intelligence, affective computing, emotion recognition, deep learning, machine learning, emotion expression.

GLOSSARY

A	Arousal.
A-V	Audio-Visual.
Acc	Accuracy.
AI	Artificial Intelligence.
ANN	Artificial Neural Networks.
AUC	Area Under the ROC Curve.
CNN	Convolutional Neural Network.
EF	Early fusion.
ER	Emotion Recognition.
F1	F1 measure.
GAN	Generative Adversarial Network.
GNN	Graph Neural Network.
KNN	k-Nearest Neighbors.
LF	Late fusion.
LLM	Large Language Model.
LSTMLong	Short-Term Memory.
MER	Multimodal Emotion Recognition.
P	Precision.
PCA	principal component analysis.
PSD	Persuasive Systems Design.

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

R	Recall.
RNN	Recurrent neural network.
ROC	Receiver Operating Characteristic.
SSL	Self-Supervised Learning.
SVM	Support Vector Machine.
TAM	Technology Acceptance Model.
V	Valence.

I. INTRODUCTION

Emotions can be regarded as a universal language, transcending cultural and linguistic boundaries, despite evident variations in its expression across different societies and contexts [1]. While there are notable differences in how emotions are manifested, such as cultural norms influencing displays of affection or grief, there are also significant similarities that underscore common human experiences [2], [3], [4]. For instance, the expression of joy often involves smiling, regardless of cultural background, highlighting a fundamental similarity in emotional expression [5]. These similarities extend beyond human societies and can be observed in certain animal species, suggesting a shared evolutionary basis for emotional behaviors [6], [7].

These parallels in emotional expression across diverse cultures underscore the universality of emotions and their evolutionary significance in facilitating social bonds and survival mechanisms [8]. As research in this field progresses, automated systems for emotion detection are gaining traction, and the need for emotion recognition (ER) is increasingly recognized across a spectrum of domains [9], particularly in the field of human-machine interactions [10].

Accurately perceiving and understanding emotions in interpersonal dynamics can enhance communication, empathy, and conflict resolution [11]. In educational settings, ER can inform personalized learning approaches, adaptive feedback mechanisms, and interventions to support emotional regulation and well-being among students [12]. Gaming platforms are integrating ER technologies to enhance user experience, adapt game-play dynamics, and create more immersive virtual environments [13]. Within the healthcare sector, ER holds promise for applications ranging from mental health monitoring and diagnosis to pain assessment and patient care [14], [15].

Among the emotion detection approaches, Multimodal Emotion Recognition (MER) offers a holistic method for capturing and interpreting human emotions through various modalities such as facial expressions, vocal intonations, gestures, and textual content [9], [16]. MER is crucial for developing empathetic and responsive AI systems, social robots, and virtual assistants that can understand and appropriately respond to human emotions [14]. The diverse applications of ER highlight the need for advanced techniques that can effectively capture and interpret emotional signals across various modalities, ranging from facial expressions and speech to physiological cues and text analysis [17].

The convergence of technological advancements, interdisciplinary collaboration, and the recognition of the importance of emotions in human-computer interaction has fueled the growth of MER as a promising area of research in recent years. The market for emotion detection and recognition was valued at USD 19.87 million in 2020 and is projected to reach USD 52.86 million by 2026, with a notable CAGR of 18.01% during the forecast period (2021-2026) [18].

The diverse applications of ER underscore its pivotal role in enhancing human-machine interactions. However, the effectiveness and reliability of ER systems depend on the quality of research and methodologies employed in this domain. Therefore, conducting a systematic review of existing studies is essential to evaluate the state of the art, identify gaps and challenges, and propose future research directions. Reviewing existing literature can help identify the limitations and challenges faced by current MER systems in accurately recognizing subtle or complex emotions. Identifying knowledge gaps concerning specific emotion categories or contextual factors can direct future research endeavors aimed at refining the precision and reliability of MER systems. Furthermore, recognizing common challenges and limitations associated with generalization and robustness can inform the development of more adaptable and context-sensitive models capable of performing effectively across diverse real-world scenarios. Examining literature on the ethical and societal ramifications of MER highlights critical concerns related to privacy, consent, bias, and the potential misuse of emotional data. Investigating studies exploring user perceptions, attitudes, and preferences regarding the ethical and societal implications of emotion detection systems, alongside the technical and scientific outcomes, offers valuable insights into stakeholders' perspectives. This, in turn, facilitates the development of responsible research practices and guidelines with respect to human-centric approaches.

A. EXISTING EMOTION RECOGNITION REVIEW STUDIES

Although some reviews have explored ER, recent years have witnessed a heightened focus on the topic. However, there remains a notable scarcity of reviews specifically dedicated to multimodal approaches. We aim to fill this gap by examining the recent reviews in MER, identifying overlooked areas and gaps, and incorporating these insights into our research questions and review objectives.

Recent reviews have made significant strides in this area. For example, [19] explores contactless open datasets and unique modality combinations, identifying unaddressed research gaps and suggesting new avenues for research. Similarly, [20] summarizes recent advances in deep learning-based multimodal emotion recognition (DL-MER) across audio, visual, and text modalities. Ahmed et al. [21] provide an overview of emotion acquisition tools, a comparison of datasets, and an exploration of machine and deep learning classifiers for feature extraction, along with an explanation of data fusion methods, aiming to better understand applied ER. Pan et al. [22] review recent

advancements in datasets, preprocessing, feature extraction, and fusion methods, guiding researchers and highlighting future directions in MER. Khare et al. [23], though not specifically focused on MER, offer a thorough review of ER techniques covering various modalities such as physical and physiological signals, alongside different emotion models and stimuli, analyzing 142 journal articles, presenting challenges in the field, and suggesting future research directions.

Earlier reviews have also contributed valuable insights. Siddiqui et al. [24] explore databases used in developing MER systems, covering various data types like facial expressions, speech, physiological signals, body movements, and gestures, including the VIRI database, showcasing its advantages over existing ones. Gu et al. [25] focus on ER, categorized into physiological signals and non-physiological signals, emphasizing direct ER through video and music. They compare fusion techniques in MER, review deep learning-based approaches, and offer suggestions for future research. Another review in MER (2021) [26] introduces a new categorization of methods based on temporal dynamics handling for speech, discusses feature representation methods for each modality, and presents challenges related to validation procedures, representation learning, and method robustness.

The previous surveys on MER have provided valuable insights and contributions to the field. However, we identified that the below needs and gaps remain unaddressed. This highlights the motivation for a new comprehensive survey. These needs and gaps include:

- Comprehensive Multimodal Focus that thoroughly addresses the complexities and advantages of MER.
- Detailed taxonomy Feature Extraction, Fusion Techniques, Evaluation Metrics, Datasets, and classifications of MER.
- An evaluation of the existing modalities in MER from a technical, scientific, and human-centric point of view.
- Human-centric approach with consideration of Ethical and Privacy Considerations.
- In detail Challenges, Open Issues, and future direction for each reviewed article.

B. THE MAIN CONTRIBUTIONS

Our review study employs a comprehensive search strategy, focusing primarily on Scopus for selecting relevant research studies.

- Literature Review Methodology: We follow PRISMA [27] guidelines for selecting pertinent research articles and use decision-making systems like topic modeling and clustering to enhance paper selection accuracy.
- Temporal Scope: We focus on articles published within the last decade to ensure a thorough review of recent literature.
- Abstract Text Analysis: We analyze abstracts of selected papers, categorizing of frequency repetition of words into Mono, Bi, and Trigram groups, and present the results accordingly.
- Focus on Multimodal: Our review includes physiological signals, textual, body gestures, and behavioral indicators (speech, facial expressions), exploring AI applications with both machine learning (ML) and deep learning (DL) techniques for emotion recognition.
- Emotion Model Diversity: We review articles covering various emotional models, including discrete and multidimensional frameworks, sourced exclusively from peer-reviewed journals.
- Categorization of Emotional Expressions: We categorize human emotional expressions, helping individuals discern sources of emotions and make informed selections.
- Comprehensive Analysis on Addressed Gaps and Possible Future Directions: We examine existing studies, focusing on fusion techniques and classification methods, addressing gaps, and proposing future directions for each reviewed paper.
- Building Blocks of MER: We present the essential steps and techniques of MER extracted from the reviewed literature.
- Current State Analysis and Modality Evaluation: We establish criteria for modality selection, considering device availability, expertise, societal factors, and human-centered elements in experiments.
- Challenges and Application Areas: We identify current obstacles, discuss challenges in-depth, and explore the broad spectrum of application domains for MER.

In this regard, the structure of this review is as follows:

Section II provides a comprehensive overview of key concepts, along with a discussion on the classification of emotional expression and emotion models essential with an overview of the available models and criteria for analyzing the current state of MER. In Section III, we outline the scope, objective, and main research questions based on the background review and gap analysis for the current review alongside the methodology and decision-making approach that has been employed in conducting this study. Section IV showcases the findings from the review process in graphical and tabular formats. Section V presents concluding remarks and a more in-depth analysis of the reviewed papers. Finally, Sections VI and VII wrap up the work with remarks on future direction, impediments, and limitations of the presented works.

II. BACKGROUND

This section outlines the characteristics of emotion, the existence of emotional models, and ways of expressing emotions. Also, it provides the basic study and seven selected criteria which later on are used to evaluate the modalities individually to be selected for MER. The outtake from this section, in summary, is in Fig. 1

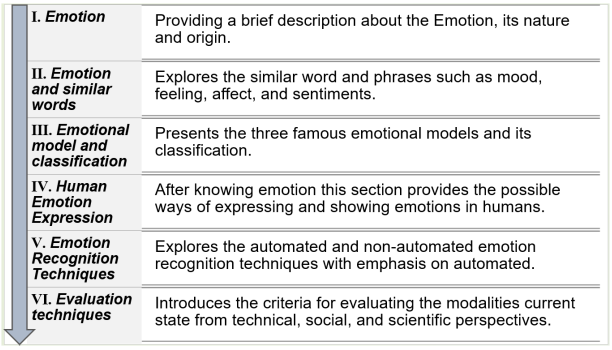

FIGURE 1. Concepts outtakes.

A. EMOTION

Emotions are often referred to as the "*motus anima*," or the driving force within us. Despite our efforts to be rational and logical, the diverse and inspiring nature of emotions remains profoundly impactful [3], [28].

Emotions are complex psychological and physiological states that are triggered by various stimuli or situations either internal or external. They involve a range of subjective experiences, such as feelings, moods, and affective states, that can influence a person's thoughts, behaviour, and physiological responses. Hence, despite being a commonly used term that appears to be easily understood, the definition of emotion has been a subject of debate for the past century, with an ever-expanding number of scientific definitions proposed [1].

The sheer volume of these definitions has grown to the point where attempting to count them seems pointless [29], [30]. For instance, in 1981, Kleinginna and Kleinginna [31] reviewed over a hundred definitions. Consequently, while it's impossible to compile a comprehensive definition, it is feasible to gain insight into the various perspectives on this topic by exploring a selection of the most influential definitions put forth by psychologists and others [31]. Therefore, we can gain insight into how psychologists and others have approached this subject by exploring some of the more impactful definitions and models.

Theories about emotions hold that, emotion is a complex entity with many components [4], [32]. Our emotions significantly influence how we think, behave, decide, and interact with others. They may inspire us to act, facilitate communication with others, and function as indicators of significant information in our surroundings. Emotions can also affect our cognitive functions, including memory, attention, and decision-making which refers to emotional decisions [33], [34]. Emotions play an important role in human social and cognitive functioning, as they help individuals navigate social interactions, make decisions, and regulate their behaviour and responses. They also play a critical role in mental health and well-being, as emotional disturbances or disorders can have significant negative impacts on a person's functioning and quality of life [35].

1) ORIGIN AND SIGN OF EMOTIONS

Over the years, numerous theories have emerged to explain the origins, mechanisms, and characteristics of emotions. While classic theories have faced criticism, many contemporary scholars still use them as foundational frameworks for their research [5].

The evolution of these theories began with the James-Lange Theory [36], [37], [38], which posited that emotions result from our physiological responses to external stimuli. This theory was later challenged by the Cannon-Bard Theory, which argued that emotions and physiological responses occur simultaneously but independently when the brain processes stimuli through the thalamus [38].

The James-Lange theory was further refined by the Schachter-Singer model [39], which combines physiological arousal and cognitive interpretation. According to this model, individuals experience physiological arousal in response to stimuli and then interpret this arousal based on environmental cues, assigning an emotional label accordingly [40].

More recent contributions include the Affective Events Theory, which explores the relationship between time and emotional reactions, and modern Cognitive Appraisal Theories [41], such as Richard Lazarus's Cognitive-Mediational Theory. These theories emphasize the role of cognitive appraisal in shaping emotional experiences.

The terminology surrounding emotions, including emotion, feeling, mood, affect, sentiment, and emotional dimensions, has roots dating back to ancient philosophical perspectives. While these terms are closely related, each carries distinct connotations.

B. EMOTION, FEELING, MOODS, AFFECT, SENTIMENTS, AND EMOTIONAL DIMENSION

The inception of these keywords can be traced back to the 18th century in scientific corpus and far beyond that in ancient philosophical perspectives like Socrates [42]. These terms are closely related but with distinct concepts and in-depth meanings.

Emotion refers to a complex, often involuntary psychological and physiological response to a particular stimulus. It involves bodily changes such as heart rate, facial expressions, and hormone release, and is typically considered universal and biologically rooted [43], [44].

Feeling is the conscious awareness or subjective experience of an emotion. It involves the mental interpretation and personal experience of the physiological and psychological changes that occur during an emotional response. While emotions are the broader processes including physiological responses, feelings are the conscious experiences and interpretations of those emotions [43], [44].

Moods are extended emotional states without a specific trigger. They are less intense than emotions but can last for hours, days, or longer, affecting a person's overall emotional state and how they perceive and react to events [45].

Affect is an unconscious, intense experience that is unshaped and unstructured, existing before or beyond

conscious awareness. It prepares the body to respond to situations with a measure of intensity, incorporating broader contexts beyond individual stimuli [44].

Sentiments are complex, long-lasting emotional attitudes or opinions toward people, objects, or concepts, influenced by emotions, feelings, and beliefs. They include enduring social connections that trigger emotions, emphasizing their social origins and influence [46], [47].

Emotional Dimensions include arousal and valence. *Arousal* refers to the level of physiological and psychological activation associated with an emotional state, ranging from calm to excited. *Valence* represents the positive or negative nature of emotion, ranging from pleasant to unpleasant [48], [49].

C. EMOTION MODELS

Different regions of the brain elicit various emotions. Emotional responses typically fall into three categories: reactive, hormonal, and automatic. In psychological terms, emotions are reactive phenomena triggered by stimuli and linked with qualitative physiological alterations. Researchers employ two primary methodologies to investigate the essence of emotions: the discrete method and the multidimensional approach.

In **Discrete emotions theory**, emotions are distinctly categorized, each characterized by its own set of cognitive, psychological, and behavioural elements. These emotions can be either positive or negative, with proponents of this idea suggesting the existence of a small number of fundamental emotions that are universally recognized across cultures. These basic emotions include happiness, sadness, anger, surprise, fear, and disgust. Robert Plutchik expanded on this concept with his comprehensive emotional model known as Plutchik's wheel of emotions, which comprises eight primary emotions: fear, joy, sadness, trust, anger, surprise, anticipation, and disgust. Additional emotions, which blend these primary ones, are determined by their position on the wheel, with intensity increasing towards the centre and decreasing towards the periphery. Fig. 2 illustrates an overview of Plutchik's wheel of emotions [50].

The **Multidimensional emotions theory** recognizes the complexity of emotional experiences, which are influenced by a multitude of factors including personal history, cultural context, and individual differences. This framework provides a deeper understanding of emotions, allowing for a more comprehensive analysis of emotional states. One classification within this approach is the 2-dimensional (2D) and 3-dimensional (3D) emotional space models. In the 2D model, emotions are categorized based on valence (positive or negative) and arousal (high or low activation). Russell's 2D emotional space model also known as the Circumplex Model of emotion illustrates this mapping [51].

The 3D model further incorporates dominance (feeling in control or controlled) alongside valence and arousal. Mehrabian and Russell's 3D emotional space model exemplifies this multidimensional mapping [52].

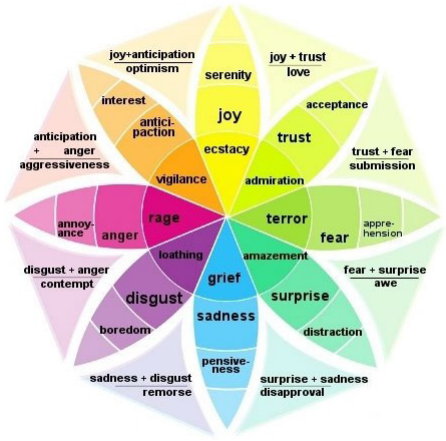

FIGURE 2. Plutchik's wheel of emotions [50].

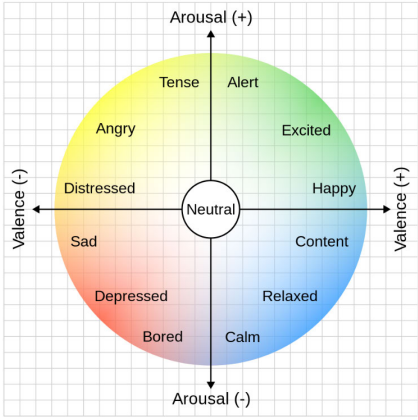

FIGURE 3. 2D VA emotion model/ Circumplex Model of emotion [51].

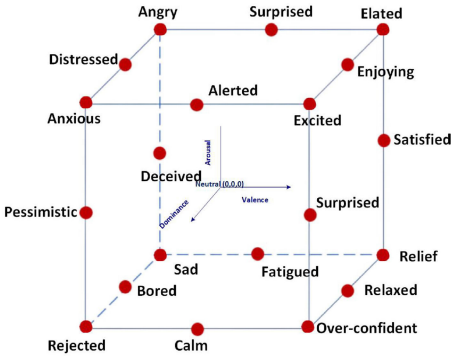

FIGURE 4. 3D VAD emotion model [52].

D. EMOTIONAL EXPRESSION

Humans express emotions through multiple channels such as facial expressions, body language, vocal intonation, gestures, and physiological responses. Emotions are complex and manifest internally and externally [53]. Darwin's two-stage model of emotional expression suggests emotions prepare us to react to our surroundings and convey social cues [54]. Keltner et al. [53] highlights the social aspect of emotions,

noting cross-cultural similarities in their production and understanding. Winkielman and Berridge [55] explores the unconscious nature of emotions, indicating they exist and can be expressed beyond our conscious control, and our brains can respond to past experiences without our awareness.

The theory of emotional regulation focuses on managing and controlling emotional responses across diverse circumstances [56], [57]. It includes strategies to comprehend, monitor, and alter emotional states [58], such as recognizing and interpreting emotions, understanding their triggers, and amplifying positive emotions or reducing negative ones [58]. Emotional regulation is often seen as the next phase of emotional expression, facilitating the transition of emotions to various manifestations and fostering artistic and aesthetic appreciation [59], [60], [61].

Charles Darwin's early work detailed over 40 emotional states, highlighting that emotional expressions manifest across multiple sensory modalities [62], [63]. Paul Ekman's [64] research demonstrated that basic emotions like happiness, sadness, anger, fear, surprise, and disgust have universal facial expressions, emphasizing their biological basis and cross-cultural recognition. This early understanding of the complexity of emotional expressions continues to influence contemporary affective science and ER research [65]. Albert Mehrabian's [52], [66] research on nonverbal communication, particularly his "7-38-55 rule," underscores the importance of tone of voice and body language in conveying emotions, though this rule is often misinterpreted and criticized.

The emotional expression goes beyond physiological gestures, finding outlets in various mediums. Langer's [67] work discusses how music, painting, and literature express human emotions. Indicators of emotion include heart rate, vocal tone, facial expressions, and sleep pattern changes [68]. Artistic expressions, linguistic forms, and communication tactics also serve as channels for emotion [69], [70], [71], [72]. Music, drawing, painting, and visual design offer unique platforms for emotive portrayal [73], [74], [75], [76]. These modalities sometimes intersect with physical expressions, such as in theatre or communication enhanced by laughter, but are often considered secondary or social forms of emotional expression. These expressions are often refined through emotional regulation [53].

Considering the two ways of emotional expression (Physiological and non-verbal expression vs communicative and artistic ways) Fig. 5 provides the two broad categories of emotional expression and expressing emotions within it.

In physiological responses [77], emotions can trigger physiological changes in the body, such as increased sweating, odour [78], trembling, changes in blood pressure, and changes in body temperature. Among them we can mention:

1) FACIAL EXPRESSIONS [79]

They are among the most recognizable signs of emotion. Different emotions, such as happiness, sadness, anger, fear, surprise, and disgust, are associated with distinct facial expressions.

2) BODY POSTURE [80]

Emotions can be conveyed through body language, including posture, gestures, and movements. For example, slumped shoulders may signal sadness, while clenched fists may indicate anger.

3) VOICE AND TONE [81]

Our pitch and rate of speech can reveal our emotional state. A high-pitched, rapid tone may indicate excitement or anxiety, while a slow, monotone voice may suggest sadness or boredom.

4) HEART RATE PATTERN [82]

The rhythmic variation in heart rate over time reflects the cardiovascular system's dynamic response to emotional experiences. For example, heart rate increases during arousal or excitement and decreases during relaxation.

5) COGNITIVE CHANGES [83]

Emotions influence our thoughts and cognitive processes. Anxiety may lead to racing thoughts and difficulty concentrating, while happiness may enhance creativity and problem-solving abilities.

6) BEHAVIORAL REACTIONS [68]

Emotions are often reflected in behaviors such as crying, laughter, aggression, or avoidance. However, these behaviors can be generalized; for instance, laughter can sometimes indicate anger or anxiety.

7) SLEEP PATTERNS [84]

Emotional states can impact sleep quality. Negative emotions like anxiety, depression, or stress are often associated with sleep disturbances such as insomnia or fragmented sleep.

8) EYE MOVEMENTS [85], [86]

Eye fixations, pupil size changes, and eye contact patterns can indicate emotional states. For example, increased pupil dilation is associated with heightened arousal.

9) BREATHING PATTERNS [87]

Different emotions can lead to changes in respiratory rate and pattern. Stress or excitement may result in faster, shallower breaths, while relaxation is associated with slower, deeper breaths.

Besides the physiological expression of emotions, another significant way emotions can be expressed is through social interactions, particularly at higher, more creative, and artistic levels. Artistic expression encompasses various visual mediums, including painting, drawing, sculpture, photography, and digital art [64], [69], [70]. Artists use color, composition, form, texture, and symbolism to convey

emotions and evoke responses in viewers. Artistic expression allows individuals to communicate complex emotions that may be difficult to convey through words alone, tapping into the universal language of imagery and visual aesthetics. Among them:

10) WRITTEN LANGUAGE AND LITERATURE [71]

Writers use prose, poetry, essays, and letters to articulate emotions, thoughts, and experiences creatively. This medium allows for the exploration and communication of complex emotions.

11) MUSICAL EXPRESSION [73]

Music conveys emotions through melody, harmony, rhythm, and tempo. Different instruments and vocal techniques can evoke various emotional responses.

12) PAINTING AND DRAWING [74], [75]

These visual arts provide powerful means for expressing emotions. Artists use elements like color, line, and form to communicate a range of feelings and inner experiences.

13) VISUAL DESIGN [76]

Designers use color theory, composition, and visual elements such as shape and texture to convey emotions. Visual design can create aesthetic experiences and emotional connections.

14) VERBAL EXPRESSION [68]

Language, both spoken and written, is used to communicate emotions. Verbal expression can facilitate social connection, empathy, and emotional support.

E. EMOTION RECOGNITION

Recognizing emotions involves the discernment and intelligent interpretation of human emotional states through various cues, including speech, facial expressions, body language, and physiological signals. This capability can be achieved either by humans or through automated systems.

1) NONE-AUTOMATED EMOTION RECOGNITION

Humans are naturally adept at recognizing emotions, often relying on a combination of cues such as facial expressions, tone of voice, body language, and context. This intuitive process involves complex cognitive and perceptual mechanisms, enabling individuals to accurately infer others' emotional states. While human emotion detection is typically spontaneous and nuanced, researchers have also developed structured techniques for assessing emotions, such as the Facial Action Coding System (FACS) for analyzing facial expressions [88], the Geneva Emotion Wheel for categorising emotional states [89], [90], and the Self-Assessment Manikin (SAM) for measuring subjective emotional experiences [91].

Moreover, several widely used questionnaires and scales have been developed to assess emotional experiences and states systematically. For instance, the Positive and Negative Affect Schedule (PANAS) measures positive and negative affectivity through self-reported ratings of various emotional states [92]. The Beck Depression Inventory (BDI) evaluates the severity of depressive symptoms, while the State-Trait Anxiety Inventory (STAI) assesses both temporary (state) and enduring (trait) anxiety levels [93], [94]. Additionally, the Profile of Mood States (POMS) measures transient mood states across various affective dimensions [95].

2) AUTOMATED EMOTION RECOGNITION

Automated ER is the process of identifying and interpreting the emotional state of an individual based on various cues by machines. It spans across human-computer interaction, psychology, education, health, and entertainment [21]. Due to the intricate and varied nature of human emotions, this area poses challenges and necessitates employing a range of techniques, including multimodal data analysis, machine learning, computer vision, and natural language processing. The objective of emotion detection is to develop and implement intelligent systems within various applications and devices.

Unimodal, early efforts in human ER primarily focused on single modes of expression an detection. Within this context, facial recognition followed by speech and text emotion recognition stood out as the most extensively studied modality, largely because of the accessibility of datasets and the advancements in computer techniques [96], [97].

Bimodal systems typically integrate information from two different sensory modalities, such as vision and audition (A-V). A bimodal system might combine visual cues from facial expressions with auditory cues from vocal intonation to infer individuals' emotional states. Bimodal systems provide a relatively simple approach to integrating two sources of information and are often used when the combination of two modalities is sufficient for the task at hand [98].

Multimodal means utilizing various information sources like speech, facial expressions, body language, and physiology to detect human emotions [17]. It is a challenging task due to the complexity and context-dependency of human emotions and the way of expressing emotions and detecting them. To address this challenge, machine learning, computer vision, and natural language processing (NLP) techniques are often employed to integrate and analyze these different modalities, however, these techniques might increase the complication even more. If multimodal detection can overcome its complexity, it has the potential to provide the most accurate means of detecting emotions, as humans typically express emotions through multiple modalities [66].

The choice of ER technique—whether unimodal, bimodal, or multimodal—depends on various factors such as the context of the application, the complexity of the emotional states to be detected, and the available resources. Here's a discussion on each technique:

In addition to choosing the appropriate emotion recognition technique—whether it be unimodal, bimodal, or multimodal—the selection of the modality itself poses

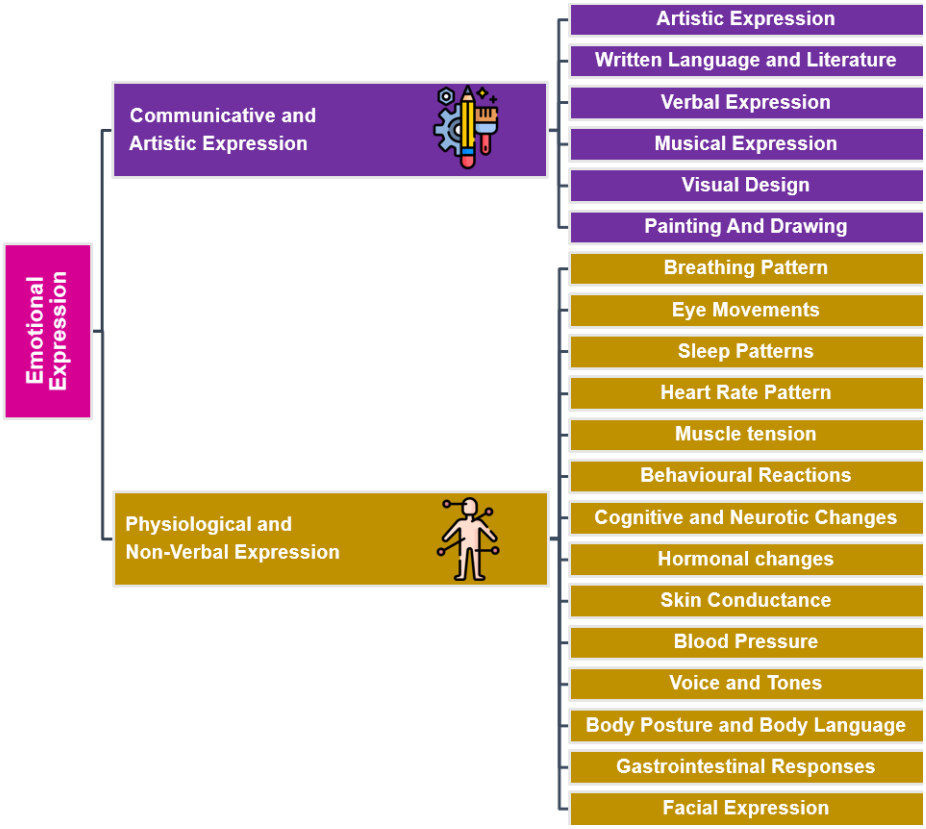

FIGURE 5. Human emotional expression.

TABLE 1. Comparison of emotion recognition techniques.

Approach	Advantages	Limitations
Unimodal	They are often simpler to implement and require less computational resources compared to bimodal or multimodal approaches. Unimodal techniques are sufficient for tasks where emotional cues are predominantly expressed through one modality.	Unimodal techniques may struggle to capture the full complexity of human emotions, as emotional cues can be expressed through multiple modalities simultaneously. They may also be less robust in diverse contexts where emotional expressions vary across individuals or cultural backgrounds.
Bimodal	The approaches capture a broader range of emotional signals and reduce ambiguity inherent in unimodal systems. Bimodal techniques offer a balance between simplicity and complexity, making them suitable for applications where two emotional cues are present but additional modalities may not significantly enhance performance.	Bimodal still overlook important emotional cues conveyed through other modalities beyond the two integrated. They also require more sophisticated algorithms and computational resources compared to unimodal approaches.
Multimodal	This approach can capture the complexity and subtlety of human emotions across diverse contexts and individuals. They offer the highest potential for accuracy and robustness, particularly in applications where emotional cues are subtle or context-dependent and are expressed through multiple channels.	The technologies are the most complex and resource-intensive, requiring complicated algorithms, extensive data collection, and integration across multiple modalities. Implementing and maintaining multimodal systems may be challenging, particularly in real-world settings with limited resources or technical expertise.

a complex challenge, influenced by a myriad of factors and criteria. These factors encompass the nature of the task at hand, the characteristics of the target audience, the available resources and technology, and the specific context in which the emotion detection system will operate. Moreover, considerations such as the reliability, accuracy, and interpretability of the chosen modality, as well as its compatibility with existing infrastructure and systems, must be taken into account. Additionally, the ease of data collection, processing requirements, computational efficiency, and ethical considerations further contribute to the complexity of modality selection. Therefore, a comprehensive evaluation and thorough analysis of these factors are essential to make informed decisions regarding the most suitable modality for a given ER task.

The next subsection aims to establish the path for suitable criteria that can navigate the researchers for evaluation and comparison of each modality for selection.

F. ESTABLISHING GUIDELINES FOR ANALYZING THE CURRENT STATE OF AUTOMATED EMOTION RECOGNITION

When considering the intersection of AI and human interaction, one of the most challenging aspects is seamlessly integrating users into the process. This involves not only enabling users to interact with AI systems but also fostering engagement, encouraging the sharing of information or feedback, and promoting acceptance of AI recommendations or decisions. Alongside these user-centric considerations, there are also technological factors to address, such as the maturity level of the AI technology being deployed and the ease of use of the interface or platform. Achieving a harmonious balance between user engagement and technological complexity is essential for the successful implementation and adoption of Human AI systems [99].

Several models have been proposed to address the challenges of user acceptance and engagement in technological and AI systems. One notable model is the Technology Acceptance Model (TAM) [100], which focuses on users' perceived ease of use and perceived usefulness of technology as key determinants of its adoption. Another model is the UTAUT [101], which extends TAM by incorporating additional factors such as social influence, facilitating conditions, and behavioural intentions. Additionally, the Persuasive Systems Design (PSD) model [102] emphasizes the importance of persuasive features in technology design to influence users' attitudes and behaviours.

Ethical considerations are also an essential factor in user acceptance models, particularly in the context of AI systems. While they may not be explicitly listed as criteria in traditional models like TAM, UTAUT, or PSD, they are increasingly recognized as crucial aspects of technology adoption and user engagement [103]. Furthermore, real-time performance emerges as a pivotal criterion within user acceptance models, especially in contexts where users depend on AI systems for immediate or time-sensitive tasks. Although not explicitly addressed in traditional models like TAM or UTAUT, real-time performance often intersects with constructs such as perceived usefulness or perceived ease of use, thereby shaping users' perceptions and experiences [104].

Below, we provide a list of selected criteria along with detailed explanations for each, based on the mentioned models and reviewed articles which later on we will use in order to empower us for current state evaluation which is one of the goals of this review. In this regard, we aimed to incorporate three foundational pillars: *social*, *technical*, and *scientific* approaches, recognizing their pivotal role within the emotional recognition context (See Fig. 6).

Accessibility [100] is not explicitly included as criteria in the Models. However, it can indirectly influence users' perceptions and intentions to use a system. In the TAM and UTAUT models, accessibility can be indirectly addressed under the construct of ''Facilitating Conditions''.

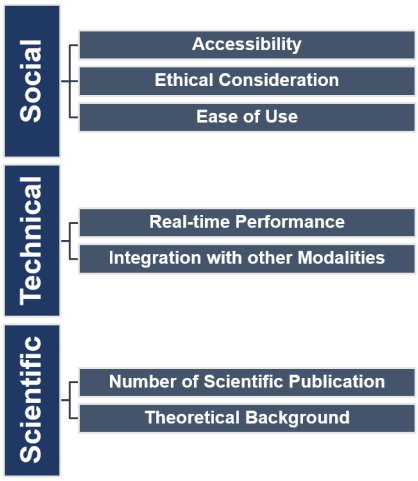

FIGURE 6. Emotional expression.

Ethical Consideration [103] encompass various aspects, including data privacy, fairness, transparency, and accountability. Users are more likely to accept and engage with AI systems that demonstrate ethical behavior, respect user privacy, and provide transparent explanations of their decision-making processes. Additionally, emotion detection data is often linked to mental data, which is regarded as sensitive information. Thus, not only the modalities from which the data are extracted and the datasets used to train the models but also the results of the analysis, are all considered highly sensitive data [105], [106].

Ease of Use [100], the degree to which users perceive a system as free from effort or difficulty to use. It is considered an important criterion in TAM model.

Real-time Performance [100], users generally prefer AI systems that provide real-time responses, quick processing times, and seamless interactions. Slow response times or delays in system performance can negatively impact user satisfaction and adoption. Therefore, in the evaluation of AI systems, real-time performance is an essential consideration to ensure that users perceive the system as efficient and effective in meeting their needs.

Integration with other Modalities, the selected modality should be feasible to integrate into the multimodal system without significant technical challenges or resource constraints.

Number of Scientific Publication [107] can be considered a criterion for evaluating the strength and background of scientific research. While a high number of publications may indicate productivity and activity in a particular research area, the quality, impact, and relevance of the publications are also crucial factors to consider. However, this review assesses productivity and the attention garnered by scientists in the field based on the number of publications.

Theoretical Background [108], as we discussed earlier emotions are conveyed through diverse channels of human communication. Hence, in the context of emotion detection,

it's important to remember that relying on just one or two ways to detect emotions might not give us all the information we need, even if we achieve high accuracy rates. For instance, when considering the 7% contribution of words to emotional expression (based on the Mehrabian rule [66]), even if we're very accurate in detecting those words, we're only capturing a small part of all the emotions being expressed.

While numerous criteria exist across the social, technical, and scientific domains, it is not possible to consider them all in one work. The selection of these specific criteria is informed by expert opinion garnered from researchers actively engaged in this field of study.

III. SCOPE AND METHODOLOGY

A. PURPOSE AND SCOPE OF THE REVIEW

Understanding and appropriately responding to human expressions is increasingly vital in human-computer interaction, computational creativity, affective computing, and cognitive computing, driven by the rapid expansion of technology and the internet. The recognition of emotions is becoming prevalent in both scientific research and industry, although it remains challenging for computers due to the intricate nature of emotions.

There have been reviews conducted in the field of emotion detection, including MER. However, to our knowledge, previous studies lack a distinct perspective, often emphasizing the human aspect of emotional expression without providing a comprehensive theoretical background or research overview. Furthermore, the absence of concrete criteria for modality selection is evident, with modalities often chosen based on preferences or availability in the laboratory. Notably, reviews often overlook the overarching framework or building blocks of MER. Additionally, there is a lack of in-depth analysis regarding identified gaps and potential future research directions across the reviewed papers.

By addressing the available gaps in the reviewed articles, this article aims to offer a unique perspective by focusing more on the human aspect of emotional expression and emotions alongside the technical sides. It investigates philosophical discussions and the origins of emotions in humans, exploring how individuals express their emotions. Additionally, the article aims to establish primary criteria for evaluating different modalities, considering human acceptance as a crucial factor in selecting specific modalities for MER. To ensure clarity of objectives and address existing gaps in the review, the following questions have been formulated:

Q1: What are the main building blocks of multimodal emotion recognition?

Q2: What is the current state of multimodal emotion recognition, and how does each modality contribute to this field, from technical, scientific, and social perspectives?

Q3: What are the possible applications for multimodal emotion recognition?

Q4: What are the current challenges for multimodal emotion recognition?

These questions will direct the subsequent development of this review.

B. METHODOLOGY

The literature review was conducted systematically, according to the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)* guidelines [27]. We selected this methodology for its predefined and well-established procedures, ensuring a transparent and comprehensive reporting of our research findings. However, when considering a novel approach a decision-making approach as assistance has also been used and proposed alongside human intelligence for more accurate selection and data extraction alongside the review selection process. Fig. 7 provides an overview of the entire process, with further detail provided in subsequent sections.

1) IDENTIFICATION, SCREENING, AND DETERMINING ELIGIBILITY IN THE STUDY

In essence, a systematic literature review has been undertaken for the study. While the primary database used for this research is Scopus, there were instances where it became necessary to search in additional databases, such as Google Scholar and Web of Science, to retrieve supplementary scientific papers, specifically to cover the literature review in the section II.

A set of keywords has been chosen considering relevant terminology in the area. Core concepts reflected here are ''multimodal'', ''emotion'' and ''recognition''. Those are accompanied by complimentary terminology i.e., ''feeling'', and ''detection''. Table 2 displays the query strings utilized in the initial paper selection stage. Each row in the table is connected by the logical operator ''AND''.

TABLE 2. Query string to be adapted for the repository.

Query String
(multimodal OR "multi modal" OR Multi-modal)
AND
(detection OR recognition)
AND
(emotion OR feeling)

Table 3 displays another level of the applied criteria, which were utilized to narrow down the selection of papers.

Despite applying various filters, some irrelevant papers remained. This cleaning phase involved human screening to identify all relevant papers. Below are the inclusion criteria used to narrow down the selection. The inclusion criteria to follow and make it narrow down while screening the articles are shown in Table 4.

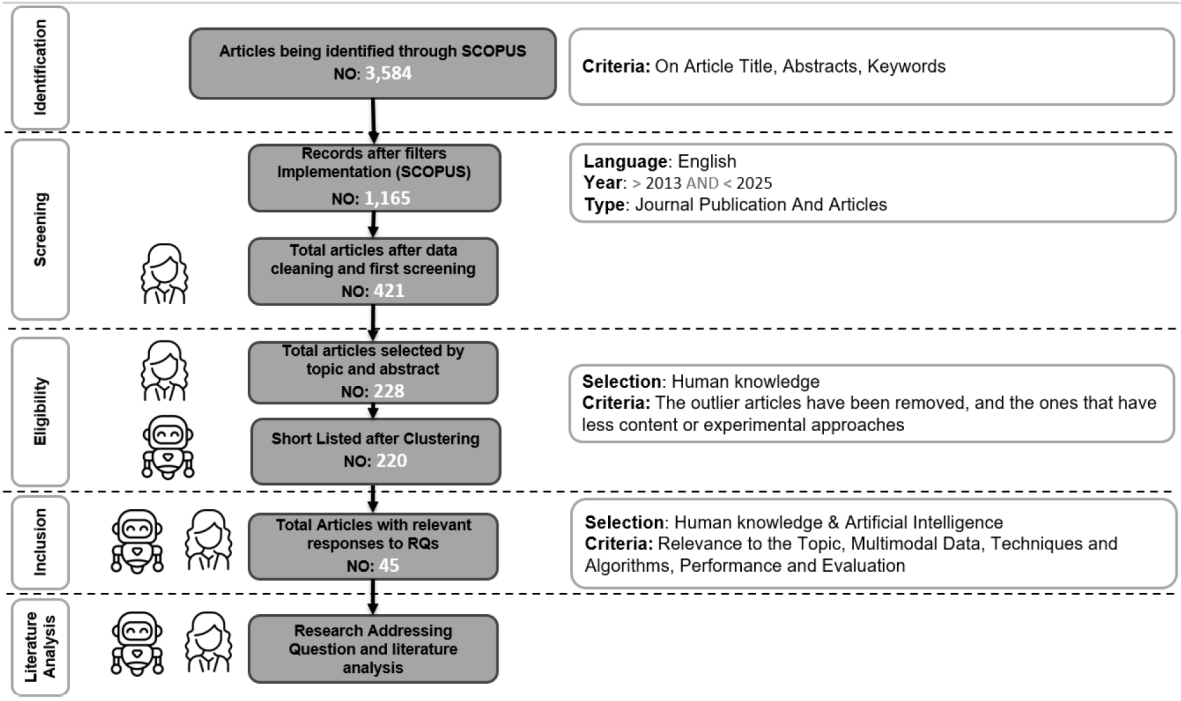

FIGURE 7. Methodology for the literature review.

TABLE 3. Applied selection criteria.

Category Selection Criteria
("Language is English")
AND
(Years are between 2014 and 2024)
AND
(Type should be Journal Publication or Articles)

TABLE 4. Inclusion criteria.

Inclusion Criteria
The entire document is downloadable.
The document is written in English.
The selected papers are peer-reviewed works.
The document is pertinent to the research inquiries.
The focus of the document is primarily on methods for MER.
The document is well-written and contains all the specified details (Datasets, modalities, feature extraction methods and techniques, classification technique, evaluation technique, and evaluation results).

2) ANALYSIS BASELINE

To analyze the research data and the papers extracted, various analytical tools such as Power BI, Jupyther, python libraries, and finally a combination of automated NLP and ML techniques and human screening were employed to analyze a large corpus of digital research publication abstracts, conclusions, and whole text to assess the current research interest and emerging trends around Identifying Emotion Using AI-Based Emotional Recognition.

In this literature review for a more accurate selection of the articles and to increase the efficiency of the decision-making, a combination of human and machine learning collaboration is being applied.

Generally, to conduct the literature review in each area the following steps have been taken (Fig. 9):

Step 1: A Suitable combination of keywords has been selected and searched in Scopus.

Step 2: Necessary criteria and filter being applied to narrow down and also reduce the number of papers. Table 2 and 3 show the selection criteria alongside the inclusion ones.

Step 3: Then the Titles, abstracts, and keywords are exported and extracted, and stored in CSV format for the human screening selection phase.

Step 4: To verify human selections and refine categorization, a K-means clustering model was utilized. This helped identify potential outliers and offered deeper insights into the categorized articles.

Step 5: In this step, abstracts are cleaned and prepared for further evaluation and analysis. To do this, the text is converted to lowercase, punctuation is removed, and common English stop-words and specified conjunctions such as "and," "or," "a," "the," "in," "on," etc., are eliminated. The final results are stored in separate columns for subsequent steps.

Step 6: For the final selection and analysis, both human judgment and machine assistance are employed. We use topic modelling, a statistical technique for identifying themes

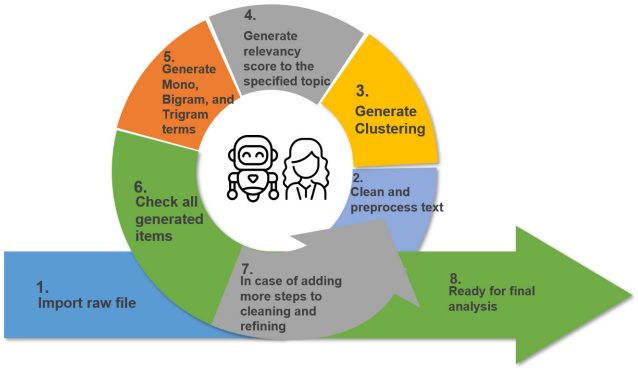

FIGURE 8. Support system decision making process.

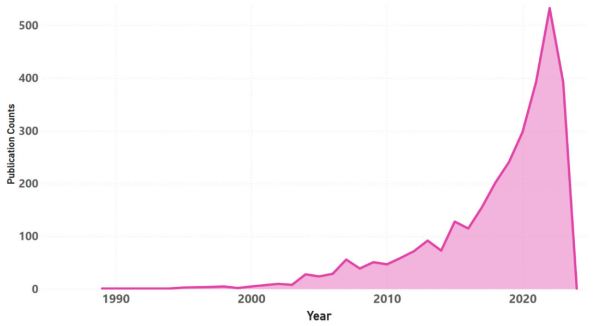

FIGURE 9. Present a histogram with the distribution of published articles in the field of Multimodal emotion recognition.

within large text collections, to enhance decision-making accuracy. This approach is essential for managing extensive datasets where manual analysis would be impractical. Relevant articles are selected based on their content, relevance to "*Multimodal Emotion Detection using AI-Based Emotional Recognition*," and scores from topic modelling, ensuring they address key questions (see Fig. 8). This iterative phase aims to achieve the best selection. The detailed steps in this process include:

1) *Relevance to the Topic:* Searched for articles directly relevant to emotional recognition and multimodal data fusion, particularly AI-based recognition. Emphasized articles discussing emotional recognition across diverse contexts with potential applications. Topic modelling and clustering facilitated this phase by applying relevance scores to the data frame.
2) *Multimodal Data:* Focused on articles discussing the integration of multiple data modalities (audio, text, visual) for emotion recognition, highlighting the need for various information sources in the process.
3) *Techniques and Algorithms:* Prioritized articles presenting specific techniques, algorithms, or models for emotion recognition.
4) *Performance and Evaluation:* Considered articles that included performance evaluations and showcased results, allowing for an assessment of the practical applicability of the methods discussed.

Step 7: Finally, data analysis was approached from various angles, employing diverse visualization techniques to present the findings. These included comparative tables, temporal charts, bar charts, and radar charts, each providing unique perspectives on trends and comparisons.

The following section is dedicated to presenting the results and conducted corpus analysis.

IV. RESULTS: TECHNICAL ASPECTS, GAPS AND POTENTIAL FUTURE RESEARCH

Fig. 9 presents a histogram with the distribution of articles within years in the multimodal emotion recognition field. It reveals a notable trend in the growth of research activity in the field of ER over the years. The number of publications

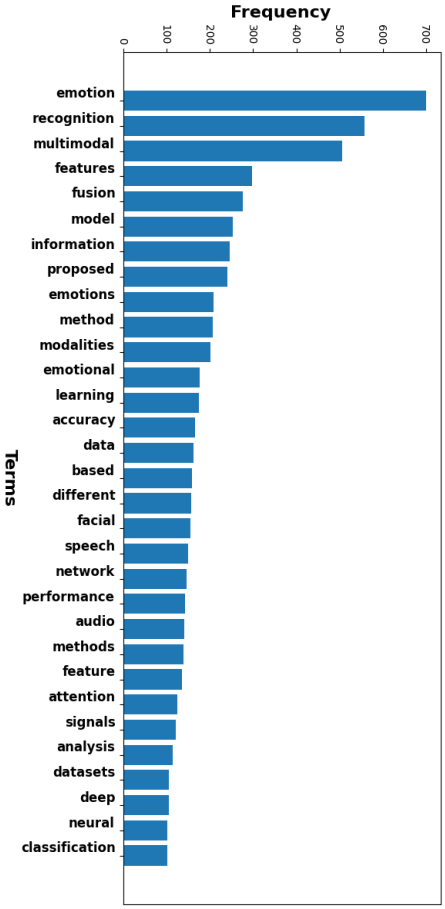

FIGURE 10. Top 30 Most Frequent unigram terms in abstracts in the field of multimodal emotion recognition.

exhibits a gradual increase from the late 1990s, with a significant surge observed from the mid-2000s onwards. Particularly noteworthy is the exponential growth observed in recent years, with a substantial rise in publications from 2010 onwards. This growth trajectory underscores the increasing interest and importance of ER research across various disciplines. The peak in publications is observed in 2023, followed by a slight decline in 2024 which is accepted

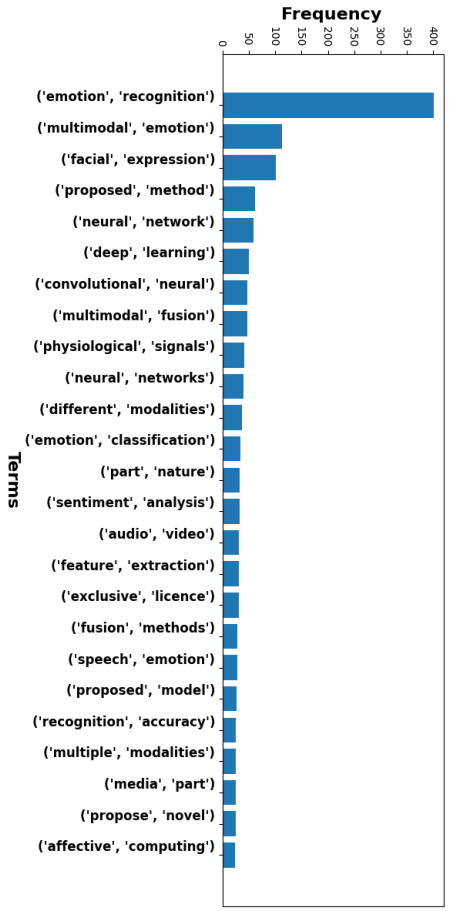

FIGURE 11. Top 25 Most Frequent bigrams in abstracts in the field of multimodal emotion recognition.

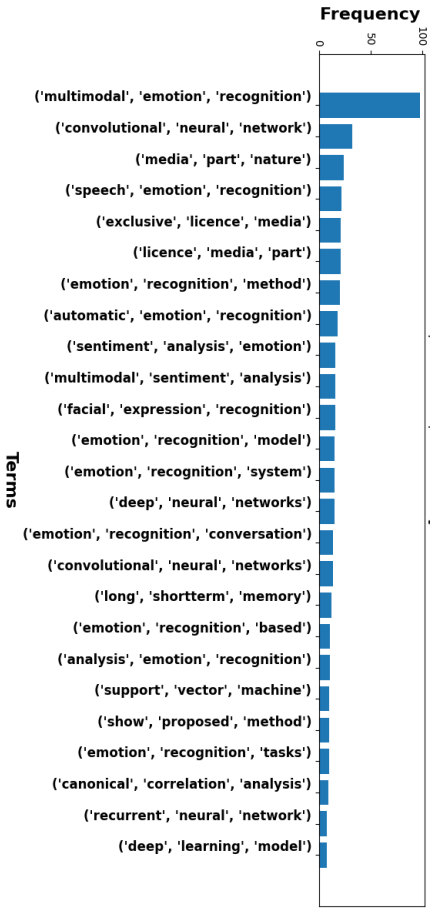

FIGURE 12. Top 25 Most frequent terms and trigrams in abstracts in the field of multimodal emotion recognition.

as we are conducting this review almost at the halfway of the year.

Figures 10, 11, and 12 represent the frequency and co-occurrence of terms extracted from the abstracts of the corpus related to ER research. The **unigram** analysis (Figure 10) highlights the emphasis on integrating multiple modalities for ER, with terms like '*emotion*', '*recognition*', '*multimodal*', '*features*', and '*fusion*' being most prevalent. Additionally, terms such as '*model*', '*information*', '*emotions*', and '*method*' signify efforts to develop robust approaches for emotion analysis, utilizing diverse modalities such as '*facial*', '*speech*', '*audio*', '*video*', '*text*', and '*visual*', often employing advanced techniques like '*deep learning*' and '*classification*'.

In the **bigram** analysis (Fig. 11), notable pairs like (*'emotion', 'recognition'*) and (*'multimodal', 'emotion'*) underscore the focus on recognizing emotions through multimodal approaches, particularly leveraging techniques like '*neural networks*' and '*facial expressions*'. Terms like (*'deep learning'*) and (*'convolutional neural'*) highlight the utilization of advanced techniques, while pairs such as (*'physiological signals'*) and (*'speech emotion'*) indicate exploration into physiological and speech signals for ER, intersecting with fields like *'affective computing'*.

The **trigram** analysis (Fig. 12) further emphasizes the interest in MER, particularly through CNN. The presence of trigrams like (*'speech emotion recognition'*) and (*'facial expression recognition'*) underscores the focus on recognizing emotions from speech and facial expressions, while terms such as (*'sentiment analysis emotion'*) suggest intersections between sentiment analysis and ER methodologies. These analyses collectively reflect the diverse research landscape and ongoing efforts to advance MER methodologies and techniques.

Finally Fig. 13 the analysis of publications group by country, indicates a notable dominance of China in the field, with 186 publications, followed by the United States with 50 publications and India with 49. These three countries exhibit a significant presence in ER research, reflecting their substantial contributions to the field. Other notable contributors include the United Kingdom, Japan, and South Korea, each with around 20 publications. Interestingly, while several European countries such as Germany, Spain, and the Netherlands show moderate participation, there is

also representation from diverse regions like Saudi Arabia, Pakistan, and Singapore.

The accuracy and effectiveness of the classification process rely heavily on the quality of the training data. Various datasets are available, tailored to different input modalities and classification criteria, which can be utilized for training the models. The overall summary of the reviewed datasets are shown in Table 5.

The review of the final 45 selected articles is presented in Table 6 provides more details from the papers such as used algorithms, fusion techniques, accuracy techniques, and used modalities. Table 7 presents a more analytical approach toward the targeted research gap and implications for future research for each reviewed paper. Each method employs unique fusion techniques, feature extraction methods, classification algorithms, and evaluation metrics to achieve varying levels of accuracy.

Dataset and Emotions:

The datasets used across the studies range from publicly available datasets such as AffectNet [131], IEMOCAP [109], and MELD [120], to proprietary or specifically generated datasets. The emotions targeted also vary significantly, with some studies focusing on basic emotions like happy, sad, angry, and neutral, while others include a broader spectrum such as arousal, valence, and even sentiment analysis.

Modalities and Fusion Techniques:

Modalities explored in these studies encompass audio-visual (AV) signals, physiological signals (e.g., EEG, ECG, GSR), and text. Fusion techniques include modality-adaptive fusion, ensemble models, cross-attention mechanisms, and transformer-based methods. Kang et al. [139] utilize modality-adaptive fusion for AV and physiological signals, while Miah et al. [140] use an ensemble model of transformers and a large language model for text-based sentiment analysis.

Feature Extraction and Classification:

Feature extraction methods are varied, from adversarial learning and plain regression tasks to more complex methods like PCA, Q-learning techniques, and multi-scale MFCCs. Classification approaches similarly range from traditional methods like SVMs and decision trees to deep learning-based methods and hybrid deep learning models. Selvi and Vijayakumaran [141] employ Q-learning techniques with target Q-networks for classification, achieving high accuracy.

Evaluation Metrics and Achieved Accuracy:

Evaluation metrics used include accuracy (ACC), F1 score, precision (P), recall (R), and more specific measures like weighted accuracy (WA) and unweighted average recall (UAR).

Besides technical approaches, the reviewed methods in Table 7, while providing a broad range of potential future research directions in MER, have revealed several limitations that need to be addressed.

One major limitation highlighted in several papers, such as [139] and [142], is the challenge of handling out-of-distribution inputs and data imbalance. While these studies propose enhancing models through domain adaptation techniques and advanced learning algorithms, they often do not consider the substantial computational resources required for these approaches. For instance, implementing sophisticated algorithms to balance data can be computationally intensive and may not be practical for real-time applications or in environments with limited resources.

Another limitation is the integration of physiological signals into MER systems, as suggested in [140] and [141]. Although incorporating signals like EEG and heart rate can potentially improve the accuracy of ER, these methods face practical challenges. The need for specialized sensors and the complexity of synchronizing physiological data with other modalities can hinder their widespread adoption. Additionally, such approaches might not be user-friendly or scalable in everyday applications, where non-intrusive methods are preferred.

When comparing the proposed multimodal fusion techniques, such as those in [143] and [144], it becomes evident that while multimodal approaches can enhance ER accuracy, they also introduce complexity in data processing and model training. Multimodal fusion requires careful alignment and synchronization of different data streams, which can be technically demanding and time-consuming. Moreover, the computational load increases with the addition of multiple data sources, potentially limiting the applicability of these methods in real-time systems.

The proposal in [142] to develop comprehensive multimodal datasets also underscores a significant gap in the current research. While larger and more diverse datasets can improve model training and validation, creating such datasets is resource-intensive. Collecting and annotating data across various modalities and emotional states requires substantial effort and coordination. Additionally, there is a need for standardized protocols to ensure the consistency and reliability of these datasets.

In terms of practical applications [134], the suggestions for improving real-time ER systems often overlook the trade-offs between accuracy and processing speed. Techniques that offer higher accuracy may not meet the latency requirements for real-time applications, posing a challenge for their implementation in scenarios like interactive user interfaces or adaptive learning environments.

V. DISCUSSION

Reviewed research studies have explored different modalities and techniques to enhance the accuracy and robustness of ER systems. This discussion compares and contrasts the achievements and methodologies of several notable studies, grouped by the modalities they have utilized.

Text and Speech:

Incorporating text data has also been a significant focus. Le et al. [153] applied a transformer-based fusion method on video frames, audio signals, and text subtitles from IEMOCAP and CMU-MOSEI, achieving weighted accuracies of 78.98% and 79.63%. Miah et al. [140] focused on

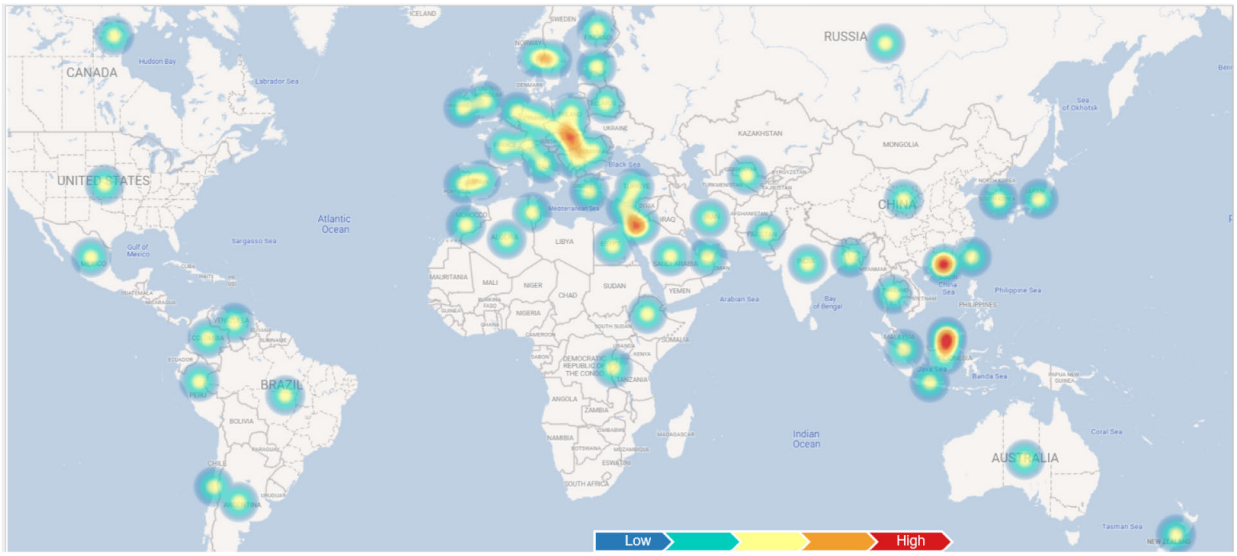

FIGURE 13. Heat map of publication counts group by country in the field of multimodal emotion recognition.

TABLE 5. Collection of datasets.(Audio-Visual (A-V)).

Ref	Datasets	Emotions	Type of data
[109]	*IEMOCAP*	Happy, Sad, Angry, Disgusted, Surprised, Fear, Excited, Frustration, Neutral	A-V
[110]	*CMU-MOSI*	Happy, Sad, Angry, Disgusted, Surprised, Fear, Excited, Neutral	A-V
[111]	*CMU-MOSEI*	Happy, Sad, Angry, Disgusted, Surprised, Fear, Excited, Neutral	A-V
[112]	*MODMA*	Happy, Sad, Angry, Disgusted, Surprised, Fear, Excited, Neutral	A-V
[14], [113]	*DAIC-WOZ*	Happy, Sad, Angry, Disgusted, Surprised, Fear, Excited, Neutral	A-V
[114]	*Extended DAIC*	Happy, Sad, Angry, Disgusted, Surprised, Fear, Excited, Neutral	A-V
[115]	*D-Vlog*	Positive, Negative	Video
[116]	*K-EmoCon*	Joy, Sad, Anger, Surprise, Fear, Disgust, Neutral	Text
[117], [118]	*SEED*	Happy, Sad, Angry, Disgusted, Surprised, Fear, Neutral	Physiological Signals
[119]	*SEED-IV*	Happy, Sad, Angry, Disgusted, Surprised, Fear, Neutral	A-V
[120]	*MELD*	Happy, Sad, Angry, Disgusted, Surprised, Fear, Excited, Neutral	A-V
[121]	*RAVDESS*	Happy, Sad, Angry, Neutral, Fear, Disgust, Surprise	A-V
[122]	*MEISD*	Joy, Anger, Sad, Fear, Disgust, Surprise	A-V
[123]	*AFEW*	Happy, Sad, Surprise, Fear, Disgust, Anger, Neutral	A-V
[124]	*SFEW*	Happy, Sad, Surprise, Fear, Disgust, Anger, Neutral	Static Images
[125]	*AffWild2*	Happy, Sad, Surprise, Fear, Disgust, Anger, Neutral	A-V
[126]	*RECOLA*	V, A	A-V
[127]	*MUStARD*	Happy, Sad, Surprise, Fear, Disgust, Anger, Neutral	A-V
[128]	*SAVEE*	Happy, Sad, Surprise, Fear, Disgust, Neutral	Audio
[129]	*TESS*	Happy, Sad, Angry, Disgust, Fear, Neutral	Audio
[130]	*CREMA-d*	Happy, Sad, Angry, Disgust, Fear, Surprise, Neutral	A-V
[131]	*Affectnet*	V, A, Happy, Sad, Surprise, Fear, Disgust, Anger, Neutral, Contempt, Uncertain	Video
[132]	*ASCERTAIN*	Happy, Sad, Surprise, Fear, Disgust, Anger, Neutral, Confusion, Frustration, Excitement, Boredom, Engagement, Pain, Interest, Concentration	Physiological Signals, Video
[133]	*M-LFW-FER*	Happy, Sad, Surprise, Fear, Disgust, Anger, Neutral	Video
[134]	*LUMED-2*	Sad, Neutral, Happy	Physiological Signals and Video
[135]	*DEAP*	Angry, Disgust, Afraid, Happy, Neutral, Sad, Surprised	Physiological Signals
[136]	*AMIGOS*	V, A, Personality traits, Fear, Anger, Disgust, Sad, Happy, Surprise	Physiological Signals, Video
[137]	*AVEC*	V, A, Happy, Sad, Surprise, Fear, Disgust, Anger, Neutral	A-V
[138]	*HAPPEI*	Positive emotions (Happy)	Video

sentiment analysis using text data in multiple languages. They employed an ensemble model of transformers and a large language model (LLM), achieving over 86% accuracy in sentiment analysis. Zhang et al. [145] developed an adaptive attention mechanism for speech and text modalities, extracting unimodal features and achieving notable

TABLE 6. Summary of technical approaches of the selected and reviewed papers.

Ref	Datasets	Emotions	Modalities	Fusion Technique	Feature Extraction	Classification	Evaluation Metric	Achieved Accuracy
[139] 2024	AffectNet	Happy, Sad, Angry, Neutral, Arousal, Valence	A-V, physiological signals	Modality-adaptive fusion	Adversarial Learning, Plain Regression Task, Extraction from Speech Signal	Deep learning-based	Pseudo labels	Acc of up to 33% than emotion recognition using only external signals
[140] 2024	The datasets generated	Sentiment analysis (positive, negative, or neutral sentiments)	Text (in various languages: Arabic, Chinese, French, Italian)	Ensemble model of transformers and a large language model (LLM)	Translation using neural machine translation models (LibreTranslate and Google Translate)	Ensemble of pre-trained sentiment analysis models (Twitter-Roberta-Base-Sentiment-Latest, bert-base-multilingual-uncased-sentiment, GPT-3)	Acc	Over 86%
[141] 2023	ASCERTAIN	Sad, Angry, Happy, Surprise, Disgust, Hatred	ECG, EEG, GSR	FOX-optimized (DDQ)		Q-learning technique, Target Q-network	F1 , Acc	greater than 95%
[143] 2023	RAVDESS, CREMA-D	Neural, Anger, Disgust, Fear, Happy, Sad, Surprise, Calm	A-V	A-V cross-attention	Spatial channel, Temporal attention mechanisms	Novel classification loss based on an emotional metric constraint	Average Acc	Average Acc scores 89.25% RAVDESS and 84.57% CREMA-D
[144] 2023	The 240 sets of data obtained from the emotion-inducing experiment	2D emotional model	Physiological signals	Multimodal information fusion		Binary tree support vector machine	Experimental result	Increased by 37.58% on average
[146] 2023	IEMOCAP, MELD	Happy, Sad, Neutral, Angry, Frustrated, Excited, Surprise, Joy, Disgust, Fear	Speech and Text	Adaptive attention mechanism	Unimodal feature extraction		Acc and F1	Sad 80%, Happy 59.72%, neutral 82.56%, fear 14%, disgust 16.18%
[147] 2023	M-LFW-FER, CREMA-D	Neutral, Negative, Positive	A-V	Multimodal neural network using fusion techniques	Deep learning-based methodology	Multimodal neural network	Experimental evaluations	Achieved 79.81%
[148] 2023	MELD, RAVDESS	Happy, Sad, Neutral, Angry, Frustrated, Excited, Surprise, Joy, Disgust, Fear	A-V, Text	Multi-set Integrated Canonical Correlation Analysis (MICCA)	Video: Shot length, lighting key, motion, color, Audio: Zero-crossing rate, Mel frequency cepstral coefficient (MFCC), energy, pitch, Text: Unigram, bigram, TF-IDF	Hybrid deep learning model (variants of the CNN model)	P, R, F-score, Acc	MELD : 74.87%, RAVDESS: 95.25%

TABLE 6. *(Continued.)* **Summary of technical approaches of the selected and reviewed papers.**

[149] 2023	AFEW, SFEW, MELD, AffWild2		A-V, Texts, others	Technique: Multimodal architecture		Deep Learning	Acc, F1	Face AffWild2 53.98%, Audio AffWild2 46.93%, Text AffWild2 60.69%
[150] 2023	Generated by an emotive device	Sad, Happy, Neutral	EEG and audio signals	Integration of multi-channel information	PCA and Grey Wolf optimization algorithm	CNN	Acc, confuion Matrix	Achieved 94.44%
[151] 2023	MODMA, DAIC-WOZ, Extended DAIC, D-Vlog, K-EmoCon, SEED, SEED-IV, MELD, RAVDESS, MEISD	Sadness, lack of interest, and pleasure (associated with depression)	Multimedia (text, A-V) and unobtrusive physiological signals (e.g., electroencephalography)	Temporal convolutional transformer with knowledge embedding	Multimodal embeddings across domains via the temporal convolutional transformer	Joint task of depression detection and ER	P, R, F-score, Acc	MODMA 99%, DAIC-WOZ 92%, Extended DAIC 93%, D-Vlog 65%, K-EmoCon 60%, SEED 88%, SEED-IV 87%, MELD 73%, RAVDESS 92, MEISD 58%
[152] 2023	IEMOCAP, MSP-IMPROV	Happy, Angry, Sad, Neutral	A-V, Text	Multi-view attention mechanism	Multi-scale MFCCs	Multi-task learning with softmax cross-entropy loss and center loss	WA,UA	Happy and Angry are increased by 51% and 12% on IEMOCAP, 60% and 49% on MSP-IMPROV. Neural and Sad on IEMOCAP are increased from 79% and 72% to 82% and 74%
[153] 2023	MUStARD, Memotion, CMU-MOSEI, MELD	Sarcasm, sentiment, emotion	A-V, Text	Encoder-decoder architecture with intramodal and intermodal attention mechanisms	Separating task-specific representations from task-shared representations during the training process	Multitask learning model with single-level and multilevel decoders	P, R, micro-F1	Improvements of 1.9% to 2.8% in terms of Micro F1
[154] 2023	IEMOCAP, CMU-MOSEI	Happy, Sad, Angry, Fearful, Disgusted, and Surprised	Video frames, audio signals, text subtitles	Transformer-based fusion and representation learning method	Unified transformer architecture for learning a joint multimodal representation	Label-level representation approach for multi-label classification	Weighted accuracy (WA) and Un-average recall (UAR)	BERT+Finetune WA 78.98% UAR 78.98%, Pre-train+Finetune WA 79.63% UAR 80.61%,
[155] 2023	FABO		Facial, Body gestures	BDFN, CCA	BDFN	Coupling network	Simulation experiments	Increased by 1.15% compared to SVMRFE.
[156] 2022	HDT-BR, VIRI public	Negative mood states (tension, depression, anger, fatigue, confusion)	Visible and infrared thermal videos	Data reliability-focused multi-modal	Temporal-spatial	POMS-net model	Acc, P, R, F1	Tension 90%, depression 86%, anger 90%, fatigue 85%
[157] 2022	AMIGOS, DEAP	A,V	GSR, ECG, and EEG	Early sensor fusion		Two-class multi-modal classification	Acc	Achieved >76% for V, >73% for A
[158] 2022	Social media and Twitter		Multimodal		FFT-based Convolutional Neural Network (FFT-CNN)	Transfer Learning-based Support Vector Machine (TL-SVM)		Achieved 98%

TABLE 6. *(Continued.)* **Summary of technical approaches of the selected and reviewed papers.**

[159] 2022	IEMOCAP, MOSEI	Happy, Sad, Angry, Fear, Disgust, Surprise	A-V, Tex	COntextualized Graph Neural Network based Multimodal Emotion recognitioN (COGMEN)	Graph Neural Network (GNN) based architecture	Linear Classifier (Shared)	Acc, F1	COGMEN outperforms 7.7% F1-score increase for IEMOCAP
[160], 2022	CMU-MOSI, CMU-MOSEI (for sentiment analysis), AVEC2019 (for depression estimation)		Sentiment analysis and depression estimation	CubeMLP, a multimodal feature processing framework based entirely on MLP	CubeMLP consists of three independent MLP units, each with two affine transformations	Task predictions using flattened mixed multimodal features	Mean Absolute Error (MAE), Acc, F1, Pearson correlation (Corr), Concordance Correlation Coefficient (CCC)	State-of-the-art performance with a much lower computing cost
[161] 2022	CMU-MOSEI, MUStARD, Memotion	Positive, Negative, Neutral, Joy, Sad, Anger, Fear, Disgust, Surprise	Multi-modal records of human communication	External knowledge enhanced multi-task representation learning network (KAMT)	External knowledge augmentation layer	SVM, RF, CNN, BiLSTM, cLSTM, MHA-LSTM, SVM+BERT, RCNN-RoBERTa, EfficientNet, UPB-MTL, QMSA, A-MTL	P, R and micro-F1, balanced Acc	Higher F1
[142] 2022	IEMOCAP, MELD, CMU-MOSI	Neutral, Sad, Anger, Happy	Text, Audio	Adaptive interactive attention network (AIA-Net)	Text and acoustic features with different dimensions	Multiple collaborative learning (co-learning) layers	Acc , F1 , binary Acc , seven-class Acc	Higher Acc
[162] 2022	eNTERFACE'05 RAVDESS, CMEW	Anger, Disgust, Fear, Happy, Sad, and Surprise	A-V	Multimodal conditional Generative Adversarial Network (GAN)	Generators and discriminators for audio and visual modalities, Hirschfeld-Gebelein-Rényi (HGR) maximal correlation	DNN model	Acc	Increased
[163] 2021	IEMOCAP, CMU-MOSEI, CMU-MOSI	Positive, Negative, Neutral, Neutral, Happy, Sad, Anger, Frustrated, Excited	A-V, Text	Pair-wise attention mechanism, Trimodal fusion	RNN	Multinomial logistic regression	Acc and F1	
[164] 2021	RECOLA	Happy, Sad, Angry, A, V	A-V	Multimodal Markovian affect model	Bag-of-Audio-Words (BoAW) features, MFCCs, PCA	OMSVM (modeling static aspects), RankSVM (modeling dynamic aspects)	UAR, weighted Kappa	Quantified
[165] 2021	MELD	Anger, Disgust, Sad, Joy, Neutral, Surprise, Fear	Text and Acoustic	Concatenation operation	GRU cells for extracting global contextual information	Reinforcement learning framework	w-average Comparison	60.2%

TABLE 6. *(Continued.)* **Summary of technical approaches of the selected and reviewed papers.**

[166] 2021	IEMOCAP	Happy, Exited, Neural, Sad, Angry, Frustrated	Text, A-V	Cross-modal attention fusion module	Unimodal feature extraction	Gated recurrent units (GRUs)	Acc , F1	Acc 65%, F1 64%
[11] 2021	AVEC 2016	A, V	A-V	Multimodal attention mechanism	OpenSMILE, CNN	CNNs	R2 coefficient, CCC, PCC, root mean square error (RMSE)	CCC 0.729 for A, 0.718 for V
[167] 2021	IEMOCAP	Angry, Happy, Neutral, Sad	A-V	Multiobjective optimization algorithm	(DCNN) for speech modal, (DSCNN) for video image modal	Decision level fusion	Acc, R	Acc is 2.88% higher than ISMS-ALA model
[168] 2021	CMU-MOSI, POM	Anger, Contempt, Disgust, Fear, Joy, Sad, Surprise, Frustration and Confusion	Text, A-V	Attention fusion network	Bidirectional gated recurrent unit (Bi-GRU), new network initialization method	Sentiment classification and sentiment regression	Precision (PR), False positive rate (FPR), Recall (RE), Acc, F1 score, Mean absolute error (MAE), and Pearson product-moment correlation coefficients (Corr)	Improved accuracy of ER in three single modalities and overall video MER, outperforming
[134] 2020	LUMED-2, DEAP	Sad, Neutral, Happy, Angry, Disgust, Afraid, Surprised	GSR, EEG	Hybrid strategy (feature- and decision-level)	Inception ResnetV2 CNN model	Decision Tree	Maximum Acc, mean Acc	Maximum one-subject-out Acc of 81.2% and mean Acc of 74.2% on LUMED-2, Maximum one-subject-out Acc of 91.5% and mean Acc of 53.8% on DEAP
[169] 2020	IEMOCAP, MELD, CMU-MOSEI, CMU-MOSI	Happy, Sad, Angry, Neutral	Text, A-V	Transformers and Attention-based fusion mechanism	Independently pre-trained Self Supervised Learning (SSL) models, Self-Attention (SA) Transformer, Inter-Modality-Attention (IMA) based Transformer blocks		Acc, F1	The model achieves 53.3% for seven class sentiment Acc and 84.6% for binary sentiment Acc.
[170] 2020	IEMOCAP, CMU-MOSEI	Angry, Happy, Neutral, Sad, Disgust, Fear, Surprise	Text, A-V	Data-driven multiplicative fusion method	Canonical Correlational Analysis (CCA) for differentiating effective and ineffective modalities, proxy features for ineffectual modalities	Fully connected layers LSTMs (64-dimensional and 6-dimensional)	F1 scores, mean classification accuracies (MAs)	82.7% on IEMOCAP, 89.0% on CMU-MOSEI, which is an improvement of about 5% over prior work

TABLE 6. *(Continued.)* **Summary of technical approaches of the selected and reviewed papers.**

[171] 2019	IEMOCAP	Angry, Excited, Neutral, Sad	Speech and text	Combination of CNN and LSTM, and Bi-LSTM	CNN, LSTM, Bi-LSTM	Deep neural network	WA, UA	Acc increased by 6.70% for text and 13.85% for speech
[119] 2019	Dataset created	Happy, Sad, Fear, and Neutral	EEG and eye movements	Multimodal deep neural networks	Power spectral density (PSD) and differential entropy (DE), pupil diameter, fixation, saccade, and blink	SVM	Acc	Achieved 85.11
[172] 2018	HAPPEI, GAFF	Six level happiness intensity, Positive, Neutral, Negative	Face, Upper body, Scene	Localized multiple kernel learning	Feature descriptions of face, upper body, and scene	Group-level ER (GER)	four-fold cross-validation protocol	Achieved 66.67%
[173] 2018	Marathi, Benchmark	Angry, Happy, Fear, Neutral, Sad, Surprise	Speech (Cepstral feature, NMF feature, and pitch feature)	Hybrid PSO-FF algorithm	Multimodal fusion of speech features	Artificial neural network	Acc, Sensitivity, Specificity, P, FPR, FNR, NPV, FDR, F1, MCC	10.85% better than conventional models
[174] 2017	Profile of the participants	Sad, Angry, Calm, Surprised, Happy	A-V, everyday activities		Automatic Facial Features Extraction (AFFE)	SVMs and Decision Trees	R, Acc, F-measure, P	
[126] 2017	RECOLA	A, V	A-V	End-to-end training with correlation of each stream	CNN, Deep Residual Network	LSTM	CCC	Performance outperforms traditional approaches
[12], 2016	Generated	Happy, Sad, Surprise, Fear, Disgust, Angry, Neutral	A-V		Facial feature extraction		Validation by two raters from the psychology department	Achieved 72%
[175] 2016	Speech and facial datasets were built with Mexican users.	Angry, Happy, Neutral, Sad	A-V	Weighted integration of MER system	ANNs, HMMs	ANNs and HMMs	Cross-validation	Up to 97.00%
[176] 2015	IEMOCAP	V, A, Anxiety, Cold anger, Despair, Elation, Hot anger, Interest, Panic fear, Pleasure, Pride, Relief, Sad	A-V	Multiple classifiers trained on different modalities at various temporal lengths	Multi-timescale	SVM	LOOCV, UA	improved the Acc by 7.3% and 6.9%, For activation , 5.5% and 4.4% improvements. For V, 3.8% and 1.9% improvements.
[177] 2015	SAVEE	Anger, Fear, Disgust, Sad, Happy, Surprise	A-V	Feature-level, decision-level	Pitch, energy, duration, MFCC, visual features (2D marker coordinates)	Gaussian classifier	Average classification Acc	Comparable to human performance

TABLE 6. *(Continued.)* **Summary of technical approaches of the selected and reviewed papers.**

[178] 2014	AVEC 2012	V, A, Power, Expectancy	A-V, lexical	temporal Bayesian	OpenSMILE, lexical, LBP	Continuous real-value estimation	Cross-validation	96% improvement
[179] 2014	Their Own experimental Dataset	Positive, Neutral, and Negative	EEG signals and pupillary response from eye tracker	Feature level fusion and decision level fusion	Emotion-relevant features extracted from EEG signals and eye tracking data	Support vector machine (SVM)	Average Acc	EEG signals (71.77%) and eye tracking data (58.90%), and average Acc for feature level fusion (73.59%) and decision level fusion (72.98%)

accuracies for emotions such as happy, sad, and neutral on datasets like IEMOCAP and MELD. Similarly, Feng et al. [151] used a multi-view attention mechanism for audio-visual-text modalities, significantly improving recognition rates for happy and angry emotions on the IEMOCAP and MSP-IMPROV datasets. These studies demonstrate the effectiveness of text-based fusion techniques in capturing the nuances of emotions expressed through language.

Audio-Visual (A-V) and Physiological Signals:

Kang et al. [139] combined A-V modalities with physiological signals using a modality-adaptive fusion technique. They employed adversarial learning and plain regression tasks for feature extraction from speech signals, achieving an accuracy of up to 33% on the AffectNet dataset for emotions such as happy, sad, and angry. Similarly, Mocanu et al. [143] utilized A-V cross-attention mechanisms to process spatial and temporal features, achieving average accuracy scores of 89.25% on RAVDESS and 84.57% on CREMA-D for emotions including anger, disgust, and fear [143]. Shahzad et al. [146] also leveraged A-V modalities, incorporating deep learning-based fusion techniques to achieve an accuracy of 79.81% on datasets like M-LFW-FER and CREMA-D. Alamgir et al. [147] used multi-set integrated canonical correlation analysis (MICCA) on MELD and RAVDESS, achieving 74.87% and 95.25% accuracy, respectively. Aguilera et al. [148] achieved varying accuracies on AFEW, SFEW, MELD, and AffWild2 datasets using a multimodal architecture, with the highest being 60.69% for text on AffWild2. Feng et al. [151] applied a multi-view attention mechanism to IEMOCAP and MSP-IMPROV datasets, showing significant improvements in accuracy, particularly for happy and angry emotions. Zhang et al. [152] employed encoder-decoder architectures with attention mechanisms, achieving micro-F1 score improvements of 1.9% to 2.8% across multiple datasets.

Text, Audio-Visual (A-V), and Physiological Signals:

Other researchers have integrated text, AV, and physiological signals to capture a broader range of emotional cues. Zheng et al. [150] utilized a temporal convolutional transformer for depression and ER across multiple datasets, achieving state-of-the-art accuracy rates, such as 99% on MODMA and 92% on RAVDESS. Le et al. [153] employed a transformer-based fusion and representation learning method on IEMOCAP and CMU-MOSEI, achieving weighted accuracy (WA) of 79.63% and un-average recall (UAR) of 80.61%. [161] used multimodal conditional Generative Adversarial Networks (GANs) with increased accuracy across multiple datasets. Jaswal et al. [149] integrated multi-channel EEG and audio signals with PCA and Grey Wolf optimization, achieving 94.44% accuracy.

Physiological Signals:

Studies using physiological signals like ECG, EEG, and GSR have shown promising results. Selvi and Vijayakumaran [141] utilized physiological signals such as ECG, EEG, and GSR, applying FOX-optimized (DDQ) techniques for feature extraction and Q-learning for classification. This approach achieved greater than 95% accuracy for emotions like sad, angry, and happy on the ASCERTAIN dataset. Wang et al. [144] also employed physiological signals, using a multimodal information fusion technique and binary tree support vector machine, resulting in an average accuracy increase of 37.58%. Zheng et al. [119] integrated EEG and eye movement data, achieving 85.11% accuracy using multimodal deep neural networks. These results underscore the potential of physiological signals in providing reliable ER.

Facial and Body Gestures:

Chen et al. [154] utilized facial and body gesture modalities, employing BDFN and CCA techniques for fusion and achieving a 1.15% accuracy increase compared to SVM-RFE on the FABO dataset. Darekar and Dhande [172] focused on speech features, using a hybrid PSO-FF algorithm for multimodal fusion and achieving better accuracy compared to conventional models for emotions like angry, happy, and sad.

Visible and Infrared Thermal Videos:

Rong et al. [155] analyzed visible and infrared thermal videos using a data reliability-focused multi-modal approach, achieving accuracies of up to 90% for emotions such as tension, depression, and anger.

Multimodal:

Le et al. [153] integrated video frames, audio signals, and text subtitles using a transformer-based fusion method achieving weighted accuracy (WA) of up to 80.61% on IEMOCAP and CMU-MOSEI datasets. Zheng et al. [150] combined multimedia and physiological signals using a

TABLE 7. Summary of selected and reviewed papers.

Ref	Target Research Gap	Potential future research
[139] 2024	The paper targets limitations in facial expression-based ER by proposing a modality-adaptive fusion approach to improve state prediction of internal/external emotions.	Focus on addressing vulnerabilities of the proposed system to out-of-distribution inputs and exploring privacy-aware MER to mitigate potential societal impacts.
[140] 2024	Sentiment analysis becomes challenging when dealing with foreign languages, particularly without labeled data for training models.	Address the limitations of the current study, such as the dataset size and the sentiment analysis models used. Additionally, the model could be extended to include languages other than those investigated in this study.
[141] 2023	Focuses on recognizing specific emotions using a FOX-optimized Double Deep Q learning model with physiological signal models.	Explore the model's application in real-world settings and its adaptability to diverse physiological signal sources to recognize a broader range of emotions.
[143] 2023	Develops a deep attention model for discrete ER through multimodal fusion, incorporating intra-modal and cross-modal attention mechanisms.	Expand the model to handle emotion classification and regression tasks and incorporate textual information to enhance performance.
[144] 2023	Address the need for improved emotional state recognition in pilot training simulations by proposing a multimodal psychological signal fusion approach.	Applying the proposed method to enhance pilot training simulations, aiming to improve interaction and effectiveness through real-time psychological state monitoring and enhanced ER accuracy.
[146], 2023	Aim to enhance ER in multimodal conversations.	Refine the proposed hierarchical adaptive attention network (HAAN-ERC) and explore its application in different conversational contexts and datasets to validate its effectiveness further.
[147] 2023	Seek to improve accuracy and inclusivity in facial expression analysis, particularly under mask-wearing conditions.	Addressing cultural diversity variations and dataset biases will be crucial for achieving more robust and inclusive models, considering a broader range of expressions with and without masks.
[148] 2023	Emphasize the need for further exploration into evaluating ER models using additional datasets and incorporating EEG signals.	Evaluate the model with diverse datasets, analyze computational complexity, and extend the framework to incorporate EEG signals for a comprehensive understanding of emotional states.
[149] 2023	Highlight the importance of balanced and diverse data sources to enhance the performance of MER systems.	Creating new datasets, optimizing data preprocessing, and enhancing labeling for both unimodal and MER tasks are crucial steps to improve system performance.
[150] 2023	Suggest expanding emotive datasets and exploring additional modalities to improve ER, particularly for emotions like anger.	Enhancements could involve including additional emotional expressions and diverse test data to enhance the interpretation and applicability of the models in real-world scenarios.
[151] 2023	Highlight the potential of extending intelligent tools beyond depression and ER to address broader mental and physical disorders.	Exploring unified models connecting ER with depression detection, integrating multimedia data, and physiological signals can extend the application of intelligent tools to various mental and bodily disorders beyond depression and emotion.
[152] 2023	Explore the impact of multi-scale Mel-frequency cepstral coefficients (MFCCs) on speech ER performance.	Investigations could go deeper into understanding the effects of multi-scale MFCCs and explore the relationships between different data types to advance ER techniques.
[154] 2023	MER study is hindered by the lack of labeled corpora in terms of scale and diversity, due to the high annotation cost and label ambiguity.	Further, investigate a prompt-based learning method that can better adapt the pre-trained MEmoBERT to downstream tasks. Extensive experiments on two public datasets demonstrate the effectiveness and robustness of our proposed methods.
[153] 2023	Identify gaps in character-level interactions and address potential information conflicts in multi-modal fusion for tasks involving sentiment, sarcasm, and ER.	Future studies should focus on character-level interactions and advanced approaches to manage information conflicts in multi-modal data more effectively.
[155] 2023	Address the need for improved MER by combining facial and gesture emotional features and considering their correlation.	Enhancements could involve integrating multimodal emotional features with scene information to enhance emotional intention understanding, leading to smoother human-robot interaction and improved recognition accuracy in real-world scenarios.
[157] 2022	Propose a brain-inspired early fusion hyperdimensional computing architecture and optimization techniques to tackle memory-intensive ER tasks.	Enhancing arousal classification accuracy, exploring feature reduction techniques, and implementing the proposed methods in applications with larger numbers of channels and modalities.
[156] 2022	Highlight a gap in incorporating voice data for negative mood state detection.	Concentrate on specific feature extraction techniques for single negative mood states, particularly depression, to enhance stability and accuracy.
[158] 2022	Explore cognitive computing-based multimodal sentiment and emotion classification using machine learning techniques, focusing on social media data from platforms like Facebook and Twitter.	Explore the scalability and adaptability of the proposed cognitive computing-based multimodal sentiment and emotion classification technique in diverse social media contexts. Enhancing the model's robustness and efficiency, as well as extending its applicability to emerging platforms beyond Facebook and Twitter.
[159] 2022	Considering the influence of various speakers' utterances and their own emotional states. The challenge lies in effectively modeling both local (inter/intra dependency between speakers) and global (context) information in a conversation.	The focus on addressing misclassifications between similar classes and emotion shift cases by incorporating a component for capturing emotional shifts for fine-grained emotion prediction.
[160], 2022	Multimodal sentiment analysis and depression estimation are hindered by the need for effective fusion strategies that integrate mind-related information from different modalities, and the high computational cost associated with such integration.	Future research will focus on exploring more MLP-based techniques for multimodal fusion, conducting further experiments in other multimodal fusion domains such as ER, and aiming to establish MLP as a widely used solution for various tasks.
[161] 2022	Identify ignorance of intra-modal interaction and the impact of unclear labels on classification performance.	Solutions include data cleaning approaches for improved performance, exploration of tri-modal tri-task learning, designing unified multi-task models capturing correlations among tasks like sarcasm and humor, and developing enhanced multi-modal fusion strategies for text, image, and audio.

TABLE 7. *(Continued.)* **Summary of selected and reviewed papers.**

Ref	Target Research Gap	Potential future research
[142] 2022	Address challenges posed by multimodal uncertain data with class imbalance and incomplete multimodalities.	Development of a comprehensive multimodal fusion framework using an auxiliary structure, tackling class imbalance issues and addressing incomplete multimodalities.
[162] 2022	The challenge of data scarcity and class imbalance in audio-visual ER, which hinders the effectiveness of deep learning methods due to the difficulty and expense of acquiring large, balanced multimodal datasets.	Generalize the architecture to utilize additional modalities, such as text and physiological signals, to enhance ER. Integrate previous techniques such as Attention Mechanism, Canonical Correlational Analysis (CCA), and Tensor Fusion with the proposed multimodal conditional GAN framework. Address label noise in the data collection process to validate and improve the effectiveness of theMER framework.
[163] 2021	Emphasize multimodal fusion, interlocutor state, and context understanding in multimodal sentiment analysis.	Addressing multi-participant scenarios in conversational videos and exploring feature selection methods to enhance overall classification accuracy by considering emotion-specific features.
[164] 2021	Identify a gap in integrated modeling of static and dynamic aspects of emotion perception, with a focus on understanding the relative salience of audio and video modalities.	Refinement of multimodal approaches for emotion prediction, particularly in leveraging the superior advantage of the audio modality in modeling arousal and capturing emotion changes.
[165] 2021	Introduce a novel ER Reinforcement Learning Framework (EDRLF), considering the accumulation of emotions in conversation scenarios.	Enhancements aim to study conversations with arbitrary turns and speakers for even more accurate ER.
[166] 2021	Explore cross-modal interactions for ER in Conversation (ERC).	Refining discriminative utterance features for enhanced performance, with promising results indicating a foundation for continued investigation into optimizing MER in conversations.
[11] 2021	Assess the effectiveness of an improved AlexNet network with an attention mechanism for continuous dimension ER in audio and video bimodal scenarios.	Address the absence of integration between the deep CNN and the attention mechanism for facial expression feature learning. Enriching audio information through the introduction of audio manual features to enhance accuracy.
[167] 2021	Propose a MER model based on a multi-objective optimization algorithm, optimizing Acc and uniformity in recognition results.	Further exploration of multiobjective optimization algorithms for enhancing MER models. Investigation into the applicability of different multiobjective optimization algorithms in improving accuracy and uniformity of recognition results.
[168] 2021	The challenge of improving accuracy in video MER, particularly in handling weight consistency of each modality in multimodal fusion, and enhancing the robustness and effectiveness of time-contextual learning.	To enhance ER in audio and video, it is proposed to incorporate more high-dimensional acoustic features into the audio modality subnetwork using tools like OpenSMILE. Additionally, employing stacked Bi-GRUs can improve feature extraction and recognition accuracy through contextual learning methods. For video MER, utilizing hierarchical attention networks can better capture important information within and between modalities. Furthermore, exploring new fusion techniques or architectures is essential to surpass the performance of current attention fusion networks.
[134] 2020	Stress the importance of subject-independent ER for real-time applications.	Reducing the number of modalities, improving user-based recognition, and addressing practicality, comfort, and non-intrusiveness in sensor devices.
[169] 2020	Complexity in the representation and fusion of features across different modalities (language, facial expressions, and speech), and the challenge of effectively using high-dimensional features extracted from self-supervised learning (SSL) models.	Explore ways to fuse SSL features from other modalities like electroencephalogram (EEG) data. Investigate SSL algorithms that can learn joint information between video and text, and design SSL models that can learn joint representations between audio (speech), video, and text.
[170] 2020	The need for robust MER methods that can handle sensor noise and effectively combine cues from multiple modalities (face, text, and speech).	Explore more elaborate fusion techniques to improve the accuracy of ER. Extend M3ER to handle more than three modalities. Investigate more naturalistic modalities, such as walking styles and contextual information, as suggested in psychological studies. Consider multi-class classification to better reflect the subjective nature of human emotion perception, which resembles a probability distribution over discrete emotions.
[171] 2019	Identify the need for more effective feature fusion methods to enhance model performance in MER.	Optimize the speech modal model for more effective audio emotion features and consider other modalities in MER.
[119] 2019	Focuses on the development and evaluation of a MER framework called EmotionMeter, which integrates EEG signals and eye movements.	Focuses on enhancing the accuracy and usability of the ER system.
[173] 2018	Propose a novel approach to ER using adaptive neural network architecture and a hybrid PSO-FF algorithm for training.	Explore advanced classification approaches to achieve even higher accuracy in ER from speech signals, either by refining the existing model or exploring alternative methodologies.
[172] 2018	Focus on combining face, upper body, and scene information for ER.	Develop a flexible method to mitigate limitations caused by the superpixel algorithm in group-level ER. This involves exploring alternative approaches that can effectively capture relevant features while accommodating variations in image resolutions and semantic subregions.
[174] 2017	Explore real-time interaction sensors like heart rate and EEG, adopting a microkernel for flexibility across hardware environments.	Exploring real-time interaction sensors like heart rate and EEG, employing a microkernel for flexibility across hardware environments, emphasizing resource optimization, for devices with constraints such as smartphones and sensors.
[126] 2017	Focus on raw signal processing and contextual information integration using recurrent networks.	Expand the model to incorporate additional modalities, particularly physiological data, to improve ER performance. Experiment with more diverse emotion databases, including those with discrete labels, and explore tasks beyond ER to broaden the application of the proposed architectures.

TABLE 7. *(Continued.)* **Summary of selected and reviewed papers.**

[12] 2016	Aim to continuously monitor learners' behavior and interpret it into emotional states during e-learning.	Develop voice ER, integrate face and voice ER modules, and implement them in online e-learning environments.
[175] 2016	Integrate sub-systems for speech and vision based on standard techniques and genetic algorithms, achieving high recognition rates for ER with Mexican users.	Increase recognition rates for the speech system, explore alternative optimization techniques, test with different databases and recognition techniques, investigate alternative feature extraction methods, incorporate depth cameras for facial expression modeling, integrate psychology support for human emotion verification, and deploy humanoid robotic systems in therapy and care centers.
[176] 2015	Study the effectiveness of a multitemporal approach in emotion classification, showing that different emotion-related cues are best observed at varying temporal analysis lengths.	Explore the fusion of paralinguistic cues with other emotion classifiers, leveraging the proposed fusion method to further improve emotion classification accuracy. Additionally, enhances fusion performance by acquiring more diverse classifiers trained on different subsets of data.
[177] 2015	Effectively compare various feature selection criteria and fusion techniques for emotion classification on the SAVEE database.	Apply the current method to all speakers of the SAVEE database and extend it to other databases for broader validation. Investigate emotion classification in both speaker-dependent and speaker-independent scenarios.
[178] 2014	Propose a Bayesian fusion approach for continuous affect estimation from multiple modalities, focusing on video, audio, and lexical inputs.	Apply the proposed Bayesian fusion approach to other types of unimodal predictors, including temporal predictors, to assess its generalization capability. Additionally, investigate how the framework performs with different datasets and consider alternative fusion methods.
[179] 2014	The study demonstrates the effectiveness of combining EEG signals and pupillary response from eye tracking for ER.	Exploring the generalizability and robustness of the fusion model across diverse populations and real-world settings. Examining the feasibility of integrating additional modalities data to further accurate recognition.
End of Table		

temporal convolutional transformer, achieving remarkable accuracies across various datasets, with up to 99% on MODMA and 92% on RAVDESS for emotions associated with depression and interest.

Huan et al. [167] utilized an attention fusion network on CMU-MOSI and POM, showing improved accuracy in MER. Mittal et al. [169] employed a data-driven multiplicative fusion method on IEMOCAP and CMU-MOSEI, achieving 82.7% and 89.0% accuracy, respectively. These approaches highlight the effectiveness of integrating diverse modalities to enhance ER systems' performance.

Expanding beyond bimodal systems, some research has focused on MultiModal approaches, integrating audio, visual, and textual data. For instance, Poria et al. [179] fused these three modalities to create a robust ER system that outperformed traditional methods by capturing a richer set of emotional cues. Their work demonstrated that leveraging multiple modalities provides a more comprehensive understanding of human emotions, leading to superior performance. Similarly, Zadeh et al. [111] extended their earlier work by introducing the Multimodal Transformer, which effectively integrates audio, visual, and textual features through cross-modal attention mechanisms, achieving remarkable improvements in ER accuracy.

These studies collectively highlight the diverse approaches and modalities in MER. The integration of different modalities, sophisticated fusion techniques, and advanced machine learning models have significantly contributed to the improvement of ER systems, achieving notable accuracies across various datasets and emotional categories.

The next section provides a more in-depth discussion about the building blocks of MER.

A. MULTIMODAL EMOTION RECOGNITION BUILDING BLOCKS: ANSWER TO Q1

The foundation of MER lies in understanding the intricate interplay between various modalities and their contributions to the recognition of human emotions. This section offers the fundamental building blocks that underpin the development of robust MER systems. From feature extraction methods to fusion techniques. Through a comprehensive exploration of these building blocks, we aim to address question **Q1** and provide the MER building blocks. In addition to detailing the foundational building blocks, this section offers a comprehensive mapping of emotional expressions and the various methods employed for their detection alongside a discussion of the reviewed and available datasets.

1) STIMULI

Stimuli are different events, circumstances, or factors, either external or internal, that spark emotional responses in people. These stimuli can be of different types and sources, including sensory experiences (such as sight, Video [134], [141], sound, touch, taste, and smell), social interactions [11], [154], memories, thoughts, or physiological changes within the body. For example, seeing a loved one smile may evoke feelings of joy, while encountering a threatening situation might lead to feelings of fear or anxiety. Stimuli are important in initiating emotional reactions and influencing subsequent emotional experiences and behaviors. Stimuli can also be context-based, varying according to the environment in which they are presented, such as classrooms or business meetings [144], [180].

2) DATA ACQUISITION

Data acquisition is the process of collecting raw data from various sources, including sensors, instruments, databases, or external systems, with the aim of capturing information relevant to a specific domain or problem [142]. This phase involves identifying the sources of data, designing and deploying data collection methods, and retrieving the required information in a structured format. Data acquisition encompasses activities such as sensor calibration, signal

processing, and data recording, ensuring the accuracy, reliability, and integrity of the collected data [152]. The acquired data serves as the foundation for subsequent analysis, modeling, and decision-making processes in various fields, including scientific research, industrial applications, and data-driven decision support systems [154].

The following methods for ER, derived from reviewed papers, correspond to the emotion expressions discussed in the background analysis section of this paper. While sketch-based approaches [181], [182], are not widely prevalent, their inclusion in the review and analysis reflects the increasing research interest in these areas.

Various methods have been explored to capture and interpret emotional cues from different sources. Facial emotion and audio detection methods, as evidenced by studies such as [143] and [146], utilize **facial expressions** and **audio** signals to infer emotional states. Similarly, the integration of speech, voice, and tone detection with facial and textual cues, as discussed in research like [147], provides a comprehensive approach to capturing emotional nuances. **Physiological signals**, including **Electrocardiography (ECG)**, **Galvanic Skin Response (GSR)**, **Electroencephalography (EEG)**, and **Blood Pressure (BP)**, offer valuable insights into emotional states, as demonstrated in studies such as [139], [141], and [144]. **Eye-tracking** techniques, as explored in [119] and [178], provide additional cues about where individuals focus their attention, contributing to a more holistic understanding of emotional responses. Text and linguistic analysis, combined with speech signals, as investigated in [145] and [150], enables the extraction of emotional content from written and spoken language. **Gesture and body movement** recognition, coupled with facial expressions, as explored in [154], further enriches the multimodal approach to ER. Finally, emerging techniques such as **sketch recognition**, as discussed in [181], offer new paths for understanding and interpreting emotional expression in diverse contexts.

Among the reviewed papers, facial, speech, and textual modalities emerged as the most extensively studied and experimentally explored examples. From a psychological standpoint, they offer complementary information for emotional detection [66]. Based on the modes of emotional expression outlined in the background section, a mapping is provided in Fig. 14. This mapping illustrates how various techniques enable researchers to detect expressed emotions and the corresponding modalities used for detection. Expressions can manifest in various modalities, categorized as unimodal, bimodal, or multimodal. For example, speech can be deconstructed into text, tone, and pitch, while painting analysis may involve capturing body movements alongside the artwork itself. Alternatively, expressions can remain unimodal, such as heart rate. Part of this discussion, addressing Q3, explores analytical approaches and modalities selection methods in more detail.

3) DATASETS

Datasets play a crucial role in the development, testing, and benchmarking of models. Table 5 lists reviewed datasets, each offering unique characteristics that collectively cover a broad spectrum of emotions and data modalities. These datasets incorporate multiple types of data, including A-V, text, physiological signals, and video, allowing for comprehensive studies that leverage the strengths of each modality to enhance ER accuracy. A-V datasets, such as IEMOCAP, CMU-MOSI, and AFEW, are prevalent due to the rich information provided by both speech and facial expressions, which are essential for capturing nuanced emotional states. Physiological signal datasets like SEED and DEAP offer insights from biological responses, providing objective measures of emotional states. Text datasets, such as K-EmoCon, focus on the linguistic content, capturing the emotional tone and context of communication.

The datasets cover a wide range of emotions, from basic ones like happiness, sadness, and anger to more complex states such as frustration, excitement, and neutrality. This extensive coverage is vital for training models that can generalize across different emotional expressions. Some datasets, like LUMED-2, focus on fewer emotions to provide depth in specific emotional categories, while others, like CMU-MOSEI, offer a broader spectrum. Larger datasets, such as CMU-MOSEI and SEED, provide extensive training data beneficial for deep learning models that require large amounts of data to achieve high accuracy. The complexity of datasets varies, with some including multimodal data types, such as physiological signals and video in ASCERTAIN, making them suitable for studies on integrating different data sources.

Key datasets in MER include IEMOCAP, CMU-MOSEI, SEED, SEED-IV, DEAP, AFEW, and SFEW. IEMOCAP is widely used for its detailed annotations and multimodal recordings, covering a comprehensive range of emotions. CMU-MOSEI is known for its large size and diversity, offering over 23,000 sentence-level annotations with A-V data covering a wide spectrum of emotions. SEED and SEED-IV are notable for their use of EEG signals along with video data, providing a unique perspective on how physiological signals can enhance ER. DEAP, focusing on physiological signals and video, is instrumental for studies exploring the integration of biological and behavioral data. AFEW and SFEW, derived from real-world movie clips, provide challenging, in-the-wild data that help develop models robust to naturalistic variations in emotional expression.

However, integrating diverse data types from various datasets remains a challenge but also an opportunity. Effective fusion techniques can exploit the complementary nature of audio, visual, and physiological signals to improve ER performance. The variety in emotional labels and annotation methods across datasets can lead to inconsistencies, highlighting the need for standardizing annotations and emotional categories to enhance the comparability of results.

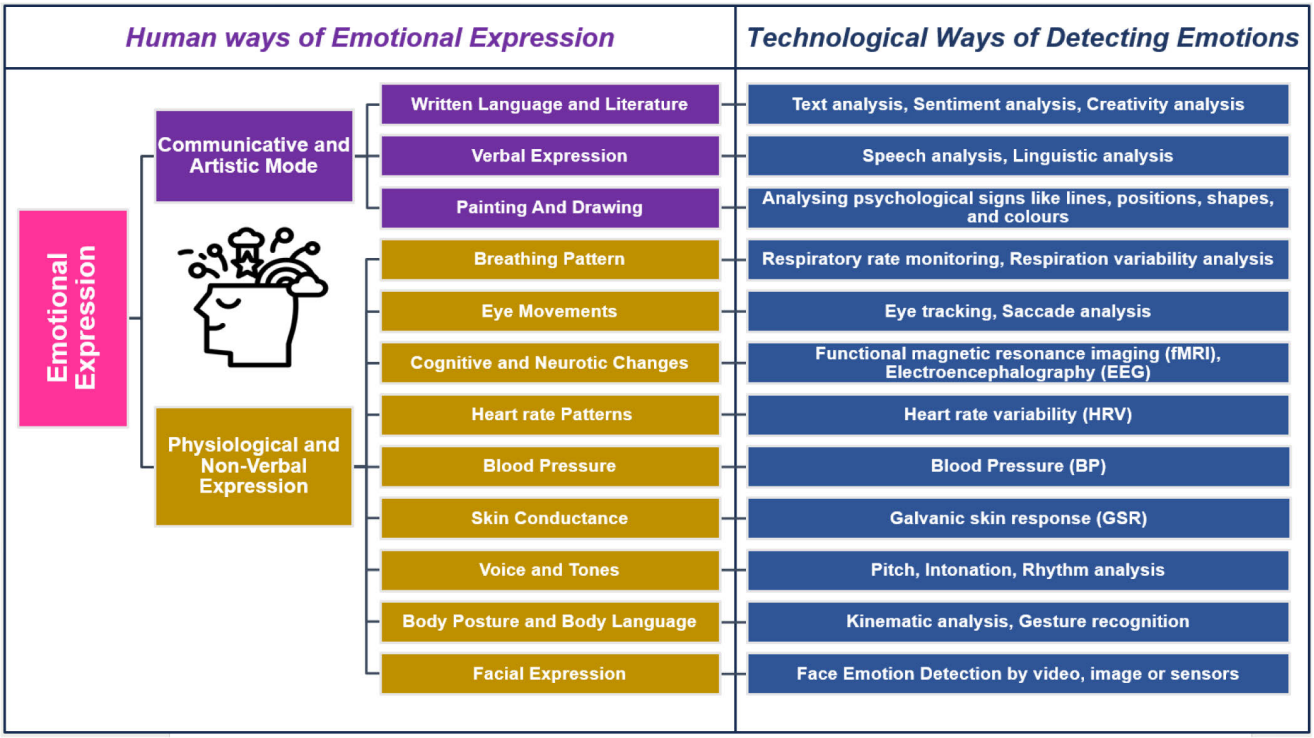

FIGURE 14. Human emotional expression and the corresponding modalities used for emotion detection.

Datasets like AFEW, collected in naturalistic settings, pose additional challenges due to their variability but are crucial for developing models that perform well in real-world applications. In contrast, controlled datasets like IEMOCAP and SEED provide high-quality, consistent data but may not capture the full range of variability found in everyday emotional expressions. The increasing adoption of self-supervised learning (SSL) methods can leverage the vast amounts of unlabeled data in these datasets, particularly useful for those with limited labeled examples [168]. SSL can help in pretraining models that are then fine-tuned on smaller labeled datasets, improving generalization.

4) PREPROCESSING AND FEATURE EXTRACTION

It is the initial steps taken to prepare the data obtained from different modalities before it is fed into the ER model and classification. These steps are crucial for ensuring that the data is clean, consistent, and suitable for analysis [154], [162].

a: DATA CLEANING [126]

This step involves identifying and handling missing or erroneous data points, removing noise or outliers, and addressing any inconsistencies or inaccuracies in the data. In MER, this might include synchronizing audio and video data streams, removing irrelevant background noise from audio recordings, or correcting misalignments in visual data to ensure high-quality inputs for the models.

b: FEATURE EXTRACTION [164]

Features relevant to ER need to be extracted from the raw data. This may involve transforming the data into a more suitable representation, extracting relevant features using signal processing techniques, or applying feature extraction algorithms specific to each modality. For example, in audio data, this might involve extracting Mel-frequency cepstral coefficients (MFCCs), while in visual data, it could involve detecting facial landmarks or extracting features using convolutional neural networks (CNNs).

c: FEATURE SCALING AND NORMALIZATION [172]

Ensuring that the features across different modalities are on the same scale can help prevent certain features from dominating the analysis. Techniques such as scaling and normalization are applied to standardize the range of feature values. These steps are particularly important when dealing with multimodal data sources as they help to bring all features into a common scale, facilitating the training of more accurate models. For instance, normalization can significantly impact the performance of models like Support Vector Machines (SVMs) and Neural Networks (NNs), which are sensitive to the scales of the input features. In MER, this could involve

normalizing pixel values in images and amplitude values in audio to ensure they contribute equally to the model.

d: DIMENSIONALITY REDUCTION [172]

Multimodal data can be high-dimensional, making analysis computationally expensive and prone to overfitting. Dimensional reduction techniques such as principal component analysis (PCA) or feature selection methods may be employed to reduce the complexity of the data while retaining relevant information. These techniques not only enhance the efficiency of the models but also help in removing noise from the data, thus improving the overall model performance. In the context of MER, PCA can be used to reduce the dimensional of facial expression data, while Linear Discriminant Analysis (LDA) might be applied to find the most informative features across both audio and visual modalities.

Another useful technique for dimensional reduction and feature learning from large and complex datasets is Autoencoders [152]. They are neural networks designed to learn compressed representations of data. They are used for feature extraction and dimensionality reduction, making them useful for preprocessing multimodal data before feeding it into classification models.

e: DATA AUGMENTATION

Advanced data augmentation techniques can be employed to artificially increase the diversity of the training data. For visual data, this might include random cropping, rotation, and flipping of images. For audio data, techniques such as time stretching, pitch shifting, and adding background noise can be used. Data augmentation helps in improving the generalization of the model by exposing it to a wider variety of data samples [160].

f: MULTIMODAL DATA ALIGNMENT [26]

Proper alignment of data from different modalities is crucial for effective ER. Techniques such as Dynamic Time Warping (DTW) and Canonical Correlation Analysis (CCA) can be used to align and synchronize data streams from different sources, ensuring that the temporal relationships between modalities are accurately captured [183].

g: ADVANCED NEURAL NETWORK ARCHITECTURES FOR FEATURE EXTRACTION [161], [184], [185], [186]

More neural network architectures such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) can be used for feature extraction. These models are capable of learning complex and high-dimensional feature representations from multimodal data. VAEs can learn latent representations that capture the underlying structure of the data, while GANs can generate synthetic data to augment the training set and improve model robustness.

h: GENERATIVE MODELS FOR DATA AUGMENTATION [161], [184], [185], [186], [187]

Using GANs, Variational Autoencoders (VAEs), and Large Language Model (LLM)s to create synthetic data that can augment training datasets. These generative models help address class imbalances and improve the robustness of MER systems. This can help in augmenting datasets and addressing the scarcity of labeled data in some modalities.

5) DATA FUSION TECHNIQUES

Multi-modal fusion aims to enhance sample representation efficiency by combining knowledge from different modalities. Two widely used paradigms for multi-modal fusion are early fusion (EF) and late fusion (LF). EF captures relations across multi-modal features, and since fusion occurs early, only one classifier needs to be learned later. Various methods have been explored for EF, including feature concatenation, shared representation learning, tensor fusion, and attentive fusion. In contrast, LF combines local decisions based on individual features, offering more flexibility. Approaches for LF include maximum voting, linear weighting, and Dempster-Shafer (D-S) evidence theory. However, LF tends to overlook multi-modal feature correlations [160].

a: EARLY FUSION (EF) [156]

In EF, the features extracted from different modalities are combined at an early stage before feeding them into the classifier. This approach involves concatenating or merging the features from different modalities into a single feature vector. Techniques used in EF include:

- Feature Concatenation: Directly combining features from different modalities into a single vector.
- Shared Representation Learning: Learning a shared space where features from different modalities can be mapped and combined effectively.
- Tensor Fusion: Combining features through tensor operations to capture high-order interactions between modalities.
- Attentive Fusion: Using attention mechanisms to weigh the importance of features from different modalities dynamically.

b: LATE FUSION (LF)

In contrast to EF, LF combines the decision outputs or confidence scores from individual classifiers trained on each modality. These outputs are then fused at a later stage using techniques such as averaging, voting, or weighted summation. Techniques used in LF include:

- Maximum Voting: Combining decisions based on the majority vote from individual classifiers.
- Linear Weighting: Assigning different weights to the outputs of classifiers before combining them.
- D-S Evidence Theory: Using a probabilistic framework to combine evidence from different sources to reach a decision.

c: MID-LEVEL FUSION

This approach involves extracting higher-level features or representations from individual modalities before combining them. It allows for the integration of more abstract and semantically rich information from each modality.

d: HYBRID FUSION [134], [162], [172]

Hybrid fusion methods combine multiple fusion techniques to leverage their respective strengths. For example, a system might use EF to combine low-level features and mid-level fusion to integrate higher-level representations, providing a more comprehensive fusion strategy.

e: FEATURE-LEVEL FUSION [176]

In this approach, features extracted from different modalities are combined into a single feature vector. This can involve concatenating features, averaging them, or applying more complex operations to merge the information from each modality.

f: DECISION-LEVEL FUSION [176]

This method combines the outputs of individual models or classifiers trained on different modalities to make a final decision. This can involve voting schemes, such as averaging or taking the majority vote, or more sophisticated methods like weighted averaging or stacking.

g: ATTENTION-BASED FUSION [11], [139], [142], [145], [151], [152], [162], [165], [188]

Attention mechanisms dynamically weight the contribution of different modalities based on their relevance to the task at hand. This allows the fusion system to focus on the most informative parts of each modality while ignoring irrelevant or noisy information. Attention-based fusion has emerged as one of the most commonly utilized methods for integrating data from various modalities due to its flexibility and effectiveness.

h: GRAPH-BASED FUSION [158]

Utilizing Graph Neural Networks (GNNs) to model relationships and interactions between different modalities. GNNs can effectively capture the dependencies between various features and modalities, enhancing the robustness and accuracy of ER models.

Multimodal Fusion [168], [189] can be integrated into multimodal frameworks to facilitate the fusion of features from different modalities (e.g., audio, visual, text). Their ability to capture complex relationships and dependencies can enhance the performance of MER systems. Fusion techniques vary significantly across studies. Zhang et al. [145] used an adaptive attention mechanism on speech and text modalities, achieving specific accuracies like 80% for sad and 59.72% for happy emotions. Ren et al. [165] utilized cross-modal attention fusion with gated recurrent units (GRUs) on IEMOCAP, achieving 65% accuracy. Liu et al. [11] implemented multimodal attention mechanisms on AVEC 2016, showing high concordance correlation coefficients (CCC) for arousal and valence. For example:

- Cross-Modal Transformers: Models like MEmoBERT utilize transformer blocks to capture cross-modal interactions between text, audio, and visual features.
- Attention Mechanisms: LLMs can incorporate attention mechanisms to selectively focus on relevant parts of input data from different modalities, improving the fusion process.
- Transformers: Leveraging transformer architectures to capture long-range dependencies and interactions between modalities. Multimodal transformers use self-attention mechanisms to integrate information across text, audio, and visual data, leading to more accurate emotion predictions.

Table 8 summarizes the reviewed fusion techniques based on their suitable input modalities, strengths, limitations, and applied techniques.

6) CLASSIFICATION

Classification refers to the process of categorizing or labeling data into predefined classes or categories based on their features or attributes. Classification involves assigning emotions to different observations or instances based on their characteristics [175], [176]. This process typically involves training a machine learning model using labeled data, where each instance is associated with a specific emotion. The model learns patterns or relationships between the input features and the corresponding emotions [160], [190]. Once trained, the model can classify new, unseen data into predefined emotion categories. The accuracy and effectiveness of the classification process depend on the quality of the training data, the choice of features, and the robustness of the classification algorithm [118].

Support Vector Machines (SVM) [191], [192] is a type of supervised learning algorithm used for both classification and regression tasks. SVM aims to find the optimal hyperplane that best separates different classes in the feature space, making it suitable for both linear and non-linear relationships between features and classes. It is particularly effective with small to medium-sized datasets. Zeng et al. [119] utilized SVM classification in order to detect emotions from EEG signals and eye movements.

Decision Trees [193], [194], [195] is another popular machine learning algorithm used for classification and regression tasks. They partition the feature space into regions based on feature values, making decisions based on simple rules at each node, thus providing an intuitive model structure. Cimtay et al. [134] used decision trees in order to recognise emotions fro GSR and EEG signals.

Random Forests [196], [197], [198], [199] are an ensemble learning method consisting of multiple decision trees. This technique involves building each tree on a random subset of the training data and averaging their predictions to improve accuracy and reduce overfitting. Random Forests are robust

TABLE 8. Guidelines of the characteristics and strengths of each fusion technique.

Fusion technique	Suitable Modalities	Strengths	Limitations	Techniques
Early Fusion (EF)	Ideal for combining low-level features directly extracted from different modalities.	Captures detailed information from each modality early in the processing pipeline, allowing for comprehensive feature integration.	Can result in very high-dimensional feature vectors, leading to increased computational complexity and potential overfitting. Also, it might not handle missing or noisy data effectively.	Feature Concatenation, Shared Representation Learning, Tensor Fusion, Attentive Fusion.
Late Fusion (LF)	Best when each modality can be processed independently with separate classifiers.	Enables integration of decision outputs or confidence scores from individual classifiers trained on different modalities.	May miss out on capturing interactions between modalities since the integration occurs at a decision level. Performance depends heavily on the quality of individual classifiers.	Maximum Voting, Linear Weighting, D-S Evidence Theory.
Mid-level Fusion	Effective for integrating higher-level, semantically rich representations extracted from individual modalities.	Combines abstract features that capture deeper relationships between modalities.	Still can be computationally intensive and might require significant preprocessing to extract meaningful high-level features.	Extracts and combines features after initial processing stages to enhance interpretability and integration.
Hybrid Fusion	Combines multiple fusion techniques to leverage complementary strengths across different modalities.	Provides flexibility and robustness by integrating both low-level and high-level features effectively.	Complex to implement and optimize, potentially requiring more resources and sophisticated architectures.	Using EF for low-level feature integration and mid-level fusion for abstract representation integration.
Feature-level Fusion	Suitable when direct integration of raw or processed features is beneficial.	Facilitates comprehensive utilization of multimodal features at a fundamental level.	High-dimensional vectors can lead to increased computational cost and overfitting. Difficulty in handling missing data.	Feature concatenation, averaging, or applying more complex operations to merge the information from each modality.
Decision-level Fusion	Works well when each modality can provide an independent assessment of emotions.	Simplifies the fusion process by dealing with classifier outputs rather than raw features.	May lose nuanced interactions between modalities and rely heavily on individual classifier performance.	Voting schemes (average, majority vote), weighted averaging, or stacking to enhance decision-making.
Attention-based Fusion	Dynamically weights the contribution of different modalities based on their relevance to the task.	Effective for scenarios where modalities vary in importance or relevance over time or context.	Flexibility in focusing on informative parts while mitigating noise or irrelevant information.	Can be computationally intensive and requires careful tuning of attention mechanisms.
Graph-based Fusion	Uses GNNs to model relationships and interactions between modalities.	Beneficial when capturing complex dependencies and interactions between features from different modalities.	Enhances robustness and accuracy by integrating structured relationships in multimodal data.	Computationally expensive and requires expertise in GNNs. The performance can be sensitive to the graph structure and parameters.
Transformers for Multimodal Fusion	Utilizes transformer architectures to capture long-range dependencies and interactions across text, audio, and visual modalities.	Effective for integrating information across diverse and complex modalities.	Improves accuracy in capturing nuanced interactions and dependencies between modalities.	Very high computational requirements and complex architecture that requires large datasets and significant computational resources.

to noise and outliers and can handle both numerical and categorical data, making them effective for tasks involving complex relationships and interactions among features.

Artificial Neural Networks (ANN) [200] are a class of machine learning models inspired by the structure and function of the human brain. These models consist of interconnected nodes organized in layers and are capable of learning complex patterns in data. Deep learning models, including CNNs and Long Short-Term Memory networks (LSTM), fall under this category and have shown remarkable performance inMER. Deep learning models excel at learning hierarchical representations of data and capturing complex patterns, making them particularly effective when working with large datasets as they can automatically extract features from raw data.

Convolutional Neural Networks (CNN) [201] is a type of neural network particularly effective for processing grid-like data, such as images. CNNs utilize convolutional layers to automatically learn hierarchical features from the data, making them suitable for tasks requiring spatial hierarchies of features. In the study by Liu et al. [11], CNNs were used to process facial expressions and vocal intonations simultaneously, demonstrating improved ER accuracy through effective feature extraction from both image and audio data.

Long Short-Term Memory networks (LSTM) [202] are a specialized type of Recurrent Neural Network (RNN) designed to overcome the vanishing gradient problem in traditional RNNs. LSTMs are capable of learning long-range dependencies, making them well-suited for sequential data tasks such as text and speech processing. Poria et al. [120]

and Cai et al. [170] employed LSTMs to analyze textual, audio, and visual cues over time, achieving significant improvements in ER by capturing the temporal dynamics of emotions.

K-Nearest Neighbors (KNN) [190] is a simple and intuitive classification algorithm that works by finding the k nearest neighbors to a given data point in the feature space and using majority voting to assign a class label. This method is particularly useful for its simplicity and effectiveness in certain contexts.

Recurrent Neural Networks (RNN) [203] are designed to handle sequential data by maintaining a form of memory through connections that form loops, allowing information to persist. This looping mechanism enables RNNs to effectively model sequences of data, making them useful for tasks like time series prediction, speech recognition, natural language processing, and handwriting recognition.

MDL techniques integrate multiple modalities to leverage complementary information. For instance, Zadeh et al. [111] introduced a Tensor Fusion Network that combines text, audio, and video features into a single framework, outperforming unimodal approaches by effectively fusing multimodal data.

Transformers [153], [204] have revolutionized natural language processing and are now being adapted for MER. Tsai et al. [205], [206] developed a multimodal transformer model that integrates language, acoustic, and visual data, using attention mechanisms to capture interactions between different modalities, leading to more accurate emotion predictions.

Hybrid models combining CNNs, LSTMs, and attention mechanisms have shown promising results. Sun et al. [159] developed a hybrid architecture that integrates CNNs for feature extraction and LSTMs with attention mechanisms for sequential modeling, demonstrating enhanced performance in recognizing emotions from multimodal data. In a setting where a model must analyze video data to detect emotions, CNNs can first extract spatial features from the video frames, while LSTMs capture temporal dependencies, and attention mechanisms focus on the most relevant parts of the input data, culminating in a robust ER system. Reference [140] used ensemble models of transformers and large language models (LLM) for text in various languages, achieving over 86% accuracy in sentiment analysis. Reference [157] applied FFT-based CNN and transfer learning-based SVM on social media data, achieving 98% accuracy. Reference [169] used a data-driven multiplicative fusion method on text and AV modalities, achieving 82.7% on IEMOCAP and 89.0

Recent classification techniques reviewed in this paper include several innovative approaches. The Contextual Attention-Based LSTM (CAT-LSTM) [179] leverages attention mechanisms within LSTM networks to focus on contextually relevant features for ER. Similarly, the Multi-Modal Attention Network (MMAN) [207] employs an attention network to integrate multiple modalities, enhancing the accuracy of ER by dynamically weighting the importance of each modality. The Multiplicative ER Model (M3ER) [169] uses multiplicative interactions between different modalities to capture complex relationships, leading to improved emotion classification performance.

The Bi-GRU Based Approach [167] utilizes bidirectional GRU (Gated Recurrent Unit) networks to process sequential data from different modalities, capturing both past and future context for better ER. In contrast, the TransModality Approach [208] integrates transformer networks to handle multimodal data, taking advantage of transformers' powerful attention mechanisms for enhanced feature extraction and classification. The Interactive Multimodal Attention Network (IMAN) [165] introduces an interactive attention mechanism that allows different modalities to influence each other, leading to more robust ER.

Another notable technique is the Multi-Channel Weight-Sharing Autoencoder [209], which shares weights across multiple channels, efficiently learning shared features from different modalities to improve ER accuracy. The MEmo-BERT [189], as a pre-trained model for MER, combines the strengths of BERT with multimodal inputs to achieve state-of-the-art performance. Additionally, the Hybrid Contrastive Learning Approach [210] uses a combination of contrastive learning techniques along with fully connected (FC) and Softmax layers to refine feature representations and improve classification accuracy.

The Feed-forward Network (FFN) with a linear FC layer and a Softmax layer is employed in the Multimodal End-to-End Sparse Model (MESM). This model processes multimodal data end-to-end, leveraging sparse representations to enhance ER efficiency and performance.

SSL has emerged as a significant and impactful area of research in representation learning as well as in the MER, providing researchers with access to pre-trained SSL models that capture various data modalities, especially for tasks where labeled data may be limited or costly to obtain. [168].

Generative Adversarial Networks (GANs) [211], [212] have also been explored in MER. GANs consist of a generator and a discriminator network, where the generator aims to produce realistic data samples, and the discriminator attempts to distinguish between real and generated samples. In MER, GANs have been used to generate synthetic data for underrepresented emotion categories, improving the robustness and generalization of ER models.

Another advanced technique involves the use of Graph Neural Networks (GNNs) [158], which can model relationships between different data points in a graph structure. In the context of MER, GNNs can be employed to capture the interactions and dependencies between different modalities and features, leading to improved ER performance.

While many of the mentioned techniques are standard across various classification tasks, their application to ER tasks necessitates a careful selection of appropriate methods and references to ensure relevance and effectiveness in

capturing emotional nuances from the data. The choice of the most suitable classification technique ultimately depends on factors such as the specific characteristics of the dataset, computational resources available, and the desired balance between performance and interpretability.

Table 9 summarizes the classification techniques along with their suitable input data and modalities.

7) MODEL EVALUATION

It involves assessing the performance and effectiveness of the trained model on unseen data to ensure its reliability and generalization capability.

Acc measures the proportion of correctly classified instances out of all instances. While it's a simple and intuitive metric, it may not be suitable for imbalanced datasets. P measures the proportion of correctly predicted positive instances out of all instances predicted as positive, while R measures the proportion of correctly predicted positive instances out of all actual positive instances. These metrics are particularly useful for imbalanced datasets, where the positive class is rare. F1 is the harmonic mean of precision and recall, providing a balanced measure of a model's performance on both precision and recall. **Confusion Matrix** provides a comprehensive breakdown of the model's predictions across different classes, showing true positives, false positives, true negatives, and false negatives. ROC and AUC, curves plot the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. AUC represents the area under the ROC curve and provides a measure of the model's ability to discriminate between positive and negative instances. **Cross-Validation**, techniques such as k-fold cross-validation, partition the dataset into multiple subsets for training and testing iteratively. This helps assess the model's performance across different data samples and reduces the risk of overfitting.

B. CURRENT STATUS ANALYSIS AND MODALITY EVALUATION: ANSWER TO Q2

By a dynamic and rapidly evolving landscape marked by ongoing research, development, and deployment efforts aimed at unlocking new frontiers in understanding and harnessing human emotions in technology-mediated environments [10]. These advancements have propelled the field forward, enabling the integration of diverse modalities to capture and interpret emotional states with greater accuracy and granularity. Breakthroughs in machine learning algorithms, particularly deep learning architectures, have played a pivotal role in enhancing the sophistication and efficacy of MER systems. Furthermore, the spread of sensors, wearable devices, and multimedia technologies has expanded the scope and applicability of ER across various domains and applications. Additionally, the growing availability of large-scale annotated datasets and benchmarking platforms has facilitated rigorous evaluation and comparison of different approaches, fostering innovation and driving continuous improvement in MER algorithms and techniques [11], [82], [119], [150], [155].

1) CURRENT STATE

Recent advancements in the field of MER have focused on several key areas, each contributing to enhanced performance and robustness of MER systems.

a: DEEP LEARNING ARCHITECTURES

CNNs [162] have proven effective in extracting spatial features from images and video frames for facial ER. These networks excel at identifying patterns in visual data, making them ideal for analyzing facial expressions and subtle emotional cues present in images and videos. RNNs and LSTM [126] networks are useful for modeling temporal dependencies in sequential data such as speech and video, which enhances ER accuracy. By retaining information over time, these architectures can capture the dynamic nature of emotions as they evolve, particularly in A-V data.

Transformer models [140] are employed for their ability to handle long-range dependencies in sequential data, significantly improving ER performance. Their self-attention mechanism allows for better context understanding and integration across different modalities, leading to more accurate emotion predictions. GNNs [158] capture relational information from multimodal data, which is crucial for understanding complex emotional contexts. GNNs excel in situations where the relationships between different data points, such as interactions between speakers and listeners, play a vital role in ER.

Pretrained models and transfer learning have also made significant contributions to the field. Models like VGG and ResNet [134], which are trained on large datasets, can be fine-tuned on smaller ER datasets. This approach leverages the prior knowledge embedded in these models, leading to improved performance and faster convergence during training.

Self-supervised learning involves pre-training models on tasks that do not require labeled data, followed by fine-tuning on smaller labeled datasets. This approach significantly improves ER accuracy by leveraging large amounts of unlabeled data to learn useful representations, which are then refined with the available labeled data [168]. Semi-supervised learning combines labeled and unlabeled data to improve learning efficiency and model performance. By using a mix of both types of data, these techniques can extract more information from the available datasets, making them particularly effective in low-resource settings where labeled data is scarce [168].

Unlike traditional supervised learning, which relies heavily on human-annotated labels, SSL [168] harnesses the unlabelled data to generate its own supervisory signals. This innovative technique enables a model to predict parts of its input from other parts by finding patterns, correlations, and structures within the data. The versatility of SSL is evident in its wide range of applications, from enhancing

TABLE 9. Classification comparison.

Classification	Suitable For	Modalities	Application
SVM	Small to medium-sized datasets with a limited number of features.	Text and audio.	SVMs can be effective for combining textual and acoustic features due to their ability to handle linear and non-linear relationships. They are less effective for large-scale datasets or highly complex feature spaces typical in video data.
Decision Trees	Intuitive and interpretable models, smaller datasets.	Text and structured data.	Decision trees can work well for combining text features with other structured data (e.g., demographic information) but may struggle with high-dimensional data such as raw audio or video.
Random Forests	Handling complex relationships among features, robustness to noise. Large datasets, high dimensional data.	Text, audio, and structured data.	Random forests can effectively handle mixed types of data, making them suitable for combining text and audio features. They are also good at managing medium-sized datasets.
ANN	Learning complex patterns, handling large datasets.	Text, audio, and visual.	ANNs, especially deeper networks, are highly effective for MER involving large datasets and complex feature spaces, such as those involving both text and video data.
CNN	Spatial data, high-dimensional inputs.	Visual (images, video frames).	CNNs are particularly suited for extracting spatial features from visual data. When combined with audio (using separate branches for each modality), they can significantly enhance ER accuracy.
LSTM	Sequential data, time-series analysis.	Text and audio.	LSTMs are well-suited for handling temporal dependencies in sequential data, such as speech and text, making them effective for analyzing temporal patterns in emotions.
KNN	Smaller datasets, simple pattern recognition.	Text and audio.	KNN can be used for simple classification tasks and works well with smaller datasets, but it may struggle with high-dimensional multimodal data.
RNN	Sequential data, time-series analysis.	Text and audio.	RNNs are suitable for modeling sequential dependencies in text and audio data and useful for temporal aspects of ER.
Transformers	Large datasets, complex sequential data.	Text, audio, and video.	Transformers are highly effective for processing large-scale sequential data, leveraging self-attention mechanisms to capture long-range dependencies in multimodal inputs.
Hybrid Models (CNN + LSTM + Attention)	Complex multimodal data integration.	Text, audio, and video.	Hybrid models combine the strengths of CNNs, LSTMs, and attention mechanisms to effectively capture spatial, temporal, and contextual features from multimodal data.
GANs	Data augmentation, generating synthetic data.	Video and images.	GANs can generate realistic synthetic data to augment training datasets, improving the robustness of ER models, particularly for visual and A-V data.
GNNs	Structured data, relational data.	Multimodal sensor data.	GNNs can model relationships and interactions between different modalities, suitable for integrating structured multimodal data from various sensors.
Linear Fully Connected (FC) Layer and Softmax Layer	Output Layer for classification tasks.	All modalities.	Fully connected and softmax layers are commonly used at the output stage of neural networks to perform the final classification of emotional states from the integrated multimodal features.
SSL	Large-scale datasets where labeled data is scarce or expensive to obtain.	Text, audio, and visual	SSL techniques can learn representations from large amounts of unlabeled data, which can then be fine-tuned for specific tasks like ER. This is particularly useful in multimodal setups where acquiring labeled data for all modalities can be challenging.

computer vision systems to improving natural language processing tasks. The method is particularly advantageous in scenarios where acquiring labeled data is impractical, thus democratizing the use of deep learning across various domains.

LLMs like BERT [213], GPT [187], and their variants are highly effective in understanding and processing textual data. These models excel at extracting nuanced features and contextual information, which are crucial for accurate ER. By integrating LLMs into multimodal frameworks, the performance of ER systems can be significantly enhanced especially regarding the textual data. Case studies and applications of LLMs in ER have demonstrated notable improvements in feature fusion and overall accuracy. These advancements highlight the potential of LLMs to enhance the capabilities of multimodal frameworks, making them more robust and reliable in various practical applications.

b: FUSION TECHNIQUES

Feature-level fusion [176] involves combining features from different modalities, such as audio and visual data, to create a comprehensive representation of emotions. This approach enhances recognition capabilities by leveraging the strengths of each modality, resulting in a more robust and detailed understanding of emotional states. Decision-level fusion integrates outputs from unimodal classifiers to make final predictions, which improves the overall robustness and accuracy of ER systems. By considering the individual predictions of each modality, this technique ensures that

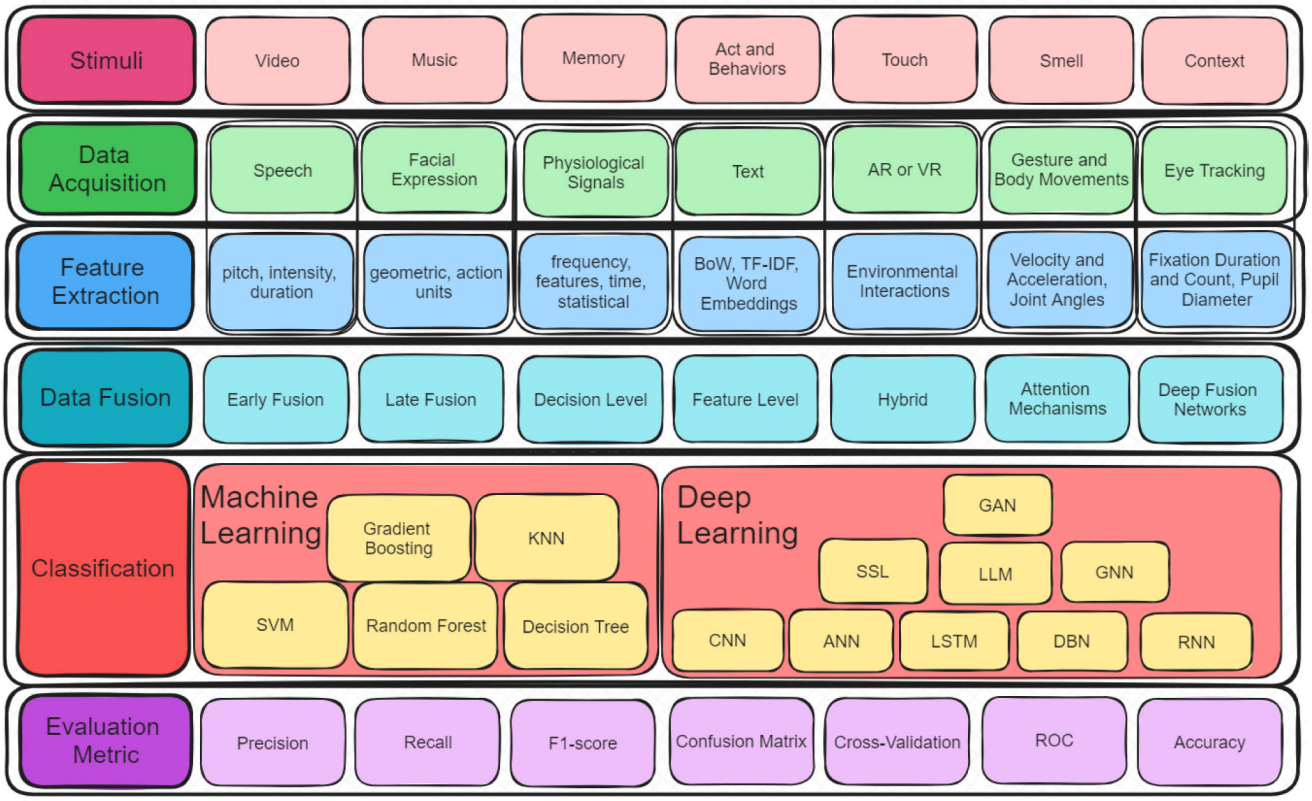

FIGURE 15. Multimodal emotion recognition building blocks.

the final decision is more reliable and less prone to errors from any single modality. Hybrid fusion approaches leverage the strengths of both feature-level and decision-level fusion [176] to enhance performance further. By combining the detailed feature integration of feature-level fusion with the robustness of decision-level fusion, these hybrid methods offer a more balanced and effective solution for MER. Advanced fusion techniques, including dynamic and adaptive fusion strategies [139], change based on the input context, thereby improving fusion effectiveness. These methods adapt the fusion process to the specific characteristics of the data being processed, leading to more accurate and contextually appropriate ER. Attention mechanisms play a crucial role in modern fusion techniques. By focusing on the most relevant parts of the data, these mechanisms enhance feature integration and overall system performance. Attention mechanisms allow the system to prioritize important features and ignore less relevant information, thus improving the accuracy and efficiency of ER.

c: DATA PROCESSING AND AUGMENTATION

GANs generate synthetic data to augment training datasets, addressing the issue of limited annotated data and improving model robustness. By creating realistic and diverse samples, GANs help in expanding the training data, which enhances the ability of models to generalize and perform well on unseen data [161]. Domain adaptation techniques adapt models trained on one dataset to work effectively on another, thereby reducing the dependency on large-scale annotated data. These techniques allow models to transfer knowledge from one domain to another, making them more versatile and applicable to various datasets without requiring extensive retraining. Data imputation techniques handle missing data in multimodal datasets through imputation or synthetic data generation, enhancing data quality and model performance. By filling in gaps in the data, these techniques ensure that the models can make use of the complete dataset, leading to more accurate and reliable ER. Various augmentation strategies, such as noise addition, time warping, and feature masking, are employed to enhance model robustness and performance. These methods introduce variability into the training data, helping models become more resilient to different types of noise and distortions that they might encounter in real-world scenarios [214].

d: REAL-TIME AND CONTEXT-AWARE SYSTEMS

Real-time ER has seen significant advancements due to improvements in hardware and software, enabling the real-time detection and interpretation of emotions. This capability is crucial for interactive applications, where timely responses to emotional cues are essential [134]. Context-aware ER incorporates contextual information, such as

environmental factors and user interactions, to enhance the accuracy of ER systems. By considering the surrounding context, these systems can make more informed and accurate predictions. Deploying MER systems on edge devices for real-time applications improves system responsiveness and usability. Edge computing allows for the processing of data close to the source, reducing latency and enabling quicker reactions to emotional inputs. Context-aware personalization tailors ER systems to individual users based on contextual information. This personalization improves the accuracy and relevance of ER by adapting to the unique characteristics and preferences of each user, resulting in more meaningful and accurate emotional insights [145].

e: REDUCING DEPENDENCY ON LARGE-SCALE ANNOTATED DATA

Data augmentation through synthetic data generation using GANs creates realistic and diverse training samples. This approach enhances model robustness and performance by providing a richer and more varied dataset, addressing the issue of limited annotated data. Transfer learning involves pre-training models on large datasets and then fine-tuning them on smaller, domain-specific datasets [157]. This method leverages the knowledge acquired from the pre-training phase, resulting in significantly improved performance even when the amount of annotated data is limited. Few-shot learning techniques enable models to learn from very few examples, which enhances their applicability in real-world settings where annotated data is scarce. By training models to generalize from minimal data, few-shot learning makes it possible to achieve high performance with limited labeled samples [153].

f: IMPROVING RECOGNITION CAPABILITIES UNDER SMALL SAMPLE OR UNSUPERVISED CONDITIONS

Synthetic data generation using GANs is a powerful method for creating additional training examples, which improves model performance and generalization. By generating realistic data that mirrors the characteristics of the original dataset, GANs help in augmenting the training set effectively. Self-supervised learning leverages unlabeled data for pre-training models, significantly improving their performance on target tasks even with minimal annotated data. This approach allows models to learn useful representations from large amounts of unlabeled data, which can then be fine-tuned with a smaller labeled dataset to achieve high accuracy. Few-shot learning techniques are designed to generalize from a limited number of samples by leveraging prior knowledge. These methods have shown competitive performance in ER tasks, making them a viable solution for scenarios with limited data availability.

g: MULTIMODAL DATASETS AND BENCHMARKING

The development of large-scale, diverse MER datasets such as the Aff-Wild2 database [125], the CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset [111], and the AFEW-VA database [123] has provided researchers with valuable resources for training and evaluating MER systems. Additionally, challenges and competitions like the ER in the Wild (EmotiW) [214] and the A-V ER Challenge (AVEC) have spurred innovation by providing standardized benchmarks for evaluating MER systems.

2) MODALITY EVALUATION AND COMPARISON

There are still several areas that need improvement including, **Accuracy and Reliability** [155], While MER systems have achieved impressive results, there is still room for improvement in terms of accuracy and reliability. Many existing systems struggle with accurately recognizing subtle or complex emotions, especially in real-world and dynamic contexts.

Generalization and Robustness [164], [172], MER systems often lack generalization across diverse populations, cultural backgrounds, and environmental conditions. There is a need to develop more robust and adaptable models that can perform effectively in various contexts and settings.

Data Availability and Diversity [155], The availability of high-quality, diverse, and well-annotated datasets is crucial for training and evaluating MER systems. However, existing datasets often suffer from limitations such as small sample sizes, limited emotional variability, and biases in demographic representation.

Interpretability and Transparency [103], Many MER systems operate as black boxes, making it difficult to interpret how they arrive at their predictions. There is a need for more transparent and interpretable models that provide insights into the features and cues used to make emotional predictions.

Ethical and Social Implications [103], MER raises important ethical and social implications related to privacy, consent, bias, and potential misuse of emotional information. There is a need for guidelines and best practices to ensure that research and application of MER are conducted ethically and responsibly.

The overall summary of the modalities covered in our review study for ER with the evaluation by the selected criteria are shown in Table 10 and Fig. 16. It is noteworthy to mention that the summary is drawn based on our observations from the papers included in the systematic review. The ratings range from Bad to Very Good, with Medium being an average rating.

From a social perspective, Facial expression recognition systems are widely accessible due to the availability of facial recognition software and hardware, which are commonly integrated into smartphones, cameras, and other devices. Ethical considerations for facial expression analysis include privacy concerns related to data collection and consent, as well as potential biases in facial recognition algorithms [105]. Facial expression analysis is relatively easy to use, especially with the availability of pre-trained facial recognition models and user-friendly software interfaces [88].

Speech and voice data are easily accessible through audio recording devices, smartphones, and online platforms,

making them readily available for analysis. Ethical considerations for speech analysis include issues of privacy, consent, and potential biases in speech recognition systems, particularly regarding dialects and accents [105]. Speech analysis requires specialized software and expertise for accurate interpretation, but user-friendly tools and APIs are available for basic speech recognition and emotion analysis [81].

Physiological sensors for detecting signals like heart rate variability and galvanic skin response may require specialized equipment and expertise, limiting their accessibility compared to other modalities. Ethical considerations for physiological detection include issues of informed consent, data privacy, and potential psychological or physical discomfort for participants during data collection. Physiological sensors require proper calibration and placement for accurate data collection, and their interpretation may require expertise in psychophysiology and signal processing [215].

Eye-tracking technology has become more accessible with the development of affordable eye-tracking hardware and software, although high-quality eye-trackers may still be expensive. Ethical considerations for eye-tracking include issues of informed consent, data privacy, and potential intrusiveness of eye-tracking technology [105]. Eye-tracking systems require proper calibration and setup for accurate gaze tracking, but user-friendly software interfaces and calibration procedures are available [216].

Gesture and body movement data can be captured using motion capture devices, cameras, or wearable sensors, but access to these tools may vary depending on the context and application. Ethical considerations for gesture and body movement analysis include issues of consent, privacy, and potential biases in gesture recognition algorithms [105]. Analyzing gesture and body movements may require specialized equipment and expertise in motion capture and biomechanics, but user-friendly software interfaces are available for basic analysis [217].

Sketch data may be less accessible compared to other modalities, as it requires manual creation or digitization of sketches, which may not be readily available in large datasets [218]. Ethical considerations for sketch analysis include issues of copyright, intellectual property, and potential biases in sketch recognition algorithms [105]. Analyzing sketches may require specialized software tools and expertise in image processing and pattern recognition, but user-friendly sketch recognition software is available for basic analysis [219].

Text and linguistic data are highly accessible through various sources such as books, articles, social media, and online platforms, making them readily available for analysis. Ethical considerations for text analysis include issues of privacy, consent, and potential biases in natural language processing algorithms [105]. Analyzing text and linguistic data is relatively easy with the availability of natural language processing tools and libraries, enabling straightforward sentiment analysis and ER [220].

From Technical aspects, Facial expressions can be effectively integrated with other modalities such as speech and physiological signals to enhance ER systems. Many research [139], [143], [146], [162], [173], [177] demonstrate successful fusion of facial expressions with other modalities. Real-time facial expression analysis systems have been developed but may face challenges in processing large datasets or complex facial expressions [154].

Speech and voice analysis can be combined with facial expressions, physiological signals, and linguistic features to improve ER accuracy [139], [142], [143], [145], [146], [149]. Real-time speech ER systems have been developed with high accuracy and low latency [149].

Physiological signals can complement other modalities like facial expressions and speech to provide a more comprehensive understanding of emotional states [134], [149], [150]. Physiological sensors can provide real-time data with minimal delay, allowing for effective real-time ER [141], [173].

Eye-tracking data can be combined with facial expressions and physiological signals to enhance ER accuracy [178]. Real-time eye-tracking systems have been developed for ER, but they may encounter challenges in handling large-scale data or complex stimuli [178].

Gesture and body movement data can be integrated with facial expressions and speech but may have limited compatibility with physiological signals [154], [156], [171]. Real-time gesture recognition systems have been developed, but they may face challenges in capturing subtle movements or recognizing complex gestures [154].

Sketch data may have limited integration potential with other modalities due to its unique format and interpretation. Research on integrating sketches with other modalities is relatively sparse [221]. Real-time sketch recognition systems have been developed, but they may have limitations in handling complex sketches or recognizing detailed features [221].

Textual and linguistic features can be easily integrated with other modalities such as speech, facial expressions, and physiological signals to improve ER accuracy [142], [145], [147], [152], [162], [164], [177]. Real-time text and linguistic analysis systems have been developed with satisfactory performance, but they may face challenges in processing large volumes of text data in real-time [152].

From the scientific point of view, Facial expressions have extensive theoretical support from psychology, including the widely cited Facial Action Coding System (FACS) [88] developed by Paul Ekman and Wallace V. Friesen. Speech and voice analysis benefit from established theories such as the emotional prosody hypothesis [222], which posits that intonation and rhythm convey emotional information. It was also evident from the corpus analysis conducted on the abstracts, as detailed in Section IV that most

publications are dedicated to Video and Speech (See Fig. 10, Fig. 11, and Fig. 12). Physiological signals like heart rate variability [223] and galvanic skin response [224] are firmly grounded in physiological psychology and neuroscience, with well-established links to emotional states. Eye-tracking techniques [225] draw upon theories of visual attention and cognitive processing, although the direct mapping to emotions may not be as well-defined as other modalities. While body language and gestures [226] are recognized as important in communication, their direct association with specific emotions may vary depending on cultural and contextual factors. Theoretical support for emotions expressed through sketches is limited in technical publications compared to other modalities, as it relies more on visual interpretation and artistic expression. There is concrete theoretical support for sketch interpretation, such as the GoodMan [227] and House Tree Person (HTP) techniques [228], which provide robust rules for the process. Text and linguistic analysis draw upon theories of semantics [229], pragmatics, and sentiment analysis [230], providing a solid theoretical and publication foundation for understanding emotional expression through language.

Although the evaluation for a single modality to combine provides useful info it is also good to consider the to choose modalities that provide complementary information about emotions. For example, combining facial expressions from video data with physiological signals like heart rate variability can offer a more comprehensive understanding of emotional states. select modalities that are contextually relevant to the task or scenario. For instance, in a customer service interaction, combining text transcripts with voice recordings can capture both verbal and non-verbal cues to assess customer sentiment effectively. Analyze the requirements of the specific ER task. Different tasks can benefit from different modalities. For example, detecting emotional sentiment in social media posts may require analyzing text, while assessing emotional engagement in multimedia content may necessitate combining video and audio modalities. Consider the complexity and performance implications of combining multiple modalities. Some modalities may require more complicated feature extraction or fusion techniques, leading to increased model complexity. Evaluate the trade-offs between model complexity and performance gains when selecting modalities.

Further elaboration on the present challenges and potential applications is provided in the subsequent two subsections in an attempt to respond **Q3** and **Q4** as well.

C. APPLICATIONS: ANSWER TO Q3

MER have found diverse applications across various domains, including learning and education, retail markets, the health sector, gaming, and tourism. The context and applications that can take advantage of the multimodal techniques are more but here are the ones which has been extracted from the reviewed papers.

Learning and Education [12], [144], [180], understanding students' emotional states during learning activities can be used. By analyzing multimodal emotional inputs, educators can gain insights into students' engagement, interest, frustration, and other emotional states. This information can inform instructional strategies, personalized learning experiences, and interventions to support students' socio-emotional development.

Health Sector [14], [112], MER can be used in various applications, such as mental health assessment, patient monitoring, and therapeutic interventions. By analyzing patients' facial expressions, voice tone, physiological signals, and verbal responses, healthcare providers can assess patients' emotional states, detect signs of distress or agitation, and tailor treatment plans accordingly. This technology can also support health services, remote patient monitoring, and virtual therapy sessions. Specifically in mental health, This technology can support early intervention, personalized treatment planning, and remote monitoring in mental healthcare settings.

Gaming [183], MER can enhance user engagement, immersion, and personalized gaming experiences. By analyzing players' facial expressions, body movements, voice intonations, and physiological responses during gameplay, game developers can adapt game content, difficulty levels, and narrative elements in real time to match players' emotional states and preferences. This can lead to more dynamic and emotionally resonant gaming experiences that increase player satisfaction and enjoyment.

Tourism [13], MER can enhance travelers' experiences and satisfaction during their journeys. This information can be used to personalize travel itineraries, recommend personalized activities, and provide targeted assistance and support to enhance tourists' overall satisfaction and enjoyment.

Decision-Making Assistance [14], MER detection can reveal the individuals' emotional states, preferences, and decision-making processes. By analyzing the data, decision-makers can better understand stakeholders' emotional responses to different options, assess their levels of engagement and satisfaction, and make more informed and empathetic decisions. This technology can be particularly useful in contexts such as customer service interactions, team collaborations, and leadership roles.

Human-computer interaction [10], [154], [174], In the design of human-computer interaction systems, MER can enhance user experiences, engagement, and satisfaction. This technology can improve the usability, accessibility, and effectiveness of interactive systems across various domains, including entertainment, education, healthcare, and productivity.

Social Robots ad social Interactions [11], [154], [171], [231], MER can enable social robots to perceive and respond to human emotions more effectively, enhancing their social capabilities and interaction experiences. It can applied in various contexts, such as companion

TABLE 10. Collection of recognition modalities and comparing them by selected criteria.

Method Criteria	Accessibility	Ethical Consideration	Ease of Use	Integration with other Modalities	Real-time Performance	Scientific Publication	Theoretical Background
Facial Expression	Very Good	Good	Good	Good	Medium	Very Good	Very Good
Speech and Voice	Very Good	Good	Very Good	Very Good	Good	Very Good	Good
Physiological Detection	Medium	Medium	Medium	Very Good	Good	Good	Very Good
Eye-Tracking	Medium	Medium	Medium	Good	Medium	Good	Good
Text and Linguistic	Good	Good	Very Good	Good	Very Good	Very Good	Good
Gesture and Body movements	Medium	Medium	Medium	Medium	Medium	Good	Medium
Sketch	Medium	Medium	Bad	Bad	Bad	Medium	Good

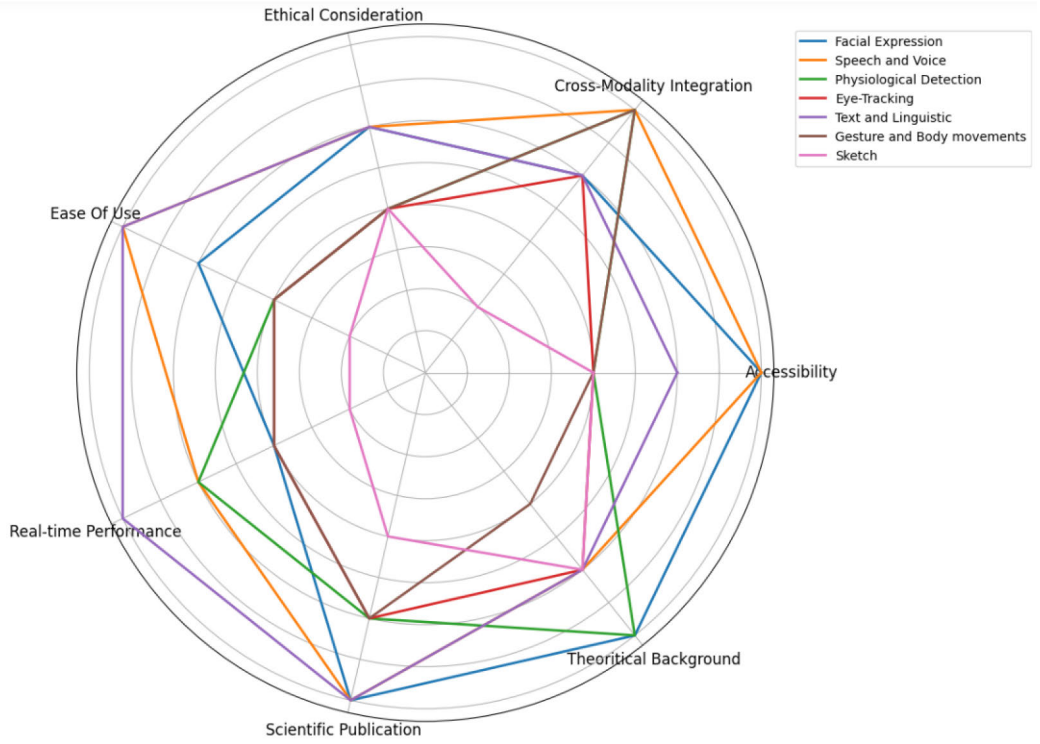

FIGURE 16. Radar graph of emotion recognition modalities evaluation based on the selected criteria.

robots for the elderly, educational robots for children, and service robots in healthcare, retail, and hospitality settings.

Creativity Detection [232], MER can be used to assess and analyze individuals' creative processes and outcomes. By analyzing various expressive modalities, such as facial expressions, verbal narratives, gestures, and physiological responses, researchers and practitioners can identify patterns, markers, and indicators of creativity in different domains, such as art, music, literature, and design. This technology can support the development of automated creativity assessment tools, creative collaboration platforms, and personalized creativity coaching systems.

D. CHALLENGES: ANSWER TO Q4

This section addresses the final research question (Q4) in which various challenges in the MER are discussed. Challenges need to be addressed and identified to improve the accuracy and reliability of ER systems.

Dataset availability and accuracy [155], the availability and accuracy of datasets pose significant challenges in MER. Obtaining large, diverse, and accurately annotated datasets across multiple modalities can be difficult. Moreover, ensuring the accuracy and reliability of labels for emotional states in multimodal data is crucial for training robust ER models.

Emotion Ranges and Context dependency [146], [152], emotions are complex and nuanced, and capturing their full

range presents a challenge in MER. Emotions vary across individuals, cultures, and contexts, making it challenging to define comprehensive emotion categories and accurately recognize them across different modalities.

Emotion Recognition Technique [152], selecting appropriate techniques for ER across modalities is another challenge. Different modalities may require different feature extraction methods, machine learning algorithms, or fusion strategies. Identifying the most effective techniques for each modality and integrating them cohesively poses a significant challenge.

Fusion techniques [142], [160], [166], [169] challenges include determining the optimal way to combine modalities, handling missing or incomplete data from certain modalities, and addressing the potential conflicts or redundancies between modalities. Additionally, selecting the appropriate fusion strategy and ensuring that the fusion process does not introduce noise or distort the original information are also important challenges. There is a lack of consensus on the most efficient mechanisms for combining or fusing information from different modalities in MER. Various fusion techniques, such as EF, LF, or decision-level fusion, exist, each with its advantages and limitations. Determining the optimal fusion mechanism for a given task or context remains a challenge.

Feature selection [163] challenges include dealing with high-dimensional data, selecting features that capture both individual and cross-modal characteristics of emotions, and handling the varying levels of importance or relevance of features across different modalities. Additionally, feature selection methods must be robust to noise and outliers in the data, and they should be scalable to handle large datasets efficiently. Moreover, ensuring that the selected features effectively capture the underlying emotional content while minimizing computational complexity is another important challenge.

Extension of the MER to other context specially real world [164], [172] beyond controlled laboratory settings presents challenges. Real-world environments introduce additional complexities such as noise, variability, and dynamic interactions, which may affect the performance of ER systems. Adapting MER models to diverse real-world contexts while maintaining accuracy and reliability is a significant challenge.

Complexity of Context Integration means the inclusion of contextual information (e.g., situational, environmental, or relational data) to improve the accuracy of ER. Emotions are often influenced by context, and ignoring it can lead to incorrect interpretations. Incorporating context requires additional data sources and can increase the complexity of the model.

Noisy Data refers to any data that contains errors, inconsistencies, or irrelevant information that can obscure the true signal. Noise can stem from various sources such as background noise in audio data, occlusions or lighting variations in visual data, and irrelevant or redundant information in text data. In Audio Data, background noise, speech overlaps, and low-quality recordings can significantly affect the accuracy of ER systems. For Visual Data, variations in lighting, facial occlusions (e.g., glasses, masks), and changes in head pose can introduce noise that complicates the extraction of meaningful emotional features. Text data may contain slang, typos, and ambiguities that can mislead sentiment analysis.

Imbalanced Datasets [161] occur when certain classes (emotions) are underrepresented compared to others. This imbalance can bias the learning process, causing the model to perform well in the majority classes but poorly in the minority classes. Models trained on imbalanced datasets may be biased towards the majority classes, leading to poor recognition of less frequent emotions. Standard metrics like accuracy can be misleading when datasets are imbalanced. Precision, recall, and F1-score become more relevant but are still challenging to optimize for minority classes.

VI. LIMITATIONS

The criteria used in this literature review might have led to the exclusion of some relevant publications from the analysis, mainly because of the diverse buzzwords associated with emotion, feelings, and MER.

However, the selected works over a 10-year period provide a representative sample of the research conducted, albeit dispersed. To mitigate this limitation, a combination of human knowledge and AI paper selection was developed. Also, the exclusion of non-English articles may have overlooked important publications. Furthermore, due to limited research on characterizing MER, certain sections of this analysis rely on empirical evidence from the literature. These sections offer a comprehensive perspective of the field and its core elements.

VII. CONCLUSION AND FUTURE WORK

The current work presents a systematic literature review of MER considering an important database, i.e., Scopus. The main objective is the description of the related concepts (Background description), trends (exploration of main building blocks, modalities, applications, and current challenges), and potential research directions. The review follows a PRISMA analysis which is driven by five research questions and covers 37 articles.

In section II, we investigate renowned emotional models and discuss three potential approaches for emotional detection: unimodal, bimodal, and multimodal. We also conduct a comparative analysis of each technique and subsequently propose criteria for selecting the most suitable modality, which is later employed to address *Q2*.

A comprehensive technical analysis, coupled with an analytical approach to identifying targeted gaps and potential future directions, is presented in Section IV. Additionally, this section includes corpus analysis of selected paper abstracts to extract the most repetitive phrases, including mono, bi, and trigrams. This approach aids in providing in-depth analytics for addressing the research question of the current review

paper across various modalities examined in the publications. Finally, a heatmap representing the active countries in the scientific domain is included for visualization.

The exploration of the diverse building blocks within a MER framework, addressing *Q1*, is delineated in Section V. This subsection provides an overview of emotional expressions and potential detection methods, complemented by a compilation of frequently encountered datasets identified in the review. The discussion on The examination of the current status of MER and evaluating different modalities in response to *Q2* is presented in Section V. This subsection includes a comparative table and a radar graph. To our knowledge, this study marks the first attempt to introduce concrete criteria for evaluation and analysis of the modalities in the experimental framework of MER. The discussion of the various challenges for MER, addressing *Q3*, is detailed in Section V. Additionally, the exploration of various applications for MER, addressing *Q4*, is provided in the same section. In both cases, we strive to encompass both the most trending and the latest developments to augment existing reviews.

For future directions, it would be beneficial to explore novel fusion techniques that effectively integrate multiple modalities to enhance ER accuracy and robustness would be valuable. Additionally, investigating the potential of emerging technologies such as edge computing in advancing MER systems could open up new avenues for research. Moreover, there is a growing need to address ethical considerations surrounding the collection, processing, and storage of multimodal data for ER to ensure privacy, transparency, and user trust. Lastly, fostering interdisciplinary collaborations between researchers from psychology, neuroscience, computer science, and engineering could facilitate a holistic approach to MER, leading to more comprehensive reviews.

REFERENCES

[1] M. Cabanac, "What is emotion?" *Behavioural Processes*, vol. 60, no. 2, pp. 69–83, 2002.
[2] J. T. Cacioppo and W. L. Gardner, "Emotion," *Annu. Rev. Psychol.*, vol. 50, no. 1, pp. 191–214, 1999.
[3] J. E. Laird, *Emotion*. Cambridge, MA, USA: MIT Press, 2012.
[4] B. Mesquita, N. H. Frijda, and K. R. Scherer, "Culture and emotion," in *Handbook of Cross-Cultural Psychology: Basic Processes and Human Development*, vol. 2. London, U.K.: Pearson, 1997, p. 255.
[5] D. Matsumoto, "Culture and emotion," in *The Handbook of Culture and Psychology*. Oxford, U.K.: Oxford Univ. Press, 2001, pp. 171–194.
[6] J. R. Davitz, *The Language of Emotion*. New York, NY, USA: Academic, 2013.
[7] I. J. Roseman, N. Dhawan, S. I. Rettek, R. K. Naidu, and K. Thapa, "Cultural differences and cross-cultural similarities in appraisals and emotional responses," *J. Cross-Cultural Psychol.*, vol. 26, no. 1, pp. 23–38, Jan. 1995.
[8] D. Keltner and J. S. Lerner, "Emotion," in *Handbook of Social Psychology*. Hoboken, NJ, USA: Wiley, 2010.
[9] Z. Xie and L. Guan, "Multimodal information fusion of audiovisual emotion recognition using novel information theoretic tools," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2013, pp. 1–6.
[10] S. Kalateh, L. A. Estrada-Jimenez, T. Pulikottil, S. N. Hojjati, and J. Barata, "The human role in human-centric industry," in *Proc. 48th Annu. Conf. IEEE Ind. Electron. Soc.*, Oct. 2022, pp. 1–6.
[11] M. Liu and J. Tang, "Audio and video bimodal emotion recognition in social networks based on improved AlexNet network and attention mechanism," *J. Inf. Process. Syst.*, vol. 17, no. 4, pp. 754–771, 2021.
[12] K. Bahreini, R. Nadolski, and W. Westera, "Towards multimodal emotion recognition in e-learning environments," *Interact. Learn. Environments*, vol. 24, no. 3, pp. 590–605, Apr. 2016.
[13] Y. Matsuda, D. Fedotov, Y. Takahashi, Y. Arakawa, K. Yasumoto, and W. Minker, "EmoTour: Multimodal emotion recognition using physiological and audio-visual features," in *Proc. ACM Int. Joint Conf. Int. Symp. Pervasive Ubiquitous Comput. Wearable Comput.*, Oct. 2018, pp. 946–951.
[14] D. DeVault et al., "SimSensei Kiosk: A virtual human interviewer for healthcare decision support," in *Proc. Int. Conf. Auto. Agents Multi-Agent Syst.*, 2014, pp. 1061–1068.
[15] K. H. Kim, S. W. Bang, and S. R. Kim, "Emotion recognition system using short-term monitoring of physiological signals," in *Medical and Biological Engineering and Computing*, vol. 42. Springer, 2004, pp. 419–427.
[16] N. Alswaidan and M. E. B. Menai, "A survey of state-of-the-art approaches for emotion recognition in text," *Knowl. Inf. Syst.*, vol. 62, no. 8, pp. 2937–2987, Aug. 2020.
[17] S. Zhao, G. Jia, J. Yang, G. Ding, and K. Keutzer, "Emotion recognition from multiple modalities: Fundamentals and methodologies," *IEEE Signal Process. Mag.*, vol. 38, no. 6, pp. 59–73, Nov. 2021.
[18] S. Pal, S. Mukhopadhyay, and N. Suryadevara, "Development and progress in sensors and technologies for human emotion recognition," *Sensors*, vol. 21, no. 16, p. 5554, Aug. 2021.
[19] U. A. Khan, Q. Xu, Y. Liu, A. Lagstedt, A. Alamäki, and J. Kauttonen, "Exploring contactless techniques in multimodal emotion recognition: Insights into diverse applications, challenges, solutions, and prospects," *Multimedia Syst.*, vol. 30, no. 3, pp. 1–48, Jun. 2024.
[20] S. Zhang, Y. Yang, C. Chen, X. Zhang, Q. Leng, and X. Zhao, "Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects," *Expert Syst. Appl.*, vol. 237, Mar. 2024, Art. no. 121692.
[21] N. Ahmed, Z. A. Aghbari, and S. Girija, "A systematic survey on multimodal emotion recognition using learning algorithms," *Intell. Syst. Appl.*, vol. 17, Feb. 2023, Art. no. 200171.
[22] B. Pan, K. Hirota, Z. Jia, and Y. Dai, "A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods," *Neurocomputing*, vol. 561, Dec. 2023, Art. no. 126866.
[23] S. K. Khare, V. Blanes-Vidal, E. S. Nadimi, and U. R. Acharya, "Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations," *Inf. Fusion*, vol. 102, Feb. 2024, Art. no. 102019.
[24] M. F. H. Siddiqui, P. Dhakal, X. Yang, and A. Y. Javaid, "A survey on databases for multimodal emotion recognition and an introduction to the VIRI (Visible and InfraRed Image) database," *Multimodal Technol. Interact.*, vol. 6, no. 6, p. 47, Jun. 2022.
[25] X. Gu, Y. Shen, and J. Xu, "Multimodal emotion recognition in deep learning: A survey," in *Proc. Int. Conf. Culture-Oriented Sci. Technol. (ICCST)*, Nov. 2021, pp. 77–82.
[26] P. Koromilas and T. Giannakopoulos, "Deep multimodal emotion recognition on human speech: A review," *Appl. Sci.*, vol. 11, no. 17, p. 7962, Aug. 2021.
[27] D. Moher, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *Ann. Internal Med.*, vol. 151, no. 4, p. 264, Aug. 2009.
[28] G. Hasson, *Understanding Emotional Intelligence* (International Series of Monographs on Physics). London, U.K.: Pearson Business, 2015.
[29] A. Scarantino and R. De Sousa, "Emotion," in *Stanford Encyclopedia of Philosophy*. Stanford, CA, USA: Stanford Univ., 2018.
[30] B. Fehr and J. A. Russell, "Concept of emotion viewed from a prototype perspective," *J. Experim. Psychol., Gen.*, vol. 113, no. 3, pp. 464–486, 1984.
[31] P. R. Kleinginna and A. M. Kleinginna, "A categorized list of emotion definitions, with suggestions for a consensual definition," *Motivat. Emotion*, vol. 5, no. 4, pp. 345–379, Dec. 1981.
[32] E. T. Rolls, *Emotion Explained* (Affective Science). London, U.K.: Oxford Univ. Press, 2005.
[33] A. Barnes and P. Thagard, "Emotional decisions," in *Proc. 18th Annu. Conf. Cogn. Sci. Soc.* Evanston, IL, USA: Routledge, 2019, pp. 426–429.
[34] J. S. Lerner, Y. Li, P. Valdesolo, and K. S. Kassam, "Emotion and decision making," *Annu. Rev. Psychol.*, vol. 66, pp. 799–823, Jan. 2015.

[35] R. Gil, J. Virgili-Gomá, R. García, and C. Mason, "Emotions ontology for collaborative modelling and learning of emotional responses," *Comput. Hum. Behav.*, vol. 51, pp. 610–617, Oct. 2015.

[36] W. B. Cannon, "The James–Lange theory of emotions: A critical examination and an alternative theory," *Amer. J. Psychol.*, vol. 39, nos. 1–4, p. 106, Dec. 1927.

[37] P. S. Sreeja and G. Mahalakshmi, "Emotion models: A review," *Int. J. Control Theory Appl.*, vol. 10, no. 8, pp. 651–657, 2017.

[38] N. K. Denzin, *On Understanding Emotion*. New Brunswick, NJ, USA: Transaction Publishers, 1984.

[39] A. Moors, "Theories of emotion causation: A review," *Cognition Emotion*, vol. 23, no. 4, pp. 625–662, Jun. 2009.

[40] R. Reisenzein, "The Schachter theory of emotion: Two decades later," *Psychol. Bull.*, vol. 94, no. 2, pp. 239–264, 1983.

[41] R. S. Lazarus, *Emotion and Adaptation*. London, U.K.: Oxford Univ. Press, 1991.

[42] T. C. Brickhouse and N. D. Smith, "Socrates on the emotions," *Plato J.*, vol. 15, pp. 9–28, Dec. 2015.

[43] D. R. Heise, *Understanding Events: Affect and the Construction of Social Action*. Cambridge, U.K.: Cambridge Univ. Press, 1979.

[44] E. Shouse, "Feeling, emotion, affect," *M/C J.*, vol. 8, no. 6, pp. 1–11, Dec. 2005.

[45] P. Ekkekakis, *Affect, Mood, and Emotion*. Human Kinetics, 2012.

[46] J. E. Stets, "Emotions and sentiments," in *Handbook of Social Psychology*. Boston, MA, USA: Springer, 2003, pp. 309–335.

[47] S. L. Gordon, "The sociology of sentiments and emotion," in *Social Psychology*. Evanston, IL, USA: Routledge, 2017, pp. 562–592.

[48] E. Bliss-Moreau, L. A. Williams, and A. C. Santistevan, "The immutability of valence and arousal in the foundation of emotion," *Emotion*, vol. 20, no. 6, p. 993, 2020.

[49] P. E. G. Bestelmeyer, S. A. Kotz, and P. Belin, "Effects of emotional valence and arousal on the voice perception network," *Social Cogn. Affect. Neurosci.*, vol. 12, no. 8, pp. 1351–1358, Aug. 2017.

[50] R. Plutchik and H. Kellerman, *Theories of Emotion*, vol. 1. New York, NY, USA: Academic, 2013.

[51] G. F. Wilson and C. A. Russell, "Real-time assessment of mental workload using psychophysiological measures and artificial neural networks," *Human Factors: J. Human Factors Ergonom. Soc.*, vol. 45, no. 4, pp. 635–644, Dec. 2003.

[52] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychol.*, vol. 14, no. 4, pp. 261–292, Dec. 1996.

[53] D. Keltner, D. Sauter, J. Tracy, and A. Cowen, "Emotional expression: Advances in basic emotion theory," *J. Nonverbal Behav.*, vol. 43, no. 2, pp. 133–160, Jun. 2019.

[54] A. F. Shariff and J. L. Tracy, "What are emotion expressions for?" *Current Directions Psychol. Sci.*, vol. 20, no. 6, pp. 395–399, 2011.

[55] P. Winkielman and K. C. Berridge, "Unconscious emotion," *Current directions Psychol. Sci.*, vol. 13, no. 3, pp. 120–123, 2004.

[56] J. J. Gross, "Emotion regulation," in *Handbook Emotions*, vol. 3. New York, NY, USA: Guilford Press, 2008, pp. 497–513.

[57] R. A. Thompson, "Emotional regulation and emotional development," *Educ. Psychol. Rev.*, vol. 3, no. 4, pp. 269–307, Dec. 1991.

[58] I. B. Mauss, S. A. Bunge, and J. J. Gross, "Automatic emotion regulation," *Social Personality Psychol. Compass*, vol. 1, no. 1, pp. 146–167, 2007.

[59] H. Gruber and R. Oepen, "Emotion regulation strategies and effects in art-making: A narrative synthesis," *Arts Psychotherapy*, vol. 59, pp. 65–74, Jul. 2018.

[60] E. L. Jurist, "Art and emotion in psychoanalysis," *Int. J. Psychoanalysis*, vol. 87, no. 5, pp. 1315–1334, Oct. 2006.

[61] S. Freud, *Writings on Art and Literature*. Stanford, CA, USA: Stanford Univ. Press, 1997.

[62] J. Plamper, *The History of Emotions: An Introduction*. Oxford, U.K.: OUP Oxford, 2015.

[63] U. Hess and P. Thibault, "Darwin and emotion expression," *Amer. Psychologist*, vol. 64, no. 2, p. 120, 2009.

[64] P. Ekman, "Expression and the nature of emotion," *Approaches Emotion*, vol. 3, no. 19, p. 344, 1984.

[65] D. Keltner, *Born To Be Good: The Science of a Meaningful Life*. New York, NY, USA: WW Norton, 2009.

[66] A. Mehrabian, "Communication without words," in *Communication Theory*. Evanston, IL, USA: Routledge, 2017, pp. 193–200.

[67] S. K. Langer, *Feeling and Form*, vol. 3. Evanston, IL, USA: Routledge, 1953.

[68] G. Collier and G. J. Collier, *Emotional Expression*. London, U.K.: Psychology Press, 2014.

[69] J. Hospers, "The concept of artistic expression," *Proc. Aristotelian Soc.*, vol. 55, pp. 313–344, Jun. 1954.

[70] D. Matravers, "Art, expression and emotion," in *The Routledge Companion to Aesthetics*. Evanston, IL, USA: Routledge, 2013, pp. 404–414.

[71] O. Klineberg, "Emotional expression in Chinese literature," *J. Abnormal Social Psychol.*, vol. 33, no. 4, p. 517, 1938.

[72] M. W. Morris and D. Keltner, "How emotions work: The social functions of emotional expression in negotiations," *Res. Organizational Behav.*, vol. 22, pp. 1–50, Jan. 2000.

[73] A. Gabrielsson and P. N. Juslin, *Emotional Expression in Music*. London, U.K.: Oxford Univ. Press, 2003.

[74] R.-L. Kortesluoma, R.-L. Punamäki, and M. Nikkonen, "Hospitalized children drawing their pain: The contents and cognitive and emotional characteristics of pain drawings," *J. Child Health Care*, vol. 12, no. 4, pp. 284–300, Dec. 2008.

[75] A. S. Winston, B. Kenyon, J. Stewardson, and T. Lepine, "Children's sensitivity to expression of emotion in drawings," *Vis. Arts Res.*, vol. 21, no. 1, pp. 1–14, 1995.

[76] T. Van Gorp and E. Adams, *Design for Emotion*. Amsterdam, The Netherlands: Elsevier, 2012.

[77] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, p. 2074, Jun. 2018.

[78] S. C. Roberts, J. T. Fialová, A. Sorokowska, B. Langford, P. Sorokowski, V. Třebický, and J. Havlíček, "Emotional expression in human odour," *Evol. Human Sci.*, vol. 4, p. e44, Oct. 2022.

[79] P. Ekman, "Facial expression and emotion," *Amer. Psychol.*, vol. 48, no. 4, p. 384, 1993.

[80] N. Dael, M. Mortillaro, and K. R. Scherer, "Emotion expression in body action and posture," *Emotion*, vol. 12, no. 5, p. 1085, 2012.

[81] K. R. Scherer, T. Johnstone, and G. Klasmeyer, "Vocal expression of emotion," in *Handbook of Affective Sciences*. Oxford Univ. Press, 2002, pp. 433–456.

[82] J. Zhu, L. Ji, and C. Liu, "Heart rate variability monitoring for emotion and disorders of emotion," *Physiological Meas.*, vol. 40, no. 6, Jul. 2019, Art. no. 064004.

[83] J. R. Gray, "Integration of emotion and cognitive control," *Current Directions Psychol. Sci.*, vol. 13, no. 2, pp. 46–48, Apr. 2004.

[84] G. Deliens, M. Gilson, and P. Peigneux, "Sleep and the processing of emotions," *Exp. Brain Res.*, vol. 232, no. 5, pp. 1403–1414, May 2014.

[85] B. Metternich, N. A. Gehrer, K. Wagner, M. J. Geiger, E. Schütz, A. Schulze-Bonhage, M. Heers, and M. Schönenberg, "Eye-movement patterns during emotion recognition in focal epilepsy: An exploratory investigation," *Seizure, Eur. J. Epilepsy*, vol. 100, pp. 95–102, Aug. 2022.

[86] M. W. Schurgin, J. Nelson, S. Iida, H. Ohira, J. Y. Chiao, and S. L. Franconeri, "Eye movements during emotion recognition in faces," *J. Vis.*, vol. 14, no. 13, p. 14, Nov. 2014.

[87] R. Jerath and C. Beveridge, "Respiratory rhythm, autonomic modulation, and the spectrum of emotions: The future of emotion recognition and modulation," *Frontiers Psychol.*, vol. 11, Aug. 2020, Art. no. 555957.

[88] P. Ekman and W. V. Friesen, "Facial action coding system," in *Environmental Psychology & Nonverbal Behavior*. APA PsycTests, 1978.

[89] V. Sacharin, K. Schlegel, and K. R. Scherer, "Geneva emotion wheel rating study," Dept. Affective Sci.s, Center for Person, Kommunikation, Aalborg Univ., Aalborg, Denmark, 2012.

[90] K. R. Scherer, "What are emotions? And how can they be measured?" *Social Sci. Inf.*, vol. 44, no. 4, pp. 695–729, Dec. 2005.

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

[92] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: The PANAS scales," *J. Personality Social Psychol.*, vol. 54, no. 6, pp. 1063–1070, 1988.

[93] A. T. Beck, C. Ward, M. Mendelson, J. Mock, and J. Erbaugh, "Beck depression inventory (BDI)," *Arch. Gen. Psychiatry*, vol. 4, no. 6, pp. 561–571, 1961.

[94] T. M. Marteau and H. Bekker, "The development of a six-item short-form of the state scale of the spielberger state—Trait anxiety inventory (STAI)," *Brit. J. Clin. Psychol.*, vol. 31, no. 3, pp. 301–306, Sep. 1992.

[95] D. McNair, M. Lorr, and L. Droppleman, "Profile of mood states manual," in *Educational and Industrial Testing Services*, San Diego, CA, USA, 1971.

[96] O. K. Akputu, K. P. Seng, and Y. L. Lee, "Affect recognition for Web 2.0 intelligent E-tutoring systems: Exploration of Students' emotional feedback," in *Student Engagement and Participation: Concepts, Methodologies, Tools, and Application*. Hershey, PA, USA: IGI Global, 2013, pp. 818–848.
[97] J. M. Saragih, S. Lucey, and J. F. Cohn, "Face alignment through subspace constrained mean-shifts," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 1034–1041.
[98] L. C. De Silva and P. C. Ng, "Bimodal emotion recognition," in *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Oct. 2000, pp. 332–335.
[99] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner, "Recognizing engagement in human–robot interaction," in *Proc. 5th ACM/IEEE Int. Conf. Human-Robot Interact. (HRI)*, Mar. 2010, pp. 375–382.
[100] F. D. Davis, "Technology acceptance model: TAM," in *Information Seeking Behavior and Technology Adoption*, vol. 205, M. N. Al-Suqri and A. S. Al-Aufi, Eds., Ann Arbor, MI, USA: Univ. of Michigan Press, 1989, p. 219.
[101] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS Quart.*, vol. 27, no. 3, pp. 425–478, 2003.
[102] H. Oinas-Kukkonen and M. Harjumaa, "Persuasive systems design: Key issues, process model, and system features," *Commun. Assoc. for Inf. Syst.*, vol. 24, p. 28, 2009.
[103] N. M. Safdar, J. D. Banja, and C. C. Meltzer, "Ethical considerations in artificial intelligence," *Eur. J. Radiol.*, vol. 122, Jan. 2020, Art. no. 108768.
[104] A. Haleem, M. Javaid, M. Asim Qadri, R. Pratap Singh, and R. Suman, "Artificial intelligence (AI) applications for marketing: A literature-based study," *Int. J. Intell. Netw.*, vol. 3, pp. 119–132, Jan. 2022.
[105] A. Katirai, "Ethical considerations in emotion recognition technologies: A review of the literature," *AI Ethics*, pp. 1–22, Jun. 2023, doi: 10.1007/s43681-023-00307-3.
[106] S. M. Mohammad, "Ethics sheets for AI tasks," 2021, *arXiv:2107.01183*.
[107] P. Korytkowski and E. Kulczycki, "Publication counting methods for a national research evaluation exercise," *J. Informetrics*, vol. 13, no. 3, pp. 804–816, Aug. 2019.
[108] P. Yaffe, "The 7% rule: Fact, fiction, or misunderstanding," *Ubiquity*, vol. 2011, pp. 1–5, Oct. 2011.
[109] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008.
[110] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," 2016, *arXiv:1606.06259*.
[111] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.
[112] H. Cai et al., "A multi-modal open dataset for mental-disorder analysis," *Sci. Data*, vol. 9, no. 1, p. 178, Apr. 2022.
[113] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, and L.-P. Morency, "The distress analysis interview corpus of human and computer interviews," in *Proc. Int. Conf. Lang. Resour. Eval.*, 2014, pp. 3123–3128.
[114] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, S. Song, S. Liu, Z. Zhao, A. Mallol-Ragolta, Z. Ren, M. Soleymani, and M. Pantic, "AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition," in *Proc. 9th Int. Audio/Visual Emotion Challenge Workshop*, Oct. 2019, pp. 3–12.
[115] J. Yoon, C. Kang, S. Kim, and J. Han, "D-Vlog: Multimodal vlog dataset for depression detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 11, 2022, pp. 12226–12234.
[116] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee, "K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations," *Sci. Data*, vol. 7, no. 1, p. 293, Sep. 2020.
[117] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auto. Mental Develop.*, vol. 7, no. 3, pp. 162–175, Sep. 2015.
[118] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *Proc. 6th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, Nov. 2013, pp. 81–84.
[119] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "EmotionMeter: A multimodal framework for recognizing human emotions," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, Mar. 2019.
[120] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," 2018, *arXiv:1810.02508*.
[121] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0196391.
[122] M. Firdaus, H. Chauhan, A. Ekbal, and P. Bhattacharyya, "MEISD: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 4441–4453.
[123] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Acted facial expressions in the wild database," Res. School Comput. Sci., College Eng. Comput. Sci., Austral. Nat. Univ., Canberra, NSW, Australia, Tech. Rep. TR-CS-11, 2011, vol. 2, no. 1.
[124] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 2106–2112.
[125] D. Kollias, M. A. Nicolaou, I. Kotsia, G. Zhao, and S. Zafeiriou, "Recognition of affect in the wild using deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1972–1979.
[126] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.
[127] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, "Towards multimodal sarcasm detection (An _obviously_perfect paper)," 2019, *arXiv:1906.01815*.
[128] P. Jackson and S. Haq, *Surrey Audio-Visual Expressed Emotion (SAVEE) Database*. Guildford, U.K.: Univ. Surrey, 2014.
[129] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, Sep. 2005, pp. 1517–1520.
[130] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, Oct. 2014.
[131] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan. 2019.
[132] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "ASCERTAIN: Emotion and personality recognition using commercial sensors," *IEEE Trans. Affect. Comput.*, vol. 9, no. 2, pp. 147–160, Apr. 2018.
[133] B. Yang, J. Wu, and G. Hattori, "Facial expression recognition with the advent of face masks," in *Proc. 19th Int. Conf. Mobile Ubiquitous Multimedia*, Nov. 2020, pp. 335–337.
[134] Y. Cimtay, E. Ekmekcioglu, and S. Caglar-Ozhan, "Cross-subject multimodal emotion recognition based on hybrid fusion," *IEEE Access*, vol. 8, pp. 168865–168878, 2020.
[135] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis; Using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012.
[136] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "AMIGOS: A dataset for affect, personality and mood research on individuals and groups," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 479–493, Apr. 2021.
[137] B. Schuller, M. Valstar, R. Cowie, and M. Pantic, "AVEC 2012: The continuous audio/visual emotion challenge—An introduction," in *Proc. 14th ACM Int. Conf. Multimodal Interact.*, Oct. 2012, pp. 449–456.
[138] A. Dhall, J. Joshi, I. Radwan, and R. Goecke, "Finding happiest moments in a social context," in *Proc. 11th Asian Conf. Comput. Vis.*, Daejeon, South Korea. Cham, Switzerland: Springer, 2012, pp. 613–626.
[139] D. Kang, D. Kim, D. Kang, T. Kim, B. Lee, D. Kim, and B. C. Song, "Beyond superficial emotion recognition: Modality-adaptive emotion recognition system," *Expert Syst. Appl.*, vol. 235, Jan. 2024, Art. no. 121097.

[140] M. S. U. Miah, M. M. Kabir, T. B. Sarwar, M. Safran, S. Alfarhood, and M. F. Mridha, "A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM," *Sci. Rep.*, vol. 14, no. 1, p. 9603, Apr. 2024.

[141] R. Selvi and C. Vijayakumaran, "An efficient multimodal emotion identification using FOX optimized double deep Q-Learning," *Wireless Pers. Commun.*, vol. 132, no. 4, pp. 2387–2406, Oct. 2023.

[142] T. Zhang, S. Li, B. Chen, H. Yuan, and C. L. P. Chen, "AIA-Net: Adaptive interactive attention network for text–Audio emotion recognition," *IEEE Trans. Cybern.*, vol. 53, no. 12, pp. 7659–7671, Dec. 2023.

[143] B. Mocanu, R. Tapu, and T. Zaharia, "Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning," *Image Vis. Comput.*, vol. 133, May 2023, Art. no. 104676.

[144] W. Wang, H. Zhang, and Z. Zhang, "Research on emotion recognition method of flight training based on multimodal fusion," *Int. J. Hum.-Comput. Interact.*, pp. 1–14, Sep. 2023.

[145] T. Zhang, Z. Tan, and X. Wu, "HAAN-ERC: Hierarchical adaptive attention network for multimodal emotion recognition in conversation," *Neural Comput. Appl.*, vol. 35, no. 24, pp. 17619–17632, Aug. 2023.

[146] H. M. Shahzad, S. M. Bhatti, A. Jaffar, M. Rashid, and S. Akram, "Multi-modal CNN features fusion for emotion recognition: A modified xception model," *IEEE Access*, vol. 11, pp. 94281–94289, 2023.

[147] F. M. Alamgir and M. S. Alam, "Hybrid multi-modal emotion recognition framework based on InceptionV3DenseNet," *Multimedia Tools Appl.*, vol. 82, no. 26, pp. 40375–40402, Nov. 2023.

[148] A. Aguilera, D. Mellado, and F. Rojas, "An assessment of in-the-wild datasets for multimodal emotion recognition," *Sensors*, vol. 23, no. 11, p. 5184, May 2023.

[149] R. A. Jaswal and S. Dhingra, "Empirical analysis of multiple modalities for emotion recognition using convolutional neural network," *Meas. Sensors*, vol. 26, Apr. 2023, Art. no. 100716.

[150] W. Zheng, L. Yan, and F.-Y. Wang, "Two birds with one stone: Knowledge-embedded temporal convolutional transformer for depression detection and emotion recognition," *IEEE Trans. Affect. Comput.*, 2023.

[151] L. Feng, L.-Y. Liu, S.-L. Liu, J. Zhou, H.-Q. Yang, and J. Yang, "Multimodal speech emotion recognition based on multi-scale MFCCs and multi-view attention mechanism," *Multimedia Tools Appl.*, vol. 82, no. 19, pp. 28917–28935, Aug. 2023.

[152] Y. Zhang, J. Wang, Y. Liu, L. Rong, Q. Zheng, D. Song, P. Tiwari, and J. Qin, "A multitask learning model for multimodal sarcasm, sentiment and emotion recognition in conversations," *Inf. Fusion*, vol. 93, pp. 282–301, May 2023.

[153] H.-D. Le, G.-S. Lee, S.-H. Kim, S. Kim, and H.-J. Yang, "Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning," *IEEE Access*, vol. 11, pp. 14742–14751, 2023.

[154] L. Chen, M. Li, M. Wu, W. Pedrycz, and K. Hirota, "Coupled multimodal emotional feature analysis based on broad-deep fusion networks in human–robot interaction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 7, pp. 9663–9673, Jul. 2024.

[155] Q. Rong, S. Ding, Z. Yue, Y. Wang, L. Wang, X. Zheng, and Y. Li, "Non-contact negative mood state detection using reliability-focused multi-modal fusion model," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 9, pp. 4691–4701, Sep. 2022.

[156] A. Menon, A. Natarajan, R. Agashe, D. Sun, M. Aristio, H. Liew, Y. S. Shao, and J. M. Rabaey, "Efficient emotion recognition using hyperdimensional computing with combinatorial channel encoding and cellular automata," *Brain Informat.*, vol. 9, no. 1, p. 14, Dec. 2022.

[157] A. K. Marandi, G. Jethava, A. Rajesh, S. Gupta, S. Sagar, and S. Sharma, "Cognitive computing for multimodal sentiment sensing and emotion recognition fusion based on machine learning techniques implemented by computer interface system," *Int. J. Commun. Netw. Inf. Secur. (IJCNIS)*, vol. 14, no. 2, pp. 15–32, Aug. 2022.

[158] A. Joshi, A. Bhat, A. Jain, A. Singh, and A. Modi, "COGMEN: Contextualized GNN based multimodal emotion recognition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2022, pp. 4148–4164.

[159] H. Sun, H. Wang, J. Liu, Y.-W. Chen, and L. Lin, "CubeMLP: An MLP-based model for multimodal sentiment analysis and depression estimation," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 3722–3729.

[160] Y. Zhang, P. Tiwari, L. Rong, R. Chen, N. A. Alnajem, and M. S. Hossain, "Affective interaction: Attentive representation learning for multi-modal sentiment classification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 18, no. 3s, pp. 1–23, Oct. 2022.

[161] F. Ma, Y. Li, S. Ni, S.-L. Huang, and L. Zhang, "Data augmentation for audio-visual emotion recognition with an efficient multimodal conditional GAN," *Appl. Sci.*, vol. 12, no. 1, p. 527, Jan. 2022.

[162] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, "Attention-based multi-modal sentiment analysis and emotion detection in conversation using RNN," *Int. J. Interact. Multimedia Artif. Intell.*, vol. 6, no. 6, p. 112, 2021.

[163] J. Wu, T. Dang, V. Sethu, and E. Ambikairajah, "Multimodal affect models: An investigation of relative salience of audio and visual cues for emotion prediction," *Frontiers Comput. Sci.*, vol. 3, Dec. 2021, Art. no. 767767.

[164] X. Huang, M. Ren, Q. Han, X. Shi, J. Nie, W. Nie, and A.-A. Liu, "Emotion detection for conversations based on reinforcement learning framework," *IEEE MultimediaMag.*, vol. 28, no. 2, pp. 76–85, Apr. 2021.

[165] M. Ren, X. Huang, X. Shi, and W. Nie, "Interactive multimodal attention network for emotion recognition in conversation," *IEEE Signal Process. Lett.*, vol. 28, pp. 1046–1050, 2021.

[166] M. Li, X. Qiu, S. Peng, L. Tang, Q. Li, W. Yang, and Y. Ma, "Multimodal emotion recognition model based on a deep neural network with multiobjective optimization," *Wireless Commun. Mobile Comput.*, vol. 2021, pp. 1–10, Aug. 2021.

[167] R.-H. Huan, J. Shu, S.-L. Bao, R.-H. Liang, P. Chen, and K.-K. Chi, "Video multimodal emotion recognition based on bi-GRU and attention fusion," *Multimedia Tools Appl.*, vol. 80, no. 6, pp. 8213–8240, Mar. 2021.

[168] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, and S. Nanayakkara, "Multimodal emotion recognition with transformer-based self supervised feature fusion," *IEEE Access*, vol. 8, pp. 176274–176285, 2020.

[169] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 1359–1367.

[170] L. Cai, Y. Hu, J. Dong, and S. Zhou, "Audio-textual emotion recognition based on improved neural networks," *Math. Problems Eng.*, vol. 2019, pp. 1–9, Dec. 2019.

[171] X. Huang, A. Dhall, R. Goecke, M. Pietikäinen, and G. Zhao, "Multimodal framework for analyzing the affect of a group of people," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2706–2721, Oct. 2018.

[172] R. V. Darekar and A. P. Dhande, "Emotion recognition from Marathi speech database using adaptive artificial neural network," *Biologically Inspired Cogn. Architectures*, vol. 23, pp. 35–42, Jan. 2018.

[173] V. P. Gonçalves, E. P. Costa, A. Valejo, G. P. R. Filho, T. M. Johnson, G. Pessin, and J. Ueyama, "Enhancing intelligence in multimodal emotion assessments," *Int. J. Speech Technol.*, vol. 46, no. 2, pp. 470–486, Mar. 2017.

[174] L.-A. Perez-Gaspar, S.-O. Caballero-Morales, and F. Trujillo-Romero, "Multimodal emotion recognition with evolutionary computation for human–robot interaction," *Expert Syst. Appl.*, vol. 66, pp. 42–61, Dec. 2016.

[175] J. C. Kim and M. A. Clements, "Multimodal affect classification at various temporal lengths," *IEEE Trans. Affect. Comput.*, vol. 6, no. 4, pp. 371–384, Oct. 2015.

[176] S. Haq, T. Jan, A. Jehangir, M. Asif, A. Ali, and N. Ahmad, "Bimodal human emotion classification in the speaker-dependent scenario," *Pakistan Acad. Sci.*, vol. 52, no. 1, pp. 27–38, 2015.

[177] A. Savran, H. Cao, A. Nenkova, and R. Verma, "Temporal Bayesian fusion for affect sensing: Combining video, audio, and lexical modalities," *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1927–1941, Sep. 2015.

[178] W.-L. Zheng, B.-N. Dong, and B.-L. Lu, "Multimodal emotion recognition using EEG and eye tracking data," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2014, pp. 5040–5043.

[179] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L.-P. Morency, "Multi-level multiple attentions for contextual multimodal sentiment analysis," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 1033–1038.

[180] L. Zhang, "Ideological and political empowering English teaching: Ideological education based on artificial intelligence in classroom emotion recognition," *Int. J. Comput. Appl. Technol.*, vol. 71, no. 3, pp. 265–271, 2023.

[181] Y.-G. Wu, M.-H. Yang, W.-J. Chen, Z.-Y. Lin, and K.-S. Wu, "Emotion detection by image analysis for painter," in *Proc. World Multi-Conf. Systemics, Cybern. Informat.*, Jul. 2022, pp. 1–6.
[182] T. Pan, X. Zhao, B. Liu, and W. Liu, "Automated drawing psychoanalysis via house-tree-person test," in *Proc. IEEE 34th Int. Conf. Tools Artif. Intell. (ICTAI)*, Oct. 2022, pp. 1120–1125.
[183] M. Chen, Y. Jiang, Y. Cao, and A. Y. Zomaya, "CreativeBioMan: A brain- and body-wearable, computing-based, creative gaming system," *IEEE Syst., Man, Cybern. Mag.*, vol. 6, no. 1, pp. 14–22, Jan. 2020.
[184] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.
[185] L. Mescheder, S. Nowozin, and A. Geiger, "Adversarial variational Bayes: Unifying variational autoencoders and generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2391–2400.
[186] D. Saxena and J. Cao, "Generative adversarial networks (GANs): Challenges, solutions, and future directions," *ACM Comput. Surv.*, vol. 54, no. 3, pp. 1–42, Apr. 2022.
[187] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.
[188] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[189] J. Zhao, R. Li, Q. Jin, X. Wang, and H. Li, "Memobert: Pre-training model with prompt-based learning for multimodal emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 4703–4707.
[190] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," in *Proc. OTM Confederated Int. Conf., Move Meaningful Internet Syst.*, Catania, Italy. Cham, Switzerland: Springer, Nov. 2003, pp. 986–996.
[191] D. Meyer and F. Wien, "Support vector machines," *R News*, vol. 1, no. 3, pp. 23–26, 2001.
[192] I. Steinwart and A. Christmann, *Support Vector Machines*. New York, NY, USA: Springer, 2008.
[193] L. Rokach and O. Maimon, "Decision trees," in *Data Mining and Knowledge Discovery Handbook*. New York, NY, USA: Springer, 2005, pp. 165–192.
[194] C. Kingsford and S. L. Salzberg, "What are decision trees?" *Nature Biotechnol.*, vol. 26, no. 9, pp. 1011–1013, Sep. 2008.
[195] B. De Ville, "Decision trees," *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 5, no. 6, pp. 448–455, 2013.
[196] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," in *Ensemble Machine Learning: Methods and Applications*. Springer, 2012, pp. 157–175.
[197] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.
[198] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, R. Tibshirani, and J. Friedman, "Random forests," in *The Elements of Statistical learning: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer, 2009, pp. 587–604.
[199] Y. L. Pavlov, "Random forests," in *Probabilistic Methods in Discrete Mathematics (Petrozavodsk, 1996)*. Berlin, Germany: De Gruyter, 1997, pp. 11–18.
[200] S. Agatonovic-Kustrin and R. Beresford, "Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research," *J. Pharmaceutical Biomed. Anal.*, vol. 22, no. 5, pp. 717–727, Jun. 2000.
[201] L. O. Chua and T. Roska, "The CNN paradigm," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 40, no. 3, pp. 147–156, Mar. 1993.
[202] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Phys. D, Nonlinear Phenomena*, vol. 404, Mar. 2020, Art. no. 132306.
[203] J. Koutnik, K. Greff, F. Gomez, and J. Schmidhuber, "A clockwork RNN," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1863–1871.
[204] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 5, 2020, pp. 8992–8999.
[205] Y.-H. H. Tsai, P. Pu Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," 2018, *arXiv:1806.06176*.
[206] Y.-H.-H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, p. 6558.
[207] Z. Pan, Z. Luo, J. Yang, and H. Li, "Multi-modal attention for speech emotion recognition," 2020, *arXiv:2009.04107*.
[208] Z. Wang, Z. Wan, and X. Wan, "TransModality: An End2End fusion method with transformer for multimodal sentiment analysis," in *Proc. Web Conf.*, Apr. 2020, pp. 2514–2520.
[209] J. Zheng, S. Zhang, Z. Wang, X. Wang, and Z. Zeng, "Multi-channel weight-sharing autoencoder based on cascade multi-head attention for multimodal emotion recognition," *IEEE Trans. Multimedia*, vol. 25, pp. 2213–2225, 2022.
[210] S. Mai, Y. Zeng, S. Zheng, and H. Hu, "Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2276–2289, Jul. 2023.
[211] N. Hajarolasvadi, M. A. Ramírez, W. Beccaro, and H. Demirel, "Generative adversarial networks in human emotion synthesis: A review," *IEEE Access*, vol. 8, pp. 218499–218529, 2020.
[212] Y. Luo, L.-Z. Zhu, and B.-L. Lu, "A GAN-based data augmentation method for multimodal emotion recognition," in *Proc. Int. Symp. Neural Netw.*, Moscow, Russia. Cham, Switzerland: Springer, Jul. 2019, pp. 141–150.
[213] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, vol. 1, 2019, p. 2.
[214] A. Dhall, O. V. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: EmotiW 2015," in *Proc. ACM Int. Conf. Multimodal Interact.*, Nov. 2015, pp. 423–426.
[215] W. Boucsein, *Electrodermal Activity*. New York, NY, USA: Springer, 2012.
[216] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer, *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford, U.K.: OUP Oxford, 2011.
[217] A. Kendon, *Gesture: Visible Action As Utterance*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[218] H. Wang, J. Zhang, Y. Huang, and B. Cai, "FBANet: Transfer learning for depression recognition using a feature-enhanced bi-level attention network," *Entropy*, vol. 25, no. 9, p. 1350, Sep. 2023.
[219] J. Zhang, Y. Yu, V. Barra, X. Ruan, Y. Chen, and B. Cai, "Feasibility study on using house-tree-person drawings for automatic analysis of depression," *Comput. Methods Biomechanics Biomed. Eng.*, vol. 27, no. 9, pp. 1129–1140, Jul. 2024.
[220] S. Bird, E. Klein, and E. Loper, *Natural Language Processing With Python: Analyzing Text With the Natural Language Toolkit*. U.K.: O'Reilly Media, 2009.
[221] P. Xu, C. K. Joshi, and X. Bresson, "Multigraph transformer for free-hand sketch recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 5150–5161, Oct. 2022.
[222] S. Mozziconacci, "Prosody and emotions," in *Proc. Speech Prosody*, Apr. 2002, pp. 1–9.
[223] B. M. Appelhans and L. J. Luecken, "Heart rate variability as an index of regulated emotional responding," *Rev. Gen. Psychol.*, vol. 10, no. 3, pp. 229–240, Sep. 2006.
[224] D. Ayata, Y. Yaslan, and M. Kamaşak, "Emotion recognition via galvanic skin response: Comparison of machine learning algorithms and feature extraction methods," *IU-J. Electr. Electron. Eng.*, vol. 17, no. 1, pp. 3147–3156, 2017.
[225] J. Z. Lim, J. Mountstephens, and J. Teo, "Emotion recognition using eye-tracking: Taxonomy, review and current challenges," *Sensors*, vol. 20, no. 8, p. 2384, Apr. 2020.
[226] F. Noroozi, C. A. Corneanu, D. Kaminska, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 505–523, Apr. 2021.
[227] N. Goodman, *Languages of Art: An Approach to a Theory of Symbols*. Indianapolis, IN, USA: Hackett, 1976.
[228] R. C. Burns, *Kinetic-House-Tree-Person Drawings (KHTP): An Interpretative Manual*. USA: American Psychological Association, 1987.
[229] F. R. Palmer, *Semantics*. Cambridge, U.K.: Cambridge Univ. Press, 1981.
[230] R. A. Stine, "Sentiment analysis," *Annu. Rev. Statist. Appl.*, vol. 6, pp. 287–308, Jan. 2019.

[231] S. Kalateh, L. A. Estrada-Jimenez, T. Pulikottil, S. N. Hojjati, and J. Barata, "Feeling smart industry," in *Proc. 62nd Int. Scientific Conf. Inf. Technol. Manage. Sci. Riga Tech. Univ. (ITMS)*, Oct. 2021, pp. 1–6.

[232] D. M. Jankowska, M. Czerwonka, I. Lebuda, and M. Karwowski, "Exploring the creative process: Integrating psychometric and eye-tracking approaches," *Frontiers Psychol.*, vol. 9, Oct. 2018, Art. no. 399258.

SEPIDEH KALATEH received the B.Sc. degree in electronic and electrical engineering and the M.Sc. degree in information technology management from IAU, Iran, in 2012 and 2017, respectively. She is currently pursuing the Ph.D. degree in electrical and computer engineering with the NOVA University of Lisbon, Portugal. She is also a Researcher with the CTS-UNINOVA Research Institute. With professional experiences in various industrial sectors alongside her academic background, including ICT and data management, her research interests include affective computing, computational creativity, and data science in the field of human–machine emotive interaction.

LUIS A. ESTRADA-JIMENEZ received the B.Sc. degree in electronic and control engineering from the Escuela Politecnica Nacional, Ecuador, in 2016, and the M.Sc. degree in mechatronics engineering from the University of Oviedo, Spain, in 2019. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the NOVA University of Lisbon, Portugal. His master's thesis was developed with the Department of Modular Automation, FESTO Company, Germany. He is also with the CTS-UNINOVA Research Institute. Currently, his main role is as an Early Stage Researcher with the Digital Manufacturing and Design Training Network (DIMAND) funded by European Union. His research interests include self-organization and automation in smart manufacturing systems and the application of artificial intelligence in industrial environments.

SANAZ NIKGHADAM-HOJJATI (Member, IEEE) received the Ph.D. degree in information technology management (business intelligence) from IAU, in 2017. She is currently a Senior Researcher with the CTS-UNINOVA Institute, NOVA University of Lisbon. She was a Postdoctoral Researcher with the Nova School of Science and Technology, NOVA University of Lisbon, from 2018 to 2019. Her research interests include computational creativity, affective computing, business intelligence, human behavior, emerging technologies, ICT, and innovation management. She has published several books and academic papers in a number of peer-reviewed journals and presented various academic papers at conferences. She has led and participated in several European Union Projects and Portuguese and Iranian National Projects. In addition, she was an University-Invited Professor. She is also the Director of the Women In Science, Technology, Engineering, and Mathematics (WoSTEM) Program, CTS-UNINOVA.

JOSE BARATA (Member, IEEE) received the Ph.D. degree in robotics and integrated manufacturing from the NOVA University of Lisbon, in 2004. He is currently a Professor with the Department of Electrical Engineering, NOVA University of Lisbon, and a Senior Researcher with the UNINOVA—Instituto de Desenvolvimento de Novas Tecnologias. He has participated in more than 15 international research projects involving different programs, including NMP, IST, ITEA, and ESPRIT. Since 2004, he has been leading the CTS-UNINOVA participation in EU projects, namely, EUPASS, self-learning, IDEAS, PRIME, RIVERWATCH, ROBO-PARTNER, and PROSECO. In the last years, he has participated actively in researching SOA-based approaches for the implementation of intelligent manufacturing devices, such as within the Inlife Project. He has authored or co-authored over 100 original papers in international journals and international conferences. His main research interest includes intelligent manufacturing, with an emphasis on complex adaptive systems, involving intelligent manufacturing devices. He is a member of the IEEE Technical Committee on Industrial Agents (IES), Self-Organization and Cybernetics for Informatics (SMC), and Education in Engineering and Industrial Technologies (IES).