



# Emotional Security for Sustainable Future: Multimodal Affective Computing

Nisha Rathi\*<sup>1</sup>, Parul Saran<sup>2</sup> and Satyam Shrivastava<sup>3</sup>

<sup>1,3</sup>Acropolis Institute of Technology and Research, Indore (M. P.) India

<sup>2</sup>Treasure Data, CA, USA

<sup>1</sup>nisharathi@acropolis.in

<sup>2</sup>parul.saran@treasure-data.co

<sup>3</sup>satyamshrivastava@acropolis.in

**Abstract.** Emotional well-being is quite much a significant human resource, which certainly boosts productivity, creativity as well as sustainability. As natural resources get conserved and sometimes preserved, emotional resilience would have to be preserved too against burnout and disaffection from society. One of the contributions by affective computing in this area is early detection of stress, personalized intervention, and even emotion-aware systems for emotional-social harmony among people. This goes in hand with the sustainability goals because it is actively supporting well-being as building blocks for a healthy sustainable future. Reading and comprehending emotions are complicated tasks, but technology aids us in this endeavor. Nowadays, sophisticated algorithms can extract and exploit aspects of body language to identify emotions from a wide range of data sources, such as images, videos, and biosignals. Scientists have been working on developing and analyzing techniques for automated emotion detection and recognition for decades. Extensive literature exists in the subject of emotion recognition, which suggests, evaluates, and estimates many methodologies in fields such as signal processing, machine learning, deep learning, computer vision, and speech recognition. Since the inception of affective computing, there are number of articles published on it. In this paper, we thoroughly examine cutting-edge fusion approaches as part of the review, and then critically evaluate possible performance gains from multimodal analysis over unimodal analysis. To help readers better grab this difficult and interesting study topic, a thorough narrative of these two complementing fields is presented.

**Keywords:** Affective computing, emotion recognition, multimodality, Deep Learning, Machine Learning.

## 1 The Scope of This Survey

This study examines affective computing and emotion recognition, emphasizing new developments in multimodal affect analysis. By offering a thorough summary of the most recent methods for combining textual, audio, and visual data—modalities that are most common in current research—it fills gaps in the study of existing literature [1]. The paper examines unimodal methods and assesses contemporary fusion techniques, emphasizing how they improve performance. The survey also addresses future directions in the discipline, providing new researchers with insightful information.

© The Author(s) 2025

S. Bhalariao et al. (eds.), *Proceedings of the International Conference on Recent Advancement and Modernization in Sustainable Intelligent Technologies & Applications (RAMSITA-2025)*, Advances in Intelligent Systems Research 192,

[https://doi.org/10.2991/978-94-6463-716-8\\_47](https://doi.org/10.2991/978-94-6463-716-8_47)

## 2 Introduction

Our thoughts are influenced by our emotions. When a person is emotionally healthy, he or she can perform at their best regardless of life's adversities. When emotional stress is severe and long-term, the risk of mental and physiological problems rises. As per the research published in the 'International Journal of Mental Health Systems' [2], there was 67.7% increase in suicide behaviour during India's COVID-19 shutdown.

### 2.1 Affective Computing Spread: When Machines Can Tell How You Feel

Affective computing covers human emotion, opinion, thoughts, sentiments and feelings [3], emotion identification, and sentiment analysis. The emotionless, clinical computer or robot is a staple of science fiction, but science reality is beginning to evolve; computers are improving their ability to recognize emotions. IBM's Watson group has developed a 'Tone Analyzer'[4] which can detect sarcasm and other emotions in your work. With advancements in research wearable computing, psychology, neuroscience, big data and modeling come under the domain of Affective Computing. The aim is to advance the knowledge, understanding, and development of systems to sense, identify, classify, and respond to human emotion [5, 6].

### 2.2 Emotion Models

Psychologists developed categorical and dimensional emotion models respectively [7] to characterize human emotion to recognize and compute emotion or sentiment. Six basic emotions i.e. happiness, surprise, fear, sorrow, disgust and anger are divided by categorical models into a set of distinct classifications that are straightforward to express. Dimensional models depict emotion as a point in multidimensional space, with valence, control and activation as dimensions. Valence-Arousal based four quadrant circumplex model (Figure 1(a)) by [8] captures complex emotions. The first quadrant has the feelings of happy emotions, the second depicts furious emotions, and the third depicts the sensations with sad emotions while the last one depicts peaceful emotions. Researchers have mostly relied on the simplified 2D model of arousal and valence. The Pleasure-Arousal-Dominance (PAD) paradigm [9] (Figure 1(b)) is a popular continuous multi-dimensional model to solve the limitations faced by discrete emotion models. The PAD model features three-dimensional spaces i.e. Mehrabian's 3-D space theory of emotion [10]. Plutchik [11] proposed the most popular component model, which is founded on evolutionary principles and comprises eight core bipolar emotions. The eight primary emotions like fear, joy, surprise, trust, sadness, disgust, anticipation, wrath, and their relationships are mentioned in Plutchik's wheel model (Figure 1(c)).

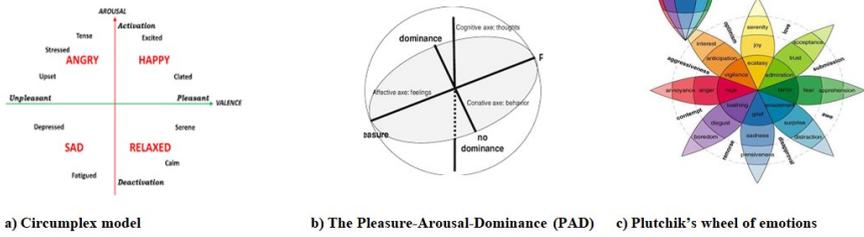


Fig.1. Emotion Models [8, 9, 11]

### 2.3 Modalities of Affect Recognition

The automated recognition of human affect is governed by the usage of various emotional channel modalities. Humans use both verbal and nonverbal channels to judge expressive behavior as predictors of interpersonal consequences [12]. Verbal channels correlate to speech and Nonverbal channels include speech prosody, eye look and blink, body gestures and face expressions. Humans are said to rely on facial expression and verbal prosody the most when interpreting emotions. Nonverbal signals are less susceptible to deliberate manipulation and are effectively representative of real affective status.

**Unimodal Affect Recognition.** The basis for Unimodal affect computing is Physical modalities like textual, audio, visual and physiological modalities like EEG or ECG [13, 14] (Figure 2). Research paper [15] depicts that the most accurate method for study and analysis of sentiment and recognizing emotions through sentiment is textual-based affective analysis. The auditory signals are sensitive to noise. The visual-based emotion recognition [16, 17] is more successful as they are easier to record and its emotional information is more helpful in understanding the humans' emotion state. The physiological signals are more challenging to be captured through wearable devices but as of their objective and trustworthy results physical modalities like EEG-based and ECG-based [18, 19] are much explored by the researchers.



Fig.2. Modalities of Affect Recognition [13, 14]

*Audio Modality.* Affective information in Speech is conveyed by linguistic information (message's semantics through Linguistic speech channel) and implicit paralinguistic information (prosody through Paralinguistic speech-prosody channel). Linear and non-linear acoustic quantifying methods are applied for Audio Emotion Recognition [20]. Machine learning based Speech Emotion Recognition is based on acoustic features extraction and classifiers selection while DL (Deep Learning) based Speech Emotion Recognition creates CNN (Convolutional Neural Network) architecture to emotion prediction [21]. Apart from prosodic and spectral features, voice-quality, frequency-domain, time-domain, hybrid, deep, statistical, and other features are essential for speech emotion recognition [22, 23]. k-Nearest Neighbor, NN (Neural Network), MLB (Maximum-Likelihood Bayes), KR (Kernel Regression), GMM (Gaussian Mixture Model), and HMM (Hidden Markov Model) are some of the most often utilized classifiers in Speech Emotion Recognition systems [22]. OpenSMILE, YAAFE, Timbre, MIR, jAudio, Librosa are the well-known audio feature extraction toolbox [24].

*Textual Modality.* Researchers extract a variety of text sentiments from blogs, product reviews, social media, news, YouTube comments [25, 26, 27, 28, 29]. To date, methods for emotion and sentiment recognition using text are rule-based methods, statistical methods, bag of words (BoW) modeling [30, 31]. The automatic identification of fine-grained emotions, those are conveyed directly or indirectly in text, have been tackled by various researchers [32]. To identify emotional content in texts, several supervised and unsupervised classifiers have been developed [33].

*Visual Modalities.* Visual emotional recognition encompasses emotion recognition through expressions captured from face and body gesture [34,35]. Body expressions reflect more real feelings but are more difficult to manage. The selection of the right modeling of taking input i.e. human traits, feature representation, and output emotions is a crucial step in the preparation for emotion body gesture recognition. In Deep Learning based Emotion Body Gesture Recognition systems the input data is pre-processed through low-level feature extractors. Deep Learning networks like CNN, LSTM (Long Short-Term Memory) or CNN-LSTM based networks can be used to learn high-level characteristics in spatial, temporal, or spatial-temporal dimensions and thus raise the Emotion Body Gesture Recognition performance [36,37]. Being the most natural marker of emotion, Facial expression analysis is the most well-studied nonverbal affect recognition method [38]. Facial emotion recognition uses Images or videos with face emotional clues [39]. Based on static photos or dynamic videos they are classified as static and dynamic facial expression recognition. Macro and micro facial expression on duration and intensity of facial expression and 2D/3D/4D on dimensions are the further categories of facial expression recognition. [40, 41]. Figure 3 depicts the various Learning method and models for Face Emotion Recognition [42-48].

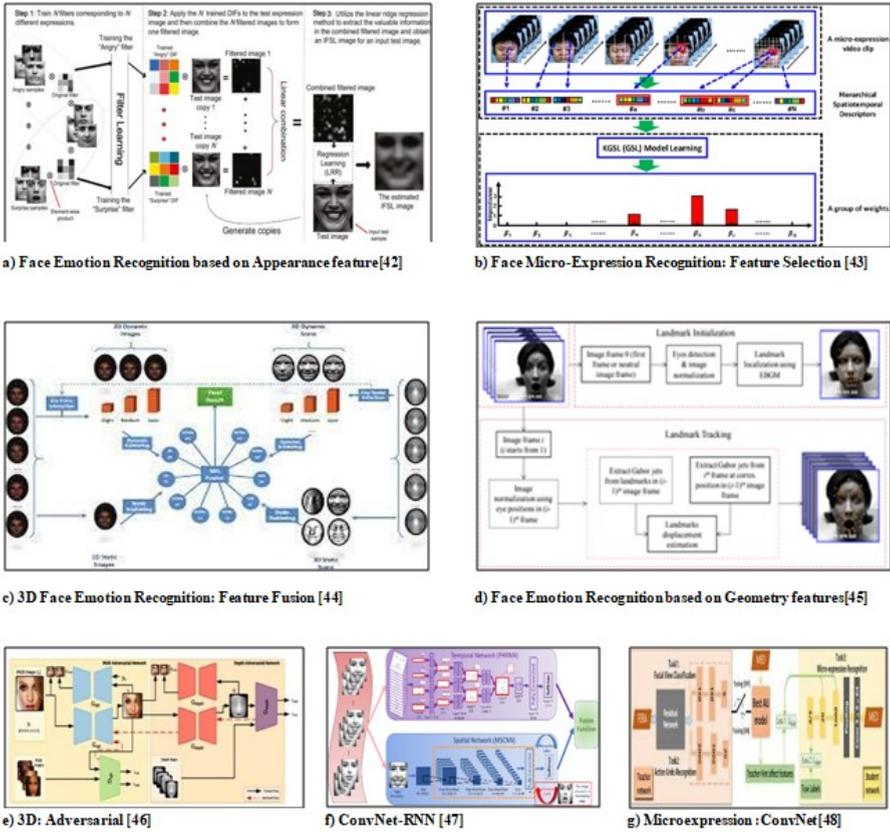


Fig. 3. Learning Method and Models for Face Emotion Recognition

*Physiological Modalities.* Physiological signals captured through non-invasive sensors can be studied and can represent emotions by detecting biological patterns [49]. Noncontact multiparameter physiological measurements, brain imaging and remote cardiopulmonary measurement are now available for capturing signals remotely [50, 51]. Various forms of physiological signals used for emotion detection are Electrocardiography (ECG), Galvanic Skin Response (GKR) and Electrodermal activity (EDA), electromyography (EMG), electroencephalography (EEG), Respiration Rate (RR), and Skin Temperature (TSK). The electrical signal from the heart is recorded by an ECG to check for different heart conditions and calculates the heart rate (HR) and heart rate variability (HRV) [52]. EEG reveals intrinsic aspects of emotional states and allows for real-time monitoring of emotional states [53].

## 2.4 Multimodalities of Affect Recognition

Affect-recognition algorithms that incorporate numerous modalities of emotion expressions are the Multimodal methods. Combining multiple-affective signals gives a more diverse set of data as well as helps to mitigate the consequences of raw signal ambiguity. Two crucial components of the multimodal emotional analysis are the fusion techniques (fusion based on decision, feature, model and hybrid) and the blending of several modalities (blending various physical modalities, various physiological modalities, and blending physical modalities with physiological modalities).

**Multimodal Fusion Techniques.** The goal of Multimodal fusion is to improve estimation precision and reliability [54]. Most recent studies of multimodal affective analysis concentrate on multimodal fusion procedures [55], which are divided into four categories: decision-level, feature-level, model-level, and hybrid-level fusion respectively. In early fusion or Feature-level fusion [56, 57], the input features are taken from several modalities, such as visual and auditory text characteristics. These features are integrated and combined into a general feature vector and then forwarded to a classifier for study and further investigation. In Decision-level fusion or late fusion, the decision vectors are generated from selected modality, which are then combined into a single general feature vector.

Fusion at the model level [58, 59] is a technique that combines data from several modalities and uses correlation to create a relaxed fusion such as HMM. Hybrid fusion blends decision-level and feature-level to take advantage of the benefits of both and avoid the drawbacks of each. [60, 61]. In Rule-based fusion methods, the data belonging to multiple modalities is combined using statistical rule-based methods. The widely used rule-based method i.e. linear weighted fusion approach [62] is less computationally expensive because before merging multimodal data the normalized weights are given to each modality and it employs sum or product operators.

**Blending of Several Modalities.** The concepts of blending of multiple physical modalities are AV (audio-Visual) [63], AT (Audio-Textual) [64], and AVT (Audio-Video-Textual)) [65]. Performance over unimodal affect recognition can be improved by combining visual and aural information [63] or by combining multiple physical modalities with the text modality [66]. In the concept of blending various physiological modality fusions, EEG and ECG [67], EMG, GSR, EOG, and BVP are integrated in investigations for affective analysis. For visual-physiological emotional analysis the multimodal physiological signals [68, 69] and the visual modalities like facial expression, gesture, and voice may be combined for affective analysis.

### 3 Databases Available for Different Modalities and Their Feature Fusion

Emotional computing databases can be divided into Textual Sentiment Database, Speech/Audio Databases, Facial Expression Recognition Databases, Physiological Databases, and Multimodal Databases. The network architecture and model design for emotional computing are profoundly impacted by the characteristics of emotional computing databases.

#### 3.1 Textual Sentiment Database

IMDB [70] is a binary sentiment analysis dataset consisting of 50,000 reviews for natural language processing or Text analytics from the Internet Movie Database (IMDb) labeled as positive or negative. Multi-domain Sentiment (MDS) [71] features slightly older product reviews from Amazon. Seven sentiment categories i.e. positive, negative, strong positive, strong negative, weak negative, weak positive, neutral, are assigned to the reviews comprising 100,000 sentences. Stanford Sentiment Treebank v2 (SST2) [72] is a Stanford Dataset with completely labeled parse trees and has emotional labels on around 215,154 phrases in a parse tree of 11,855 sentences for predicting Sentiment from longer Movie Reviews.

#### 3.2 Speech/Audio Databases

CREMA-D [73] is an audio data set of 7,442 original clips from 91 actors (48 male and 43 female). 12 sentences were presented using one of six different emotions (Disgust, Neutral, Anger, Sad, Fear, Happy), and four different emotion levels (Low, Medium, High, and Unspecified). Emo-DB i.e. Berlin Database of Emotional Speech [74] is a freely available German emotional database that contains over 535 utterances spoken by five males and five females in 7 emotional states, including happiness, rage, anxiety, fear, boredom, and contempt. Ryerson Audio-Visual Database of Emotional Speech and Song (RAVD ESS)[75] is a dataset of speech and song, audio and video containing 1440 records and 24 professional actors (12 female, 12 male), vocalizing two statements, those are lexically-equivalent, in a unbiased North American accent.

#### 3.3 Facial Expression Recognition Databases

JAFPE [76] contains 213 photos of 7 face expressions captured from the Japanese female models. The Oulu-CASIA NIR-VIS [77] contains around 2,800 image sequences taken with different imaging systems and different lighting scenarios. Subjects were given instructions to make 7 different facial expressions in order to develop the extended Cohn-Kanade (CK+) [78]. BU-3DFE (Binghamton University 3D Facial Expression) [79] database comprises around 600 facial expression sequences recorded from hundred persons. MMI [80] is composed of onset-apex-offset sequences as op-

posed to CK+. A large-scale, unrestricted database called FER2013 [81] contains around 35,000 gray images with a resolution of 4848 pixels that were automatically gathered using the Google image search API. One million photos in EmotioNet [82] have 25,000 personally annotated Action Units and 950,000 automatically annotated Face Action Units.

### 3.4 Body Gesture Emotion Databases

As compared to the slight or fake changes in the face, it is usually easy to interpret physical expressions. Natural body activities and movements are captured spontaneously from real time scenarios like interviews, cameras or non-spontaneously from movies which are then compiled to form the corpus of video sequences and thus form Body gesture emotion databases. EmoTV1 [83], is a large collection of video clips recorded from French TV channels containing interviews. The publicly accessible bimodal FACE and BODY database (FABO) [84] contains both face expressions and body gesture. The fragments from two movie versions that make up THEATER Corpus [85] are tagged with eight emotive states that correspond to the corners of PAD [86]. Body motions in daily activities were gathered for affect analysis in Emotional Body Expression in Daily Actions (EMILYA) database [87].

### 3.5 Physiological Signals Datasets

The physiological signals are not impacted by social masking, making them more purposeful, effective and dependable for affect recognition. DSdRD (Real-World Driving Tasks) tool is applied for detection and assessment of pressure and anxiety levels in drivers. Twenty Four participants gave diverse signals that completed their driving jobs and then relaxed for at least 50 minutes. DEAP [88] includes data from 32 people, including 32-channel EEG, EOG, plethysmograph EMG, body temperature, RESP and GSR. EGG recordings from 15 participants are available in SEED [89, 90]. AMIGOS [91] dataset was created with the intention of gathering participant feelings in both individual and group settings. High-quality multimodal data of WESAD (Wearable Stress and Affect Detection) [92] is useful in detecting strain and anxiety with emotional states as neutral, stress or amusement.

### 3.6 Multimodal Datasets

Changes in emotions of around 30 people watching different films were observed and recorded to create a video-physiological database MAHNOB-HCI [93]. A multimodal dataset RECOLA (Remote Collaborative and Affective Interactions) [94] is created through spontaneous interactions from 46 participants. Largest gender balanced dataset CMU-MOSEI (CMU Multimodal Opinion Sentiment and Emotion Intensity)[95] dataset comprised 23,500 sentence utterance videos from around 1000 YouTube speakers. A multimodal corpus ICT-MMMO (Creative Technologies Multimodal Movie Opinion database) [96] contains around 300 YouTube videos and 80 ExpoTV movie review videos for affect recognition. CreativeIT [97] contains rich full-body motion, visual-audio, and text description data of 16 participants. 13 favorable, 12

negative YouTube videos and 22 neutral videos can be found in the HOW (Harvesting Opinions from the Web database) [98]. The important databases used for affective computing are listed in Table 1.

**Table 1. Multimodal Datasets**

<b>Dataset</b>	<b>Ref</b>	<b>Modality</b>	<b>Features</b>
RAVDESS	[75]	Video and Audio	440 files: 60 trials per actor x 24 actors = 1440.
YouTube dataset	[98]	Text, Video and Audio	1000 linguistic + 1941 acoustic +20 visual
Belfast	[99]	Video and Audio	Wide range of emotions
CH-SIMS	[100]	Video, face, speech, eye, text	Around 2280 segments
CMU-MOSEI Dataset	[101]	language, visual, and acoustic face, eye, speech, text, video	23453 sentences utterance videos 2199 videos
CMU-MOSI	[102]	Video, face, speech, eye, text	Around 2200 clips
DEAP	[103]	Face, GSR, EEG, BVP, ECG, RA, EOG, ST, EMG	1280 samples
Human3.6M	[104]	Images, videos	activities by 11 professional actors in 17 scenarios
Ekman-6	[105]	video (audio, image)	1637 videos
EMDB	[106]	video, Skin-Conductance-Level, and Heart-Rate	52 clips
EmoReact Dataset	[107]	Video and Audio	1102 audio-visual clips, six basic emotions, neutral, valence and nine complex emotions
eNTER-FACE	[108]	Video and Audio	Happiness, fear, anger, surprise, sadness, , disgust
HUMAINE	[109]	Video and Audio	core affect dimensions, Authenticity, context label, Emotion words
IEMOCAP	[110]	face, speech, t-text, video	10039 turns
LIRIS-ACCEDE	[111]	video (audio, image)	9800 clips
MAHNOB-HCI	[112]	face, eye, audio, EEG, ECG, GSR, ST, RA	532 samples
PATS	[113]	aligned pose, audio & transcripts	251 hours of data
SEMAINE	[114]	Audio, Video, face, speech	959 conversations
Social-IQ	[115]	videos, questions and answers	1250 videos, 7500 questions

IEMOCAP	[116]	Video and Audio	Happiness, neutral state, frustration, anger, sadness
ICT-MMMO	[95]	Text, Video and Audio	+20 visual , 1941 acoustic and 1000 linguistic
MOUD	[117]	Text, Video and Audio	28 acoustic + 40 visuals
MELD	[118]	Text, Video and Audio	1400 dialogues, 13708 utterances
K-EmoCon	[119]	EEG, and peripheral physiological recordings,	16 sessions of approximately 10-minute long paired debates
EMOTIC	[120]	images	23, 571 images and 34, 320 annotated people
Dreamer	[121]	EEG,ECG, audio-visual stimuli	Signals from 23 participants
Ascertain	[122]	EEG, ECG, GSR and facial activity data	5 personality scales & emotional self-ratings of 58 users
AffectNet	[123]	Facial expressions along with the intensity of valence and arousal.	0.4 million images manually labeled

### 4 Machine Learning and Deep Learning Based Models for Emotion Recognition

Machine Learning based techniques provides feature extractors, classifiers, and pre-processing techniques for raw signals [21, 54, 62, 69,124,125]. The Random forest (RF), the Support Vector Machine (SVM), Gaussian Mixture model (GMM), Hidden Markov Model (HMM), Artificial Neural Network (ANN) and K-Nearest Neighbor (KNN) are the most popular Machine Learning based classifiers. Due to their limitation in task and domain specific feature descriptors, Machine Learning based approaches for affective analyses are difficult to apply [81].

**Table 2.Machine Learning & Deep Learning based Techniques for Physiological emotion recognition**

Modality	Ref	Year	Feature Representation	Classifier	Modality	Ref	Year	Feature Representation	Classifier
EEG	[156]	2016	mRMR-based	SVM	ECG	[160]	2017	Feature Fusion+LLN	KNN
	[157]	2019	ADMM-based	Multi-class SGL		[161]	2017	Feature Fusion	SVM

[158]	2018	Bi-DANN	LSTM+ Soft-max	[162]	2019	EmotionalGAN	SVM
[159]	2020	Spatial temporal	Pooling+Sof tmax	[163]	2020	Self-supervised CNN	FC+ Sig-moid

**Table 3. Machine Learning and Deep Learning based Techniques for Textual, Speech and Face Emotion Recognition**

Modality	Machine Learning Techniques				Deep Learning Techniques			
	Ref	Year	Feature description	Classifier	Ref	Year	Feature description	Classifier
Textual Sentiment Analysis	[134]	2014	Multi-feature fusion	SVM - ELM	[138]	2020	CNN	HieN N-DWE CNN
	[135]	2018	Multilingual features	Metric-based	[139]	2020	Word2Vec	- LSTM
	[136]	2008	CDM	NB	[140]	2018	PF-CNN	Soft max
	[137]	2019	Features fusion	LML	[141]	2018	Word Embeddings	GCA E
	[124]	2013	MFCC	SVM	[145]	2018	Adversarial AE	SVM
Speech Emotion Recognition	[142]	2010	Acoustic feature	SVM	[146]	2016	CNNs-LSTM	LSTM
	[143]	2014	Feature selection & fusion	MKL	[147]	2018	CNN/Attention-BLSTM	DNN
	[144]	2020	Acoustic features	TL-FMR F	[84]	2018	Conditional adversarial	CNN
Face Emotion Recognition	[148]	2020	IFSL	SVM	[152]	2021	CNN-LSTM	6 Col-Cla
	[149]	2019	LPDP	SVM	[153]	2021	SCAN	FC
	[150]	2019	DSF	KNN	[154]	2020	GCN-CNN	FC
	[151]	2018	HSDS	SVM	[155]	2020	CNN-LSTM	Pooling

In most instances of affective computing, DL-based models performed better than Machine Learning based models [21, 36, 69, 126-128]. As of superior feature representation learning capabilities the DL-based algorithms are capable of automatically learning the most discriminative physical traits. Also, the impact of database size and quality is more prominent than for Machine Learning based emotion recognition. The deep spatial-temporal feature extraction is a task that can be handled by CNN-LSTM models [37]. Cross-domain learning and Adversarial learning are frequently employed to increase the robustness of models [129]. For overall improved performance, various attention mechanisms [130] and auto encoders [131] are used with DL-based approaches. The temporal dynamics for sequence information are captured through RNNs and its variants [132]. The significant and discriminative characteristics are captured through CNNs and their derivatives from static information, face and spectrogram images [133]. The ML and DL techniques are summarized in Table 2 and Table 3.

## 5 Evaluation Metrics

Evaluation metrics are vital for assessing the performance and effectiveness of models in multimodal emotion recognition and affective computing. Table 4 and Table 5 cover the metrics used for emotion recognition.

**Table 4. Multimodal Specific Evaluation Metrics**

Metrics	Description	Significant study	Ref
Modality Contribution Analysis	It calculates how much each modality—such as text, audio, and facial expressions—contributes to the total accuracy of recognition.	It measures the impact of individual modalities	[164]
Fusion Accuracy	It assesses how well the fusion method of combining a variety of modalities works.	In order to priorities each modality according to context, research has concentrated on developing strong fusion approaches, such as transformers & attention processes.	[165]

**Table 5. Common Evaluation Metrics**

Metrics	Description	Particulars	Significant study
Accuracy- a simple metric	The amount of accurate forecasts among all predictions.	In unbalanced datasets having underrepresented classes, accuracy can be unreliable.	The limitations of accuracy brought on by unbalanced data. The more appealing option to measure is that offers a more refined and detailed perspective.

Precision (Relevance)	The proportion of accurately expected positive observations to all predicted positive observations.	It measures how relevant the predicted emotions are.	
Recall (sensitivity) (completeness)	The proportion of accurately predicted positive observations to all observations in the actual class	It measures how well the model captures all relevant emotions	F1-Score is frequently chosen in recent research to provide a fair evaluation because emotion data frequently exhibits imbalance as certain emotions are uncommon.
F1-Score	Precision and Recall's harmonic mean. F1-Score balances both metrics.	It strikes a compromise between recall and precision, which is crucial when there is a class imbalance.	
Concordance Correlation Coefficient (CCC)	It combines accuracy and precision to assess how well two sets of continuous data agree	Regression-based tasks in multimodal emotion identification, such as valence-arousal prediction, benefit greatly from its utilization.	It takes bias and linear correlation into account; it is being used more and more in research to assess models on continuous emotional states.
Confusion Matrix	It compares predicted and actual classes to show how well the model performed.	It can assist in recognizing patterns of misclassification.	In order to improve misclassifications, researchers are analyzing emotion-specific errors using confusion matrices & improving model topologies.
Mean Squared Error (MSE), Mean Absolute Error (MAE)	They are often used metrics in regression-based emotion detection (e.g., predicting valence and arousal).	They assist in predicting emotional intensity on a continuous scale.	In order to demonstrate their improvement in predicting finer emotional details, models that predict continuous emotion scales (such as valence-arousal space) frequently report lower MSE or MAE.
Weighted Metrics (Weighted Accuracy,	It assigns varying degrees of priority to various classes	It improves the estimation of rare emotions. It is frequently	To provide a more impartial assessment, researchers are creating weighted measures that take the distribution of

Weighted F1)		employed when working with unbalanced datasets.	the data into account.
AUC Area Under the Curve	Evaluates how well the model can differentiate between classes.	The model performs better if its AUC is larger. AUC-ROC assesses the ability to differentiate among emotion classes at a variety of thresholds.	In multimodal systems, AUC-ROC is becoming more and more significant, particularly when the performance can be greatly impacted by the classification threshold.
ROC Receiver Operating Characteristic	A visual representation of a classifier's diagnostic performance when its discrimination threshold is changed.	It punishes incorrect classifications. It works well with complex multimodal data	It is commonly used during training in deep learning models for emotion recognition, especially those that use transformer topologies
Cross-Entropy Loss (Log Loss)	A loss function that gauges how well a classification model performs when the output is a probability value ranging from 0 to 1.	It assigns equal weight to each emotion class. it is preferred when datasets are very unbalanced.	In benchmarks requiring extremely unbalanced datasets for emotion recognition, UAR is becoming a standard metric.
Unweighted Average Recall (UAR)	The average of recall scores across all classes, excluding class imbalance		

## 6 Recent Advancements, Challenges and Future Directions

Regardless of significant advancements in affective computing, challenges exist in multimodal emotion recognition, including difficulties in extracting and handling multiple modalities, disambiguating complex emotional states, and dynamically modeling emotions. Challenges such as data incompleteness, ethical concerns, and cultural variations complicate the development of practical applications. The shortage of large-scale labeled datasets and the issues of transferring models across domains further hamper progress. In addition, ensuring user privacy and minimizing emotional strain in real-world applications are some of the serious considerations. Future research in multimodal fusion, domain adaptation, and few-shot learning may assist in overcoming these obstacles and facilitating more accurate and ethical emotion recognition.

systems. Table 6 covers the latest trends and Table 7 records future directions, challenges and probable solutions in the field of multimodal emotion recognition.

**Table 6. Latest Research Trends in Multimodal Affective Computing**

<b>Latest Trends</b>	<b>Description</b>	<b>Significant study</b>	<b>Ref</b>
Transformer-based Models	In multimodal situations, transformers such as BERT and ViT (Visual Transformers) are used to model intricate relationships between modalities.	According to recent research, transformers can greatly improve performance, particularly when attention methods are used to highlight important emotional cues.	[166]
Attention Mechanisms for Modality Fusion	It dynamically weighs every modality based on its emotional involvement, thus improves the interpretability and accurateness.	To dynamically balance the relative importance of each modality, attention techniques are commonly used.	[167]
Robustness to Noisy & Missing Data	Multimodal recognition still faces difficulties when dealing with noisy or missing data. To increase model robustness, methods like modality dropout and robust modality fusion are being researched.	It enhances the flexibility to incomplete or noisy data.	[168]
Large-Scale Multimodal Datasets	Research indicates that sizable datasets aid in capturing a wider range of emotions, resulting in more reliable models.	To enhance generalization, researchers are utilizing a variety of extensive datasets, such as MEAD, EmoReact, and MELD. It helps in capturing a broader spectrum of emotions.	[169]
Multimodal Pretraining	Pretraining on large datasets, followed by fine-tuning on specific emotion tasks, improves the generalization across contexts.	Recent research shows that pretrained multimodal models, such as MMF (Multimodal Fusion Framework), boost performance.	[170]

**Table 7. Future Directions, Challenges and Probable Solutions**

Category	Challenges and Issues	Probable Solutions
Multimodal Extraction & Handling	<ul style="list-style-type: none"> <li>-Handling multiple modalities simultaneously</li> <li>-Non-linear disambiguation</li> <li>-Dynamic modeling of emotions.</li> </ul>	<p>Use advanced fusion techniques (e.g., hybrid feature- and decision-level fusion) to manage multimodal integration.</p> <p>Prefer usage of attention-based models that helps in focusing on relevant features from each modality, thus improves the model robustness and interpretability.</p>
Growing Security Threats	<ul style="list-style-type: none"> <li>-Bypassing of current systems with the increasing complexity of security threats</li> <li>-Added risks from cloud-based solutions.</li> </ul>	<p>For the protection of data during training, incorporate secure machine learning practices like differential privacy and federated learning</p> <p>For the identification of suspicious patterns across modalities, develop and make use of multimodal anomaly detection systems</p>
Realistic and Practical Application	<ul style="list-style-type: none"> <li>-Effective use of multimodal data for stress prediction, real-time recognition, and mood impact.</li> <li>-Simplify the data collection.</li> <li>-tackle ethical considerations and misconceptions.</li> </ul>	<p>Prefer using non-intrusive, wearable sensors, and design systems with transparency and opt-in privacy features to Streamline the data collection.</p> <p>Provide extensive user education to improve misconceptions, beside clear ethical guidelines and limitations for data usage.</p>
Multimodal Fusion	<ul style="list-style-type: none"> <li>-Underutilization of Fusion techniques in multimodal affective computing, especially in NLP and sentiment analysis.</li> <li>-synchronization, noise handling, and dimensionality reduction</li> </ul>	<p>Synchronization algorithms like dynamic time warping and incorporate adaptive noise filtering mechanisms could be applied. Explore efficient fusion methods (e.g., tensor fusion networks) that reduce dimensionality while maintaining essential emotion features.</p>
Baseline Models & Datasets	<p>Lack of standardized multimodal affect databases, particularly for discrete emotion recognition.</p>	<p>Develop comprehensive multimodal datasets that cover diverse emotions, cultural nuances, and real-world variability. Utilize data augmentation strategies to expand existing datasets without additional manual labeling.</p>

Affective Gap	The emotive gap between extracted features and perceived emotions, even after closing the semantic gap, poses a challenge for multimodal emotion recognition.	Leverage hybrid affective models combining physiological signals (e.g., ECG) with other modalities to better capture underlying emotions. Use reinforcement learning to fine-tune emotion prediction models, bridging the gap between low-level features and high-level emotions.
Cultural & Personal Variability	Diverse emotional responses due to factors like culture, personality, and environment, impacting recognition accuracy.	Create models that account for cultural diversity by training on region-specific datasets. Use transfer learning to adapt a base model to new cultural and personal contexts, thus improving its generalization to varied populations.
Data In-completeness	Missing or partial data due to collection challenges, which may cause conflicts between modalities.	For the estimation of missing data, prefer using imputation methods, such as generative adversarial networks (GANs). In case of arising conflicts, prefer using adaptive weighting techniques to prioritize the most reliable modalities.
Modality Contribution Disparity	Uneven contributions of modalities, e.g., lengthy articles with limited images, or using neutral images for sentimentally charged content.	Use modality-specific attention mechanisms to weight the contribution of each modality dynamically based on the content's emotional context. The context-aware models can be developed to adjust with respect to modality imbalances, and thus allows focusing on the most informative channels.
Large-Scale Labeled Data	Need for vast, labeled datasets for multimodal affect recognition, with issues in data labeling inconsistency and noise.	Adopt semi-supervised and weakly supervised learning frameworks to minimize the need for extensive labeled data. Explore few/zero-shot learning and self-supervised methods to enable affective computing models to learn with limited labeled data.
Domain Adaptation	Transferring models across domains leads to performance degradation due to domain shift.	Employ domain adaptation methods, such as adversarial domain adaptation or transfer learning, to bridge domain gaps. Consider domain generalization approaches to create robust models that work across multiple domains without re-training.
Privacy and Ethics	Privacy concerns in gathering emotional data, with potential misuse or harmful applications impacting user consent and wellbeing.	Prioritize user privacy through differential privacy techniques and enforce data encryption. Introduce privacy-aware affective computing solutions that provide opt-in consent and allow data control to users,

Future Goal: Real-Time Re-cognizer	Developing a real-time multimodal affect recognizer remains crucial for applications like mental health support, but challenges indicate a long way to go.	minimizing unintended consequences or emotional strain.
		Focus on developing lightweight models optimized for low latency to achieve real-time performance. Advances in edge AI can enable on-device processing, maintaining speed while safeguarding user data locally for sensitive emotional applications.

## 7 Conclusion

To explain the subject of emotion recognition and affective computing, this paper described key theoretical ideas and discussed the state-of-the-art in the industry. We described recent developments in emotional computing, which may be divided primarily into unimodal and multimodal emotion recognition. We introduced category, dimensional, and component emotional models respectively. We selected significant research on unimodal affect identification because we believe these studies are essential building blocks for a multimodal affect detection framework. Voice emotion recognition, textual sentiment analysis, emotion body-gesture recognition, facial expressions, and physiological emotion recognition are the several types of unimodal affect recognition systems. Knowing the state-of-the-art in the domain of affect recognition on single modalities would make building a suitable multimodal framework easier. The multimodal affective analysis is typically divided into three categories based on blending either various physical, multiple physiological and several physical-physiological approaches. Additionally, we covered how modality fusion techniques affect multimodal emotion recognition. Both the modality combination and fusion approach have a significant impact on how well multimodal emotional analysis performs. We also discussed how Machine Learning and Deep Learning based models affect affective computing. The paper also discussed the available benchmark datasets, recent work and literature of multimodal Affective Computing. We examined the challenges and potential possibilities for future affective computing research.

## References

1. Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., Gao, S., Sun, Y., Ge, W., Zhang, W., & Zhang, W.: A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 83–84 (2022).
2. Pathare, S., Vijayakumar, L., Fernandes, T., et al.: Analysis of news media reports of suicides and attempted suicides during the COVID-19 lockdown in India. *International Journal of Mental Health Systems*, 14(1), 88 (2020).
3. Fleckenstein, K. S.: Defining affect in relation to cognition: A response to Susan McLeod. *Journal of Advanced Composition*, 11, 447–453 (1991).

4. Al-Maliky, R.S.B.: A semantic approach to emotion recognition using IBM Watson Bluemix tone analyzer and translator language. In: 2nd International Conference on Engineering Technology and its Applications (IICETA) 2019, pp. 1–1. IEEE (2019).
5. Lee, W., Norman, M.D.: Affective computing as complex systems science. *Procedia Computer Science*, 95, 18–23 (2016).
6. Korsmeyer, C. Rosalind W. Picard.: Affective Computing. *Minds and Machines* 9(3), 443–447 (1999).
7. Calvo, R. A., & Kim, M.: Emotions in text: Dimensional and categorical models. *ACM Computational Intelligence*, 29(3), 527–543 (2013).
8. Russell, J.: A circumplex model of affect. *Journal of Personality and Social Psychology*, 39, 1161–1178 (1980).
9. Bakker, I., van der Voordt, T., Vink, P., & de Boon, J.: Pleasure, arousal, dominance: Mehrabian and Russell revisited. *Current Psychology*, 33, 405–421 (2014).
10. Mehrabian, A.: Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies. Cambridge, MA: Oelgeschlager, Gunn & Hain (1980).
11. Plutchik, R.: *Emotion and life: Perspective from psychology, biology, and evolution*. Washington, DC: American Psychological Association (2003).
12. Ambady, N., & Rosenthal, R.: Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111, 256 (1992).
13. Christy, T., & Kuncheva, L.: Technological advancements in affective gaming: A historical survey. *GSTF Journal on Computing (JoC)*, (2014).
14. Zhao, S., Jia, G., Yang, J., Ding, G., & Keutzer, K.: Emotion recognition from multiple modalities: Fundamentals and methodologies. *IEEE Signal Processing Magazine*, 38(6), 59–73 (2021).
15. Akhtar, M.S., Chauhan, D., Ghosal, D., Poria, S., Ekbal, A., Bhattacharyya, P.: Multi-task learning for multi-modal emotion recognition and sentiment analysis. In: *Proc. of NAACL-HLT 2019*, vol. 1. Association for Computational Linguistics (2019).
16. Azazi, A., Lutfi, S.L., Venkat, I., Fernández-Martínez, F.: Towards a robust affect recognition: Automatic facial expression recognition in 3D faces. *Expert Systems with Applications* 42(8), 3056–3066 (2015).
17. Fang, H., Mac Parthaláin, N., Aubrey, A.J., Tam, G.K.L., Borgo, R., Rosin, P.L., Grant, P.W., Marshall, D., Chen, M.: Facial expression recognition in dynamic sequences: An integrated approach. *Pattern Recognition* 47(3), 1271–1281 (2014).
18. Alarcão, S.M., Fonseca, M.J.: Emotions recognition using EEG signals: A survey. *IEEE Transactions on Affective Computing* 10(3), 374–393 (2019).
19. Hsu, Y.-L., Wang, J.-S., Chiang, W.-C., Hung, C.-H.: Automatic ECG-based emotion recognition in music listening. *IEEE Transactions on Affective Computing* 11(1), 85–99 (2020).
20. Krishnan, P.T., Raj, A.N.J., Rajangam, V.: Emotion classification from speech signal based on empirical mode decomposition and non-linear features. *Complex & Intelligent Systems* 7, 1919–1934 (2021).
21. de Lope, J., Hernández, E., Vargas, V., Graña, M.: Speech emotion recognition by conventional machine learning and deep learning. In: *Hybrid Artificial Intelligent Systems: Proc. of HAIS 2021*, Bilbao, Spain, pp. 319–330 (2021).
22. Swain, M., Routray, A., Kabisatpathy, P.: Databases, features and classifiers for speech emotion recognition: A review. *International Journal of Speech Technology* 21(1), 93–120 (2018).

23. Akçay, M.B., Oğuz, K.: Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication* 116, 56–76 (2020).
24. Moffat, D., Ronan, D., Reiss, J.: An evaluation of audio feature extraction toolboxes. In: *Proc. of the AES International Conference* (2015).
25. Douiji, Y., Mousannif, H., Al Moatassime, H.: Using YouTube comments for text-based emotion recognition. *Procedia Computer Science* 83, 292–299 (2016).
26. Andalibi, N., Buss, J.: The human in emotion recognition on social media: Attitudes, outcomes, risks. In: *Proc. of CHI 2020* (2020).
27. Aldous, K.K., An, J., Jansen, B.J.: Measuring emotions of news posts from social media platforms over eight months. *ACM Transactions on Social Computing* 4(4), Article 15, 1–31 (2021).
28. Lin, K.-H.Y., Yang, C., Chen, H.-H.: What emotions do news articles trigger in their readers? In: *Proc. of SIGIR 2007*, pp. 733–734 (2007).
29. Jiang, Y., Wang, H., Yi, T.: Evaluation of product reviews based on text sentiment analysis. In: *Proc. of AIAIS 2021* (2021).
30. Alm, C.O., Roth, D., Sproat, R.: Emotions from text: Machine learning for text-based emotion prediction. In: *Proc. of HLT-EMNLP*, pp. 579–586 (2005).
31. Patacsil, F.: Combining rule-based and bag-of-words for phase-level sentiment analysis of blog comments. *Int. J. Eng. Technol. (UAE)* 7, 183–190 (2018).
32. Singh, G., Brahma, D., Rai, P., Modi, A.: Fine-grained emotion prediction by modeling emotion definitions (2021).
33. Azari, B., Westlin, C., Satpute, A. B., et al.: Comparing supervised and unsupervised approaches to emotion categorization in the human brain, body, and subjective experience. *Sci. Rep.* 10, 20284 (2020).
34. Gunes, H., Piccardi, M.: Bi-modal emotion recognition from expressive face and body gestures. *J. Netw. Comput. Appl.* 30(4), 1334–1345 (2007).
35. Chaudhary, A., Nunes, R., Nasrollahi, K., Rehm, M., Moeslund, T.: Deep emotion recognition through upper body movements and facial expression (2021).
36. Santhoshkumar, R., Kalaiselvi Geetha, M.: Deep learning approach for emotion recognition from human body movements with feedforward deep convolutional neural networks. *Procedia Comput. Sci.* 152, 158–165 (2019).
37. Ly, S. T., Lee, G.-S., Kim, S.-H., Yang, H.-J.: Gesture-based emotion recognition by 3D-CNN and LSTM with keyframes selection. *Int. J. Contents* 15(4), 59–64 (2019).
38. de Gelder, B.: Why bodies? Twelve reasons for including bodily expressions in affective neuroscience. *Philos. Trans. R. Soc. B Biol. Sci.* 364, 3475–3484 (2009).
39. Tian, Y.-I., Kanade, T., Cohn, J. F.: Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 23, 97–115 (2001).
40. Allaert, B., Bilasco, I. M., Djeraba, C.: Micro and macro facial expression recognition using advanced local motion patterns. *IEEE Trans. Affect. Comput.* (2019).
41. Zhen, Q., Huang, D., Wang, Y., Chen, L.: Muscular movement model-based automatic 3D/4D facial expression recognition. *IEEE Trans. Multimed.* 18, 1438–1450 (2016).
42. Yan, Y., Zhang, Z., Chen, S., Wang, H.: Low-resolution facial expression recognition: A filter learning perspective. *Signal Process.* 169, 107370 (2020).
43. Zong, Y., Huang, X., Zheng, W., Cui, Z., Zhao, G.: Learning from hierarchical spatiotemporal descriptors for micro-expression recognition. *IEEE Trans. Multimed.* 20, 3160–3172 (2018).

44. Yao, Y., Huang, D., Yang, X., Wang, Y., Chen, L.: Texture and geometry scattering representation-based facial expression recognition in 2D+3D videos. *ACM Trans. Multimed. Comput. Commun. Appl.* 14, 18:1–18:23 (2018).
45. Ghimire, D., Lee, J.: Geometric feature-based facial expression recognition in image sequences using multi-class AdaBoost and support vector machines. *Sensors* 13, 7714–7734 (2013).
46. Bodur, R., Bhattarai, B., Kim, T.-K.: 3D dense geometry-guided facial expression synthesis by adversarial learning. 2021 IEEE Winter Conf. Appl. Comput. Vision (WACV) 2391–2400 (2021).
47. Zhang, K., Huang, Y., Du, Y., Wang, L.: Facial expression recognition based on deep evolutionary spatial-temporal networks. *IEEE Trans. Image Process.* 26(9), 4193–4203 (2017).
48. Sun, B., Cao, S., Li, D., He, J., Yu, L.: Dynamic micro-expression recognition using knowledge distillation. *IEEE Trans. Affect. Comput.* (2020).
49. Dalglish, T., Dunn, B. D., Mobbs, D.: Affective neuroscience: Past, present, and future. *Emotion Rev.* 1(4), 355–368 (2009).
50. Poh, M.-Z., McDuff, D. J., Picard, R. W.: Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Trans. Biomed. Eng.* 58(1), 7–11 (2011).
51. McDuff, D., Gontarek, S., & Picard, R. W.: Improvements in remote cardiopulmonary measurement using a five-band digital camera. *IEEE Transactions on Biomedical Engineering* 61(10), 2593–2601 (2014).
52. Castaldo, R., Montesinos, L., Melillo, P., James, C., & Pecchia, L.: Ultra-short-term HRV features as surrogates of short-term HRV: A case study on mental stress detection in real life. *BMC Medical Informatics and Decision Making* 19(1) (2019).
53. Li, C., Zhang, Z., Song, R., Cheng, J., Liu, Y., & Chen, X.: EEG-based emotion recognition via neural architecture search. *IEEE Transactions on Affective Computing* (2021).
54. Zhao, H., Miao, X., Liu, R., & Fortino, G.: Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges. *Information Fusion* 80, 241–265 (2022).
55. Poria, S., Cambria, E., Bajpai, R., & Hussain, A.: A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37, 98–125 (2017).
56. Sarkar, C., Bhatia, S., Agarwal, A., & Li, J.: Feature analysis for computational personality recognition using YouTube personality dataset. In: *Proceedings of the 2014 ACM Multimedia Workshop on Computational Personality Recognition*, pp. 11–14 (2014).
57. Monkaresi, H., Hussain, M. S., & Calvo, R. A.: Classification of affects using head movement, skin color features, and physiological signals. In: *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2664–2669 (2012).
58. Lin, J.-C., Wu, C.-H., & Wei, W.-L.: Error-weighted semi-coupled hidden Markov model for audio-visual emotion recognition. *IEEE Transactions on Multimedia* 14(1), 142–156 (2012).
59. Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B., & Rigoll, G.: LSTM-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing* 31(2), 153–163 (2013).
60. Ilhan, H. O., Serbes, G., & Aydin, N.: Decision and feature level fusion of deep features extracted from public COVID-19 datasets. *Applied Intelligence* 52(11), 8551–8571 (2022).
61. Ma, Z., Ma, F., Sun, B., & Li, S.: Hybrid multimodal fusion for dimensional emotion recognition. In: *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, pp. 29–36 (2021).

62. Seng, K. P., Ang, L.-M., & Ooi, C. S.: A combined rule-based and machine learning audio-visual emotion recognition approach. *IEEE Transactions on Affective Computing* 9(1), 3–13 (2018).
63. Chang, X., & Skarbek, W.: Multi-modal residual perceptron network for audio-video emotion recognition. *arXiv Preprint* (2021).
64. Yang, K., Xu, H., & Gao, K.: CM-BERT: Cross-modal BERT for text-audio sentiment analysis. In: *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 521–528 (2020).
65. Jin, Q., Li, C., Chen, S., & Wu, H.: Speech emotion recognition with acoustic and lexical features. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4749–4753 (2015).
66. Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D.: M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34(2), 1359–1367 (2020).
67. Goshvarpour, A., Abbasi, A., & Goshvarpour, A.: An accurate emotion recognition system using ECG and GSR signals and matching pursuit method. *Biomedical Journal* 40(6), 355–368 (2017).
68. Yin, G., Sun, S., Yu, D., Li, D., & Zhang, K.: An efficient multimodal framework for large scale emotion recognition by fusing music and electrodermal activity signals. *arXiv Preprint* (2021).
69. Wu, D., Zhang, J., & Zhao, Q.: Multimodal fused emotion recognition about expression-EEG interaction and collaboration using deep learning. *IEEE Access* 8, 133180–133189 (2020).
70. Kumar, K., Harish, B. S., & Darshan, H.: Sentiment analysis on IMDb movie reviews using hybrid feature extraction method. *International Journal of Interactive Multimedia and Artificial Intelligence* (2018).
71. Kumar, P. K., Tejaswi, N. J., Vasanthi, M. L., Srihitha, L. L., & Kumar, B. P.: Sentimental analysis on multi-domain sentiment dataset using SVM and Naive Bayes algorithm. In: Garg, D., Jagannathan, S., Gupta, A., Garg, L., & Gupta, S. (Eds.), *Advanced Computing. IACC 2021. Communications in Computer and Information Science*, vol. 1528, pp. 161–170. Springer, Cham (2022).
72. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: *Proceedings of EMNLP* (2013).
73. Croitoru, F.-A., Ristea, N.-C., Ionescu, R. T., & Sebe, N.: LeRaC: Learning rate curriculum. *arXiv Preprint* (2022).
74. Dair, Z., Donovan, R., & O'Reilly, R.: Linguistic and gender variation in speech emotion recognition using spectral features. *arXiv Preprint* (2021).
75. Livingstone, S. R., & Russo, F. A.: 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* 13(5), e0196391 (2018).
76. Lyons, M., Akamatsu, S., Kamachi, M., & Gyoba, J.: Coding facial expressions with Gabor wavelets. In: *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 200–205 (1998).
77. Zhao, G., Huang, X., Taini, M., Li, S. Z., & Pietikäinen, M.: Facial expression recognition from near-infrared videos. *Image and Vision Computing* 29(9), 607–619 (2011).
78. Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I.: The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified ex-

- pression. In: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 94–101 (2010).
79. Yin, L., Wei, X., Sun, Y., Wang, J., & Rosato, M. J.: A 3D facial expression database for facial behavior research. In: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, pp. 211–216 (2006).
  80. Valstar, M. F., & Pantic, M.: Induced disgust, happiness, and surprise: An addition to the MMI facial expression database (2010).
  81. Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... & Bengio, Y.: Challenges in representation learning: A report on three machine learning contests. In M. Lee, A. Hirose, Z.-G. Hou, & R. M. Kil (Eds.), *Neural Information Processing*, 117–124 (2013). Springer, Berlin, Heidelberg.
  82. Benitez-Quiroz, C. F., Srinivasan, R., & Martinez, A. M.: EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, 5562–5570 (2016).
  83. Abrilian, S., Devillers, L., Buisine, S., & Martin, J.-C.: EmoTV1: Annotation of real-life emotions for the specifications of multimodal active interfaces (2005).
  84. Gunes, H., & Piccardi, M.: A bi-modal face and body gesture database for automatic analysis of human nonverbal affective behavior. Proceedings of the 18th International Conference on Pattern Recognition, 1148–1153 (2006).
  85. Kipp, M., & Martin, J.-C.: Gesture and emotion: Can basic gestural form features discriminate emotions? (2009).
  86. Mehrabian, A.: Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(3), 261–292 (1996).
  87. Fourati, N., & Pelachaud, C.: Emilya: Emotional body expression in daily actions database (2014).
  88. Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., ... & Patras, I.: DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1), 18–31 (2012).
  89. Duan, R.-N., Zhu, J.-Y., & Lu, B.-L.: Differential entropy feature for EEG-based emotion classification. Proceedings of the 2013 6th International IEEE EMBS Conference on Neural Engineering, 81–84 (2013).
  90. Zheng, W.-L., & Lu, B.-L.: Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(2), 162–175 (2015).
  91. Correa, J. A. M., Abadi, M. K., Sebe, N., & Patras, I.: AMIGOS: A dataset for affect, personality, and mood research on individuals and groups. *IEEE Transactions on Affective Computing*, 1–1 (2018).
  92. Chen, Z., Huang, D., Wang, Y., & Chen, L.: Fast and light manifold CNN based 3D facial expression recognition across pose variations. Proceedings of the 26th ACM International Conference on Multimedia, 229–238 (2018). Association for Computing Machinery.
  93. Soleymani, M., Lichtenauer, J., Pun, T., & Pantic, M.: A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3, 42–55 (2012).
  94. Ringeval, F., Sonderegger, A., Sauer, J., & Lalanne, D.: Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 1–8 (2013).

95. Wollmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., & Morency, L.-P.: YouTube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3), 46–53 (2013).
96. Bagher Zadeh, A., Liang, P. P., Poria, S., Cambria, E., & Morency, L.-P.: Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Vol. 1 (Long Papers), 2236–2246 (2018). Association for Computational Linguistics.
97. Metallinou, A., Yang, Z., Lee, C., Busso, C., Carnicke, S., & Narayanan, S.: The USC CreativeIT database of multimodal dyadic interactions: From speech and full body motion capture to continuous emotional annotations. *Language Resources and Evaluation*, 50, 497–521 (2016).
98. Morency, L.-P., Mihalcea, R., & Doshi, P.: Towards multimodal sentiment analysis: Harvesting opinions from the web. *Proceedings of the 13th International Conference on Multimodal Interfaces*, 169–176 (2011). ACM.
99. Douglas-Cowie, E., Cowie, R., & Schroder, M.: A new emotion database: Considerations, sources, and scope. *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 39–44 (2000).
100. Zhao, J., Zhang, T., Hu, J., Liu, Y., Jin, Q., Wang, X., & Li, H.: M3ED: Multi-modal multi-scene multi-label emotional dialogue database. *arXiv* (2022).
101. Liang, P. P., Lyu, Y., Fan, X., Wu, Z., Cheng, Y., Wu, J., ... Morency, L.-P.: MultiBench: Multiscale benchmarks for multimodal representation learning. *arXiv* (2021).
102. Zadeh, A., Zellers, R., Pincus, E., & Morency, L.-P.: MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv* (2016).
103. Ding, Y., Robinson, N., Zeng, Q., & Guan, C.: LGGNet: Learning from local-global-graph representations for brain-computer interface. *Journal of LaTeX Class Files*, 14(8) (2022).
104. Han, F., Reily, B., Hoff, W., & Zhang, H.: Space-time representation of people based on 3D skeletal data: A review. *arXiv* (2016).
105. Nandwani, P., & Verma, R.: A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11, 81 (2021).
106. Carvalho, S., Leite, J., Galdo-Álvarez, S., & Gonçalves, O. F.: The Emotional Movie Database (EMDB): A self-report and psychophysiological study. *Applied Psychophysiology and Biofeedback*, 37(4), 279–294 (2012).
107. Nojavanasghari, B., Baltrušaitis, T., Hughes, C., & Morency, L.-P.: EmoReact: A multimodal approach and dataset for recognizing emotional responses in children. *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)* (2016).
108. Martin, O., Kotsia, I., Macq, B., & Pitas, I.: The eNTERFACE' 05 audio-visual emotion database. *22nd International Conference on Data Engineering Workshops (ICDEW)*, 8–8 (2006).
109. Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., McRorie, M., ... Batliner, A.: The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. *Affective Computing and Intelligent Interaction*, 488–500 (2007). Springer.
110. Li, Z., Tang, F., Ming, Z., & Zhu, Y.: EmoCaps: Emotion capsule-based model for conversational emotion recognition. *Association for Computational Linguistics*, 1610–1618 (2022).
111. Baveye, Y., Dellandréa, E., Chamaret, C., Chen, L.: LIRIS-ACCEDE: A video database for affective content analysis. *IEEE Transactions on Affective Computing* 6(1), 43–55 (2015).

112. Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M.: A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing* 3, 42–55 (2012).
113. Ahuja, C., Lee, D. W., Ishii, R., Morency, L.-P.: No gestures left behind: Learning relationships between spoken language and freeform gestures. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 1884–1895 (2020).
114. McKeown, G., Valstar, M., Cowie, R., Pantic, M., Schroder, M.: The SEAME database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing* 3(1), 5–17 (2012).
115. Zadeh, A., Chan, M., Liang, P. P., Tong, E., Morency, L.-P.: Social-IQ: A question answering benchmark for artificial social intelligence. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8799–8809 (2019).
116. Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., Narayanan, S. S.: IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation* 42, 335–359 (2008).
117. Bong, S. Z., Murugappan, M., Yaacob, S.: Analysis of electrocardiogram (ECG) signals for human emotional stress classification. In: Ponnambalam, S. G., Parkkinen, J., Ramanaathan, K. C. (eds.), *Trends in Intelligent Robotics and Automation in Manufacturing*, pp. 198–205. Springer (2012).
118. Poria, S., Hazarika, D., Majumder, N., Naik, G., Mihalcea, R., Cambria, E.: MELD: A multimodal multi-party dataset for emotion recognition in conversation (2018).
119. Park, C. Y., Cha, N., Kang, S., Kim, A., Khandoker, A., Hadjileontiadis, L., Oh, A., Jeong, Y., Lee, U.: K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data* 7, 293 (2020).
120. Kosti, R., Álvarez, J. M., Recasens, A., Lapedriza, A.: Context based emotion recognition using the Emotic dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* (2019).
121. Katsigiannis, S., Ramzan, N.: DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE Journal of Biomedical and Health Informatics* 22(1), 98–107 (2018).
122. Subramanian, R., Wache, J., Abadi, M. K., Vieriu, R. L., Winkler, S., Sebe, N.: ASCERTAIN: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing* 9(2), 147–160 (2018).
123. Wen, Z., et al.: Distract your attention: Multi-head cross attention network for facial expression recognition. *arXiv preprint arXiv:2109.07270* (2021).
124. Milton, A., Roy, S. S., Selvi, S. T.: SVM scheme for speech emotion recognition using MFCC feature. *International Journal of Computer Applications* 69, 34–39 (2013).
125. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - EMNLP 02*, 79–86 (2002).
126. Rouast, P. V., Adam, M., Chiong, R.: Deep learning for human affect recognition: Insights and new developments. *IEEE Transactions on Affective Computing* (2019).
127. Fu, Y., Wu, X., Li, X., Pan, Z., Luo, D.: Semantic neighborhood-aware deep facial expression recognition. *IEEE Transactions on Image Processing* 29, 6535–6548 (2020).
128. Conneau, A., Schwenk, H., Barrault, L., Lecun, Y.: Very deep convolutional networks for text classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 1107–1116 (2017).

129. Sahu, S., Gupta, R., Espy-Wilson, C.: Modeling feature representations for affective speech using generative adversarial networks. *IEEE Transactions on Affective Computing* (2020).
130. Chen, M., He, X., Yang, J., Zhang, H.: 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters* 25, 1440–1444 (2018).
131. Sahu, S., Gupta, R., Sivaraman, G., AbdAlmageed, W., Espy-Wilson, C.: Adversarial Auto-Encoders for Speech Based Emotion Recognition. In: *Interspeech 2017*, pp. 1243–1247. ISCA (2017).
132. Li, C., Bao, Z., Li, L., Zhao, Z.: Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition. *Information Processing & Management* 57, 102185 (2020).
133. Yu, J., Zhang, C., Song, Y., Cai, W.: ICE-GAN: Identity-Aware and Capsule-Enhanced GAN with Graph-Based Reasoning for Micro-Expression Recognition and Synthesis. In: *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8 (2021).
134. Poria, S., Cambria, E., Winterstein, G., Huang, G.-B.: Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems* 69, 45–63 (2014).
135. Blekanov, I., Kukarkin, M., Maksimov, A., Bodrunova, S.: Sentiment Analysis for Ad Hoc Discussions Using Multilingual Knowledge-Based Approach. *Proceedings of the 3rd International Conference on Applied Information Technology (ICAIT)*, pp. 117–121. ACM Press (2018).
136. Chen, J., Huang, H., Tian, S., Qu, Y.: Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications* 36, 5432–5436 (2009).
137. Li, D., Rzepka, R., Ptaszynski, M., Araki, K.: A novel machine learning-based sentiment analysis method for Chinese social media considering Chinese slang lexicon and emoticons. *Proceedings of the 3rd International Conference on Applied Information Technology* (2019).
138. Liu, F., Zheng, L., Zheng, J.: HieNN-DWE: A hierarchical neural network with dynamic word embeddings for document-level sentiment classification. *Neurocomputing* 403, 21–32 (2020).
139. Li, W., Zhu, L., Shi, Y., Guo, K., Cambria, E.: User reviews: Sentiment analysis using lexicon integrated two-channel CNN-LSTM family models. *Applied Soft Computing* 94, 106435 (2020).
140. Huang, B., Carley, K.: Parameterized Convolutional Neural Networks for Aspect Level Sentiment Classification. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1091–1096. Association for Computational Linguistics (2018).
141. Xue, W., Li, T.: Aspect Based Sentiment Analysis with Gated Convolutional Networks. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* 2(5), 2514–2523 (2018).
142. Bitouk, D., Verma, R., Nenkova, A.: Class-level spectral features for emotion recognition. *Speech Communication* 2(5), 613–625 (2010).
143. Jin, Y., Song, P., Zheng, W., Zhao, L.: A feature selection and feature fusion combination method for speaker-independent speech emotion recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing* 2(5), 4808–4812 (2014).
144. Chen, L., Su, W., Feng, Y., Wu, M., She, J., Hirota, K.: Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction. *Information Sciences* 2(5), 150–163 (2020).

145. Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M.A., Schuller, B., Zafeiriou, S.: Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. *IEEE International Conference on Acoustics, Speech and Signal Processing* 2(5), 5200–5204 (2016).
146. Zhao, Z., Zhao, Y., Bao, Z., Wang, H., Zhang, Z., Li, C.: Deep Spectrum Feature Representations for Speech Emotion Recognition. *ASMM-C-MMAC18, Association for Computing Machinery* 2(5), 27–33 (2018).
147. Han, J., Zhang, Z., Ren, Z., Ringeval, F., Schuller, B.: Towards Conditional Adversarial Training for Predicting Emotions from Speech. *IEEE International Conference on Acoustics, Speech and Signal Processing* 2(5), 6822–6826 (2018).
148. Yan, Y., Zhang, Z., Chen, S., Wang, H.: Low-resolution facial expression recognition: A filter learning perspective. *Signal Processing* 2(5), 107370 (2020).
149. Makhmudkhujaev, F., Abdullah-Al-Wadud, M., Iqbal, M.T.B., Ryu, B., Chae, O.: Facial expression recognition with local prominent directional pattern. *Signal Processing: Image Communication* 2(5), 1–12 (2019).
150. Zhen, Q., Huang, D., Drira, H., Amor, B.B., Wang, Y., Daoudi, M.: Magnifying Subtle Facial Motions for Effective 4D Expression Recognition. *IEEE Transactions on Affective Computing* 2(5), 524–536 (2019).
151. Zong, Y., Huang, X., Zheng, W., Cui, Z., Zhao, G.: Learning from Hierarchical Spatiotemporal Descriptors for Micro-Expression Recognition. *IEEE Transactions on Multimedia* 2(5), 3160–3172 (2018).
152. Lo, L., Xie, H.-X., Shuai, H.-H., Cheng, W.-H.: MER-GCN: Micro-Expression Recognition Based on Relation Modeling with Graph Convolutional Networks. *IEEE Conference on Multimedia Information Processing and Retrieval* 2(5), 79–84 (2020).
153. Gera, D., Balasubramanian, S.: Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition. *Pattern Recognition Letters* 2(5), 58–66 (2021).
154. Liu, D., Ouyang, X., Xu, S., Zhou, P., He, K., Wen, S.: SAANet: Siamese action units attention network for improving dynamic facial expression recognition. *Neurocomputing* 2(5), 145–157 (2020).
155. Behzad, M., Vo, N., Li, X., Zhao, G.: Towards Reading Beyond Faces for Sparsity aware 3D/4D Affect Recognition. *Neurocomputing* 2(5), 297–307 (2021).
156. Atkinson, J., Campos, D.: Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers. *Expert Systems with Applications* 2(5), 35–41 (2016).
157. Puk, K.M., Wang, S., Rosenberger, J., Gandy, K.C., Harris, H.N., Peng, Y.B., Nordberg, A., Lehmann, P., Tommerdahl, J., Chiao, J.-C.: Emotion Recognition and Analysis Using ADMM-Based Sparse Group Lasso. *IEEE Transactions on Affective Computing* 2(5) (2019).
158. Li, Y., Zheng, W., Cui, Z., Zhang, T., Zong, Y.: A Novel Neural Network Model based on Cerebral Hemispheric Asymmetry for EEG Emotion Recognition. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence* 2(5), 1561–1567 (2018).
159. Gao, Z., Wang, X., Yang, Y., Li, Y., Ma, K., Chen, G.: A Channel-fused Dense Convolutional Network for EEG-based Emotion Recognition. *IEEE Transactions on Cognitive and Developmental Systems* 2(5) (2020).
160. Ferdinando, H., Seppänen, T., Alasaarela, E.: Enhancing Emotion Recognition from ECG Signals using Supervised Dimensionality Reduction. *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods* 2(5), 112–118 (2017).

161. Cheng, Z., Shu, L., Xie, J., Chen, C.L.P.: A novel ECG-based real-time detection method of negative emotions in wearable applications. *International Conference on Security, Pattern Analysis, and Cybernetics* 2(5), 296–301 (2017).
162. Chen, G., Zhu, Y., Hong, Z., Yang, Z.: EmotionalGAN: Generating ECG to Enhance Emotion State Classification. *Proceedings of the 2019 International Conference on Artificial Intelligence and Computational Science* 2(5), 115–122 (2019).
163. Elakkiya, M., Vimal, R., Udayakumar, V., Suresh, R.: ECG Signal Processing Based Emotion Recognition for Healthcare. *Informatics in Medicine Unlocked* 2(5), 100560 (2021).
164. Zhang, Q., Wang, X., Wang, Z., He, D., Sun, L.: A Novel Hybrid Model for Emotion Recognition from ECG and Respiratory Signals. *IEEE Transactions on Biomedical Engineering* 2(5), 1424–1434 (2021).
165. Shang, Y., Fu, T.: Multimodal fusion: A study on speech-text emotion recognition with the integration of deep learning. *Intelligent Systems with Applications* 2(5), 200436 (2023).
166. Hazmoune, S., Bougamouza, F.: Using transformers for multimodal emotion recognition: Taxonomies and state of the art review. *Engineering Applications of Artificial Intelligence* 2(5), 108339 (2024).
167. Gan, C., Fu, X., Feng, Q., Zhu, Q., Cao, Y., Zhu, Y.: A multimodal fusion network with attention mechanisms for visual–textual sentiment analysis. *Expert Systems with Applications* 2(5), 122731 (2024).
168. Fan, Q., Zuo, H., Liu, R., Lian, Z., Gao, G.: Learning Noise-Robust Joint Representation for Multimodal Emotion Recognition under Incomplete Data Scenarios. *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing* 2(5), 116–124 (2024).
169. Zhang, H., Wang, X., Xu, H., Zhou, Q., Gao, K., Su, J., Zhao, J., Li, W., Chen, Y.: MInt-Rec2.0: A Large-scale Benchmark Dataset for Multimodal Intent Recognition and Out-of-scope Detection in Conversations. *arXiv Preprint* 2(5) (2024).
170. Lin, Z.: Fine-Tuning Methods of Multimodal Pretraining Models for Emotion Recognition. *Applied and Computational Engineering* 2(5), 86–92 (2024).

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

