


Review

Emotion Recognition Using Different Sensors, Emotion Models, Methods and Datasets: A Comprehensive Review

Yujian Cai , Xingguang Li * and Jinsong Li

School of Electronic Information Engineering, Changchun University of Science and Technology, Changchun 130022, China

* Correspondence: leexingguang@126.com; Tel.: +86-155-2689-3169

Abstract: In recent years, the rapid development of sensors and information technology has made it possible for machines to recognize and analyze human emotions. Emotion recognition is an important research direction in various fields. Human emotions have many manifestations. Therefore, emotion recognition can be realized by analyzing facial expressions, speech, behavior, or physiological signals. These signals are collected by different sensors. Correct recognition of human emotions can promote the development of affective computing. Most existing emotion recognition surveys only focus on a single sensor. Therefore, it is more important to compare different sensors or unimodality and multimodality. In this survey, we collect and review more than 200 papers on emotion recognition by literature research methods. We categorize these papers according to different innovations. These articles mainly focus on the methods and datasets used for emotion recognition with different sensors. This survey also provides application examples and developments in emotion recognition. Furthermore, this survey compares the advantages and disadvantages of different sensors for emotion recognition. The proposed survey can help researchers gain a better understanding of existing emotion recognition systems, thus facilitating the selection of suitable sensors, algorithms, and datasets.

Keywords: sensors for emotion recognition; emotion models; emotional signal processing; classifiers; emotion recognition datasets



Citation: Cai, Y.; Li, X.; Li, J. Emotion Recognition Using Different Sensors, Emotion Models, Methods and Datasets: A Comprehensive Review. *Sensors* **2023**, *23*, 2455. <https://doi.org/10.3390/s23052455>

Academic Editor: Wataru Sato

Received: 31 January 2023

Revised: 18 February 2023

Accepted: 21 February 2023

Published: 23 February 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Emotion is a comprehensive manifestation of people's physiological and psychological states; emotion recognition was systematically proposed in the 1990s [1]. With the rapid development of science and technology, emotion recognition has been widely used in various fields, such as human-computer interactions (HCI) [2], medical health [3], Internet education [4], security monitoring [5], intelligent cockpit [6], psychological analysis [7], and the entertainment industry [8].

Emotion recognition can be realized through different detection methods and different sensors. Sensors are combined with advanced algorithm models and rich data to form human-computer interaction systems [9,10] or robot systems [11]. In the field of medical and health care [12], emotion recognition can be used to detect the patient's psychological state or adjuvant treatment, and improve medical efficiency and medical experience. In the field of Internet education [13], emotion recognition can be used to detect students' learning status and knowledge acceptance, and cooperate with relevant reminders to improve learning efficiency. In the field of criminal interrogation [14], emotion recognition can be used to detect lies (authenticity test). In the field of intelligent cockpits [15], it can be used to detect the drowsiness and mental state of the driver to improve driving safety. In the field of psychoanalysis [16], it can be used to help analyze whether a person has autism. This technique can also be applied to recognize the emotions of the elderly, infants, and those with special diseases who cannot clearly express their emotions [17,18].

A correct understanding of emotion can deepen the research on emotion recognition. Section 2 introduces the definition of emotion and famous emotion models. Each sensor has different detection emphases for emotion recognition, which can be roughly divided into three categories: The first one is to use human actions or speech signals, such as facial expressions, speech, and gestures. The second one is to use the physiological signals inside the human body, such as EEG, respiratory rate, and heart rate. The last type is multi-modal fusion emotion recognition, which uses multiple signals for emotion recognition. These three types of detection methods have their own advantages and disadvantages, which are detailed in Section 3. Preprocessing, feature extraction, and classification methods for different sensor signals are detailed in Sections 4 and 5. Section 6 presents some of the main datasets for different signals. Sections 7 and 8 are the conclusion of the overall survey and thoughts on the future development of emotion recognition.

2. Emotion Models

The definition of emotion is the basis of emotion recognition. The basic concept of emotion was proposed by Ekman in the 1970s [19]. At present, there are two mainstream emotion models: the Discrete emotion model and the dimensional emotion model.

Discrete Emotion Model

Darwinian evolution [20] holds that emotions are primitive or fundamental. Emotion as a form is considered to correspond to discrete and elementary responses or tendencies of action. The discrete emotion model divides human emotions into limited categories [21], mainly including happiness, sadness, fear, anger, disgust, surprise, etc. There are two to eight basic emotions, according to different theories. However, these discrete emotion model theories have certain common features. They believe that emotions are: mental and physiological processes; caused by the awareness of developmental events; inducing factors for changes in the body's internal and external signals; related to a fixed set of actions or tendencies. Ekman proposed seven characteristics to distinguish different basic emotions and emotional phenomena: autonomous evaluation; have specific antecedent events; also present in other primates; rapid onset; short duration; unconscious or involuntary appearance; reflected in unique physiological systems such as the nervous system and facial expressions. R. Plutchik proposed eight basic emotions and distinguished them according to intensity, forming the Plutchik's wheel model [22]. It is a well-known discrete emotion model, as shown in Figure 1 (adapted from [22]).

Dimensional emotion models view emotions as combinations of vectors within a more fundamental dimensional space. This enables complex emotions to be researched and measured in fewer dimensions. Core emotions are generally expressed in two-dimensional or three-dimensional space. The dimensional emotion model in the two-dimensional space is usually the arousal-valence model. Valence reflects the positive or negative evaluation of an emotion and the degree of pleasure the participant feels. Arousal reflects the intensity or activation of an emotion in the body. The level of arousal reflects the individual's will, and low arousal means less energy. However, dimensional emotion models in two dimensions were not able to successfully distinguish core emotions with the same degree of consistency and valence. For example, both anger and fear have high arousal and low valence. Therefore, a new dimension needs to be introduced to distinguish these emotions.

The most famous three-dimensional emotion model is the pleasure, arousal, and dominance (PAD) model [23] proposed by Mehrabian and Russell through the study of environmental psychology methods [24] and the feeling-thinking-acting [25] model, as shown in Figure 2 (adapted from [23]).

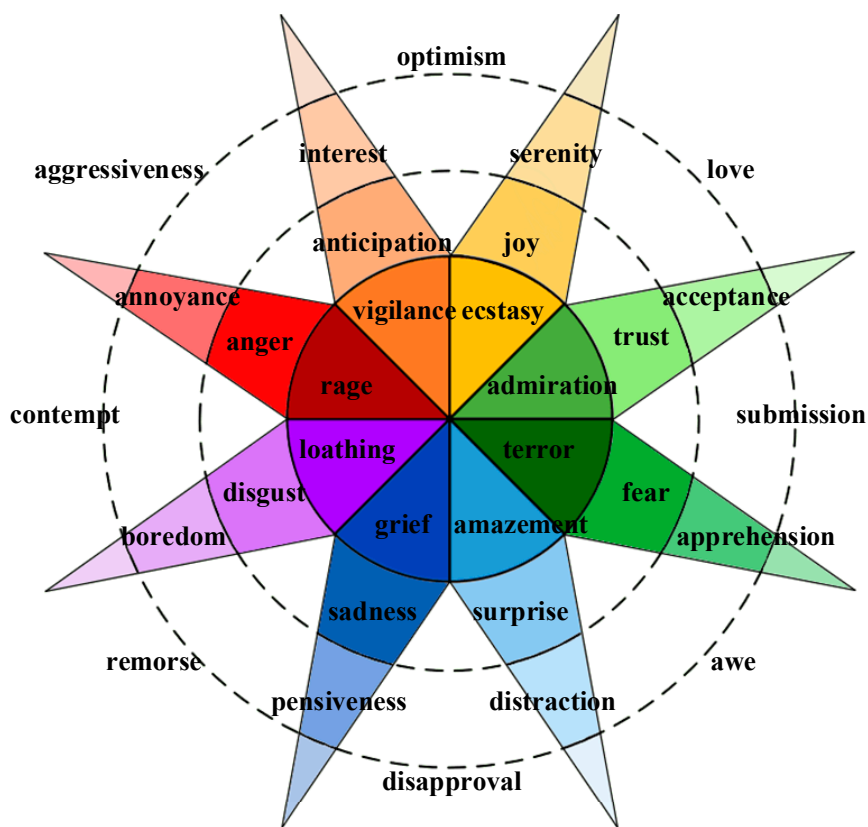


Figure 1. Piutchik's wheel model.2.2. Dimensional Emotion Model.

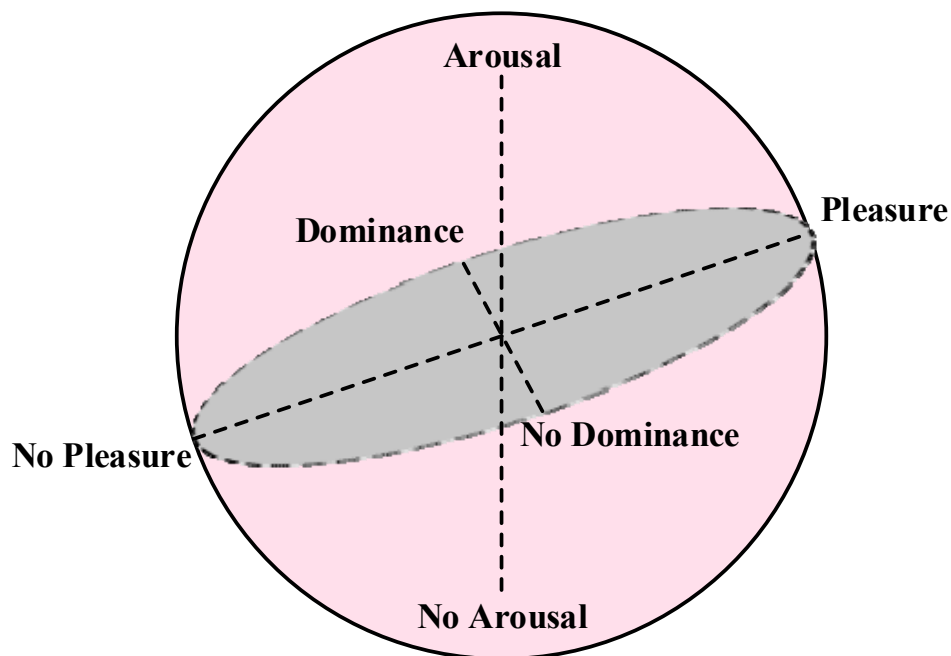


Figure 2. PAD 3D emotion model.

Dominance represents control or position, and indicates whether a certain emotion is submissive. It is worth noting the dimensional emotion model can accurately identify the core emotion. However, for some complex emotions, the dimensional emotion model will lose some details.

3. Sensors for Emotion Recognition

The sensors used for emotion recognition mainly include visual sensors, audio sensors, radar sensors, and other physiological signal sensors, which can collect signals of different dimensions and achieve emotional analysis through some algorithms. Different sensors have different applications in emotion recognition. The advantages and disadvantages of different sensors for emotion recognition are shown in Table 1.

Table 1. Advantages and disadvantages of different sensors for emotion recognition.

Sensors	Advantages	Disadvantages
Visual sensor	Simple data collection; high scalability	Restricted by light; easy to cause privacy leakage [26]
Audio sensor	Low cost; wide range of applications	Lack of robustness for complex sentiment analysis
Radar sensor	Remote monitoring of physiological signals	Radial movement may cause disturbance
Other physiological sensors	Ability to monitor physiological signals representing real emotion	Invasive, requires wearing close to the skin surface [27]
Multi-sensor fusion	Richer collected information; higher robustness	Multi-channel information needs to be synchronized; the follow-up calculation is relatively large

3.1. Visual Sensor

Emotion recognition based on visual sensors is one of the most common emotion recognition methods. It has the advantages of low cost and simple data collection. At present, visual sensors are mainly used for facial expression recognition (FER) [28–30] to detect emotion or remote photoplethysmography (rPPG) technology to detect heart rate [31,32]. The accuracies of these methods severely drop as the light intensity decreases.

The facial expression recognition process is shown in Figure 3. Facial expressions can intuitively reflect people's emotions. It is difficult for machines to capture the details of expressions like humans [33]. Facial expressions are easy to hide, which leads to emotion recognition errors [34]. For example, in some social activities, we usually politely smile even though we are not in a happy mood [35].



Figure 3. Facial expression recognition process.

Different individuals have different skin colors, looks, and facial features [36,37], which pose challenges to the accuracy of classification. Facial features of the same emotion can be different, and small changes in different emotions of the same individual are not very obvious [38]. Therefore, there is a classification challenge of large intra-class distance and small inter-class distance for emotion detection through facial expression recognition by the camera. It is also difficult to effectively recognize emotions when the face is occluded (wearing a mask) or from different shooting angles [39].

Photoplethysmography (PPG) is an optical technology for the non-invasive detection of various vital signs, which was first proposed in the 1930s [40]. PPG is widely used in the detection of physiological signals in personal portable devices (smart wristbands, smart watches, etc.) [41,42]. The successful application of PPG has led to the rapid development

of remote photoplethysmography (rPPG). A multi-wavelength RGB camera is used by rPPG technology to identify minute variations in skin color on the human face caused by changes in blood volume during a heartbeat [43], as shown in Figure 4.

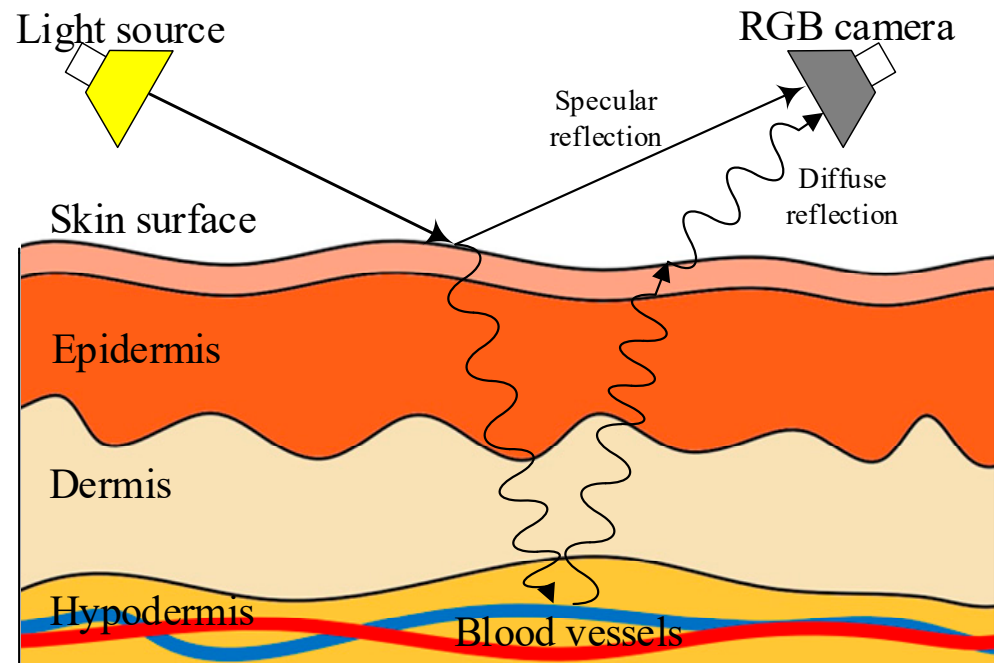


Figure 4. Schematic diagram of rPPG technology.

The rPPG technology can be used to obtain the degree of peripheral vascular constriction and analyze the participant's emotions. External vasoconstriction is considered to be a defensive physiological response. When people are in a state of pain, hunger, fear, or anger, the constriction of external blood vessels will be enhanced. Conversely, in a calm or relaxed state, this response will reduce.

With the improvement in hardware and algorithm level, rPPG technology can also realize remote non-contact monitoring and estimation of heart rate [44], respiratory rate [45], blood pressure [46], or other signals. Emotion recognition is performed after analyzing a large amount of monitoring data. These signals can classify emotions into a few types and intensities. There are certain errors in the recognition of multiple types of emotions. It is necessary to combine other physiological information to improve the accuracy rate of emotion recognition [47].

3.2. Audio Sensor

Language is one of the most important components of human culture. People can express themselves or communicate with others through language. Speech recognition [48] has promoted the development of speech emotion recognition (SER) [49]. Human speech contains rich information that can be used for emotion recognition [50,51]. Understanding the emotion in information is essential for artificial intelligence to engage in effective dialog. SER can be used for call center dialog, automatic response systems, autism diagnosis, etc. [52–54]. SER is jointly completed by acoustics feature extraction [55] and language mark [56]. The process of SER is shown in Figure 5.

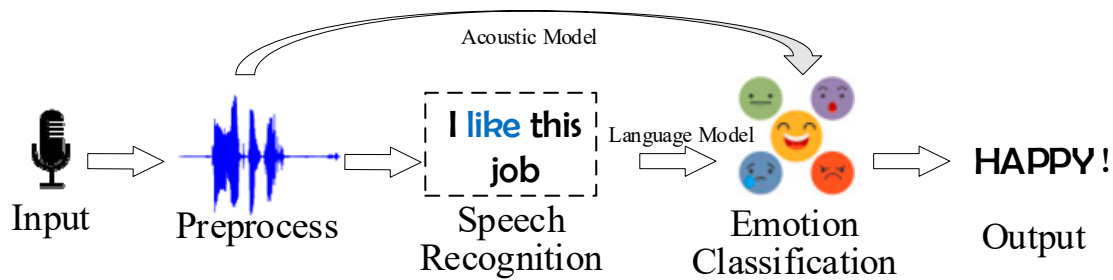


Figure 5. The process of SER.

In the preprocessing stage, the input signal is enhanced into segmentations [57] after noise reduction; and then feature extraction and classification are performed [58]. The language model [59] can identify emotional expressions with specific semantic contributions. The acoustic model can distinguish different emotions contained in the same sentence by analyzing the features of prosody or spectral [60]. Combining these two models can improve the accuracy of SER.

Understanding the emotion in speech is a complex process. Different speaking styles of different people will bring about acoustic variability, which will directly affect speech feature labeling and extraction [61]. The same sentence may contain different emotions [62], and some specific emotional differences often depend on the speaker's local culture or living environment, which also pose challenges for SER.

3.3. Radar Sensor

Different emotions will cause a series of physiological responses, such as changes in respiratory rate [63], heart rate [64,65], brain wave [66], blood pressure [67], etc. For example, the excitement caused by happiness, anger, or anxiety can lead to an increased heart rate [68]. Positive emotions can increase respiratory rate, and depressive emotions can tend to inhibit breathing [69]. Respiratory rate also affects heart rate variability (HRV), which decreases when exhaling and increases when inhaling [70]. Currently, radar technology is widely used in remote vital signs detection [71] and wireless sensing [72]. Radar sensors can use the echo signal of the target to analyze the chest micro-motion caused by breathing and heartbeats. It can realize remote acquisition of these physiological signals. The overall process of emotion recognition based on radar sensor is shown in Figure 6.

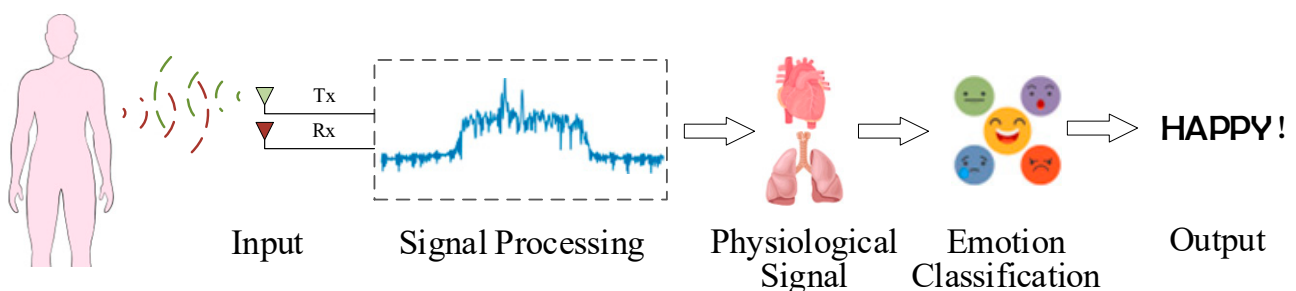


Figure 6. Process of emotion recognition based on radar sensor.

Compared with visual sensors, emotion recognition based on radar sensors is unrestricted by light intensity [73]. However, in real environments, radar echo signals are affected by noise, especially for the radial doppler motion close to or away from the radar [74], which affects the accuracy of sentiment analysis.

3.4. Other Physiological Sensors

Emotions have been shown to be biological since ancient times. Excessive emotion is believed to have some effects on the functioning of vital organs. Aristotle believed that

the influence of emotions on physiology is reflected in the changes in physiological states, such as a rapid heartbeat, body heat, or loss of appetite. William James first proposed the theory of the physiology of emotion [75]. He believed that external stimuli would trigger activity in the autonomic nervous system and create a physiological response in the brain. For example, when we feel happy, we laugh; when we feel scared, our hairs stand on end; when we feel sad, we cry.

Human emotion is a spontaneous mental state, which is reflected in the physiological changes of the human body and significantly affects our consciousness [76]. Many other physiological signals in the human body, such as electroencephalogram (EEG) [66], electrocardiogram (ECG) [77], electromyogram (EMG) [78], galvanic skin response (GSR) [79], blood volume pulse (BVP) [80], and electrooculography (EOG) [81], as shown in Figure 7.

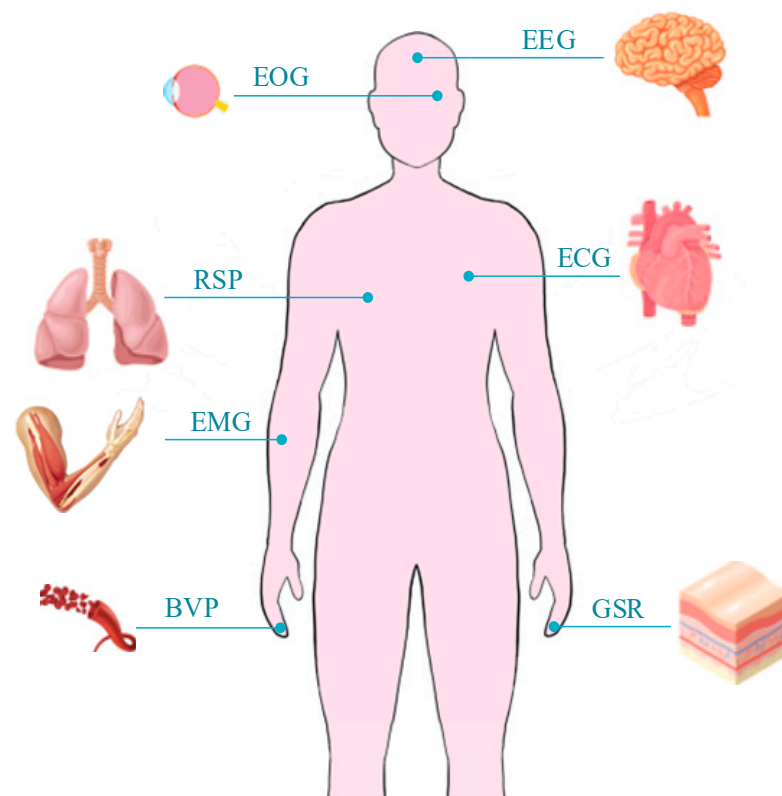


Figure 7. Physiological signals detected by other physiological sensors.

EEG measures the electrical signal activity of the brain by setting electrodes on the skin surface of the head. Many studies have shown that the prefrontal cortex, temporal lobe, and anterior cingulate gyrus of the brain are related to the control of emotions. Their levels of activity induce emotions such as anxiety, irritability, depression, worry, and resentment, respectively.

ECG is a method of electrical monitoring on the surface of the skin that detects the heartbeat controlled by the body's electrical signals. Heart rate and heart rate variability obtained through subsequent analysis are widely used in affective computing. Heart rate and heart rate variability are controlled by the sympathetic nervous systems and parasympathetic nervous systems. The sympathetic nervous system can speed up the heart rate, which is reflected in greater psychological stress and activation. The parasympathetic nervous system is responsible for bringing the heart rate down to normal levels, putting the body in a more relaxed state.

EMG measures the degree of muscle activation by collecting the voltage difference generated during muscle contraction. The current EMG signal measurement technology

can be divided into two types. The first is to study facial expressions by measuring facial muscles. The second is to place electrodes on the body to recognize emotional movements.

GSR is another signal commonly used for emotion recognition. Human skin is normally an insulator. When sweat glands secrete sweat, the electrical conductivity of the skin will change. Therefore, GSR can reflect the sweating situation of a person. GSR is usually measured on the palms or soles of the feet, where sweat glands are thought to best reflect changes in emotion. When a person is in an anxious or tense mood, the sweat glands usually secrete more sweat, which causes a greater change in current.

Related physiological signals also include BVP, EOG, etc. These signals all change with emotional changes, and they are not subject to human conscious control [82]. Therefore, these signals can be measured by different physiological sensors to achieve the purpose of emotion recognition. Using these physiological sensors can accurately and quickly obtain real human physiological signals. However, physiological sensors other than visual, audio, and radar sensors usually need to touch the skin or wear related equipment to extract physiological signals, which will affect people's daily comfort (most people will not accept this monitoring method). Contact sensors are limited by weight and size [27]. These contact devices may also cause people tension and anxiety, which will affect the accuracy of emotion recognition.

3.5. Multi-Sensor Fusion

There are certain deficiencies in single-modal emotion recognition, and it is usually unable to accurately identify complex emotions. The multi-modal emotion recognition method refers to the use of signals obtained by multiple sensors to complement each other and obtain more reliable recognition results. Multi-modal approaches can promote the development of emotion recognition. Multi-modal emotion recognition can often achieve the best recognition performance, but the computational complexity will increase due to the excessive number of channels. There are higher requirements for the collection of multi-modal datasets. Multi-modal emotion recognition has different fusion strategies, which can be mainly divided into pixel-level fusion, feature-level fusion, and decision-level fusion.

Pixel-level fusion [83] refers to the direct fusion of the original data; the semantic information and noise of the signal will be superimposed, which will affect the classification effect after fusion. Processing time is wasted when there is too much redundant information.

The feature-level fusion [84] process is shown in Figure 8. Feature-level fusion occurs in the early stages of the fusion process. Extract features from different input signals and combine them into high-dimensional feature vectors. Finally, output the result through a classifier. Feature-level fusion retains most of the important information, it can greatly reduce computing consumption. However, when the amount of data is small or some details are missing, the final accuracy rate will decrease.

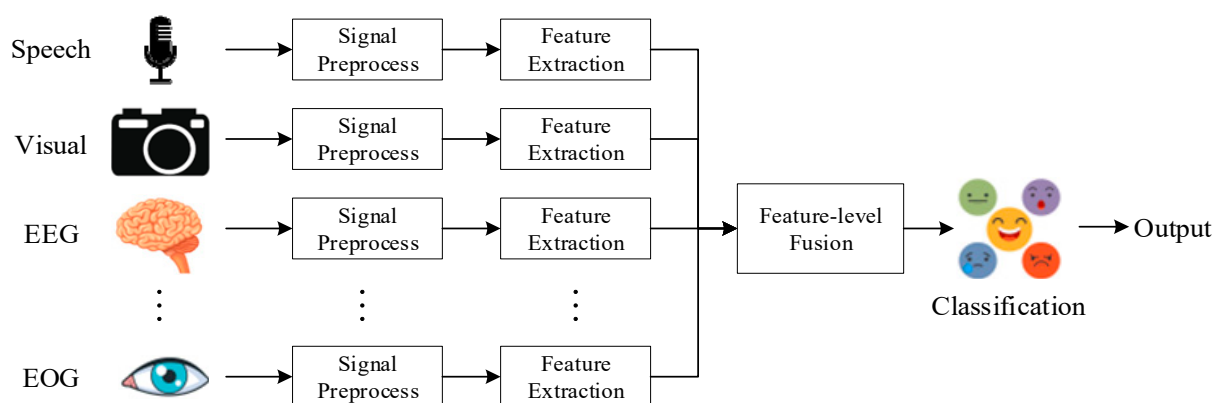


Figure 8. Feature-level fusion.

The decision-level fusion [85,86] process is shown in Figure 9. Decision-level fusion refers to the fusion of independent decisions of each part after making independent decisions based on signals collected by different sensors.

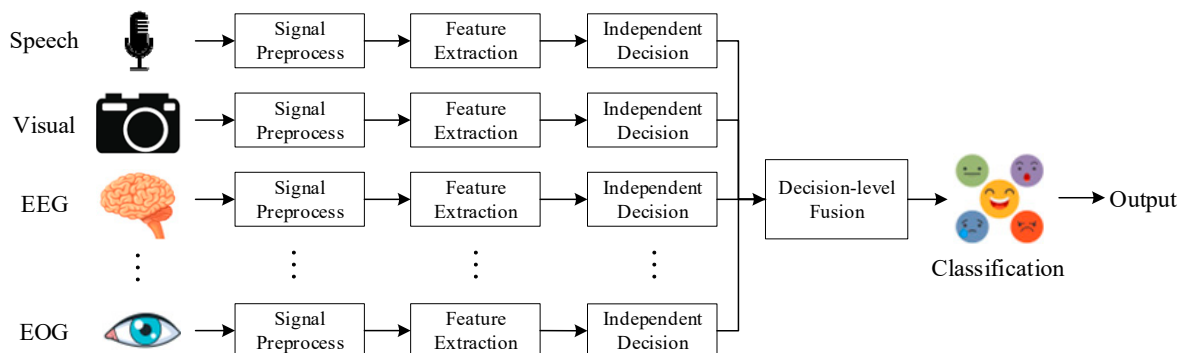


Figure 9. Decision-level fusion.

The advantage of decision-level fusion is that independent feature extraction and classification methods can be set according to different signals. It has lower requirements for the integrity of multimodal data. Decision-level fusion has higher robustness and better classification results.

4. Emotion Recognition Method

Choosing the right method can improve the accuracy of emotion recognition [87]. The emotion recognition method of different sensors is described in Figure 10.

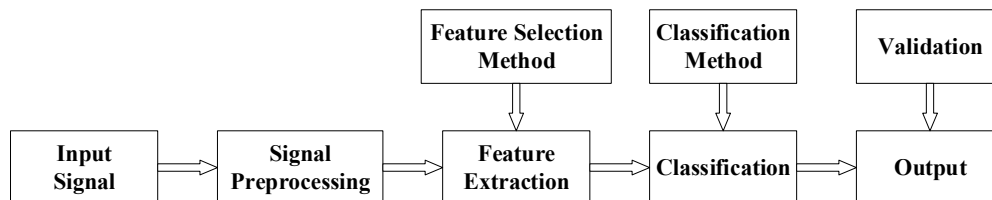


Figure 10. Process of emotion recognition method.

Signal preprocessing refers to improving signal quality and reducing noise. Feature extraction is mainly used to find the characteristics of different signals and reduce the amount of calculation required for classification. Classification refers to applying the extracted features to a certain classification model. Finally, the emotion corresponding to the signal is obtained through analysis.

4.1. Signal Preprocessing

For emotion recognition from different sensors, signal preprocessing is an important step [88]. Preprocessing can reduce the impact of noise in the early stages of emotion recognition.

For visual signals, mainly use cropping, rotation, scaling, grayscale, and other methods for signal preprocessing.

The signal preprocessing method for audio signals mainly includes:

- Silent frame removal: Remove frames below the set threshold to reduce calculation consumption [89];
- Pre-emphasis: Compensate for high-frequency components of the signal;
- Regularization: Adjust the signal to a standard level to reduce the influence of different environments on the results;
- Window: Prevent signal edge leakage from affecting feature extraction [90];
- Noise reduction algorithm: Use noise reduction algorithms such as minimum mean square error (MMSE) to reduce background noise.

The preprocessing methods for radar signals and physiological signals mainly include:

- Filtering: Remove noise, signal crosstalk [91], or baseline wander [92] by different filters;
- Wavelet transform [93]: Using time window and frequency window to characterize the local characteristics of physiological signals;
- Nonlinear dynamics: Use approximate entropy [94], sample entropy [95], transfer entropy [96] to obtain a smooth signal estimate and remove transient disturbances [97].

4.2. Feature Extraction

Feature extraction can ignore information irrelevant to the target, reduce the amount of calculation, overcome the curse of dimensionality, and improve the generalization ability of the model. Signals often require feature extraction before being input into some classical classification models.

4.2.1. For Visual Signals

Principal Component Analysis (PCA) [98]: PCA is a very common dimensionality reduction method. Keeping the most important features of high-dimensional data while removing noise and unimportant features. This can greatly improve data processing speed and save time and costs. PCA can be defined as an orthogonal linear transformation that projects data to a new coordinate system. PCA can satisfy the maximum reconfiguration, which means that the distance between the sample point and the hyperplane is close enough. At the same time, PCA can also satisfy maximum separability, which means that the projection of sample points onto the hyperplane can be separated as much as possible.

In [99], the authors simply used PCA to reduce the dimensionality of the feature vectors. The accuracy rate on the JAFFE dataset reached 74.14%. In [100], the authors combined PCA and PSO to obtain optimized feature vectors. The accuracy rate on the JAFFE dataset reached 94.97%. In [101], the authors proposed two-dimensional PCA; 2DPCA is based on 2D image matrices instead of 1D vectors, so there is no need to convert image matrices to vectors before feature extraction. Indeed, 2DPCA can directly use the original image to construct the covariance matrix, which is more effective than PCA. In [102], the authors utilized bidirectional PCA to extract visual features. The accuracy rate on the YALE multimodal dataset reached 94.01, which was an increase of 0.9% compared with the PCA method.

Histogram of Oriented Gradients (HOG) [103]: HOG was proposed based on image edge information and was first used for object detection. Each window region of an image can be described by the local distribution and gradient of edge directions. A HOG descriptor can be obtained by computing the histogram of edge directions in these cells and normalizing them. Combining these descriptors can be used to detect facial expressions. The features generated by HOG are not affected by illumination and geometric transformation.

In [104], the authors proposed a framework for emotion recognition based on HOG and SVM. The accuracy rate on the GEMEP-FERA dataset reached 70%. In [105], the authors proposed a FER framework for real-time inference of emotional states. The framework extracted HOG features from active face patches; 95% accuracy was achieved on the CK+ dataset. In [106], the authors proposed an emotion recognition framework based on HOG descriptors and the Cuttlefish algorithm. This method did not generate irrelevant or noisy features. The model achieved 97.86%, 95.15%, and 90.95% accuracy on the CK+, RaFD, and JAFFE datasets.

Other feature extraction methods for visual signals include Local Binary Patterns (LBP) [107] and Linear Discriminate Analysis (LDA) [108].

4.2.2. For Speech Signals

Speech signal features mainly include prosodic features [109], frequency spectral features, frequency cepstral coefficients [110], and energy features. These features carry both information and emotion. Therefore, some methods can be utilized to extract them.

Linear Predictor Coefficients (LPC) [111]: LPC is based on a speech production model. This model uses an all-pole filter to model the characteristics of the vocal tract. LPC is equivalent to the smooth envelope of the speech logarithmic spectrum. It can be directly computed from windowed parts of speech by autocorrelation or covariance methods. LPC can accurately and quickly estimate speech parameters.

In [112], the authors combined the features of TEO and LPC for T-LPC feature extraction. This method can accurately recognize stress speech signals. The accuracy on the Emo-DB dataset reached 82.7% (male) and 88% (female). In [113], the authors proposed a combined spectral coefficient optimization method based on LPC. The accuracy on the Emo-DB dataset reached 88%. Comparative experiments showed that this optimization method improved the accuracy by 4%. In [114], the authors measured the emotion recognition accuracy when LPC coefficients were introduced in the feature vectors. Using only the LOC coefficients, the model achieved 78% accuracy on the SROL dataset. In [115], the authors proposed a meta-heuristic feature selection model. This model took LPC features as input. The accuracy of the model on the SAVEE and Emo-DB datasets reached 97.31% and 98.46%.

Teager Energy Operator (TEO) [116]: TEO is a powerful nonlinear energy operator. It is able to extract signal energy based on mechanical and physical considerations. TEO can extract the features of speech when the utterance presents a certain stress. It measures speech non-proximity by processing the characteristics of speech signals in the frequency and time domains.

In [117], the authors proposed a two-stage emotion recognition system based on TEO. Autoencoders improved recognition rates. The accuracy on the RML dataset reached 74.07%. In [118], the authors proposed the EMD-TEO model. Experiments showed that the features extracted based on TEO were robust, and the performance of speech emotion recognition was significantly improved. The accuracy of the model on the EMO-DB dataset reached 81.34%. In [119], the authors fused TEO and MFCC to form T-MFCC feature extraction technology. TEO extracted the nonlinear features of speech and was mainly used to identify stressful emotions. Experiments showed that T-MFCC had better performance. The accuracy of the model on the EMO-DB dataset reached 93.33%.

Other commonly used speech signal feature extraction methods include Short-time Coherence (SMC), Fast Fourier Transform (FFT), Principal Component Analysis (PCA) [120], and linear discriminant analysis (LDA) [121].

4.2.3. For Physiological and Radar signals

Fast Fourier Transform (FFT) [122]: FFT is a popular signal processing method. It can be used to convert time-domain signals to frequency-domain signals, and vice versa. For spectrum analysis, the magnitude squared of the FFT is usually used to obtain the Power Spectral Density (PSD). PSD can be used to analyze the contribution of a specific frequency band to the total power of the signal.

In [123], the authors utilized FFT to analyze short-duration EEG signals for emotion classification. Through experiments, it was concluded that the short-term EEG signal characteristics reflected the changes in emotional state. The accuracy on the self-built dataset reached 91.33%. In [124], the authors built an emotion recognition model based on FFT and Genetic Programming (GP). FFT was used to convert a signal from the time domain to the frequency domain. The accuracy on the self-built dataset (four emotions) reached 89.14%. In [125], the authors utilized FFT and Wigner-Ville Distribution (WVD) methods to convert physiological signals into images. Putting the image into a CNN model could obtain excellent classification results. The accuracy on the self-built dataset reached 93.01%.

Maximal-Relevance Minimal-Redundancy (mRMR) [126]: mRMR uses mutual information as a correlation measurement with maximum dependence criterion and minimum redundancy criterion. It is capable of selecting features with the strongest correlation

with the categorical variable. The mRMR algorithm can not only reduce dimensions and improve prediction accuracy, but also obtain features with more meaning and value.

In [127], the authors researched stable patterns over time for emotion recognition from EEG. The model used the mRMR algorithm to reduce the dimension and improve the stability of the classifier. The accuracy on the DEAP dataset and SEED dataset reached 69.67% and 91.07%. In [128], the authors proposed a method that combined the feature selection task of mRMR and kernel classifiers for emotion recognition. The authors used mRMR to incorporate feature selection tasks into classification tasks. The accuracy on the DEAP dataset reached 60.7% (Arousal) and 62.33% (Valence). In [129], the authors analyzed non-stationary physiological signals and extracted features that could be used to achieve accurate emotion recognition. The authors utilized the mRMR algorithm to reduce the dimensionality of the constructed feature vectors. The average accuracy on the DEAP dataset reached 80%.

Other feature extraction methods of physiological signals and radar signals include Empirical Mode Decomposition (EMD), Linear Discriminate Analysis (LDA) [130], Locality Preserving Projections (LPP) [131], and the Relief-F algorithm [132].

5. Classification

The classifier can classify different input signals and output the corresponding emotion category. The quality of the classifier will affect the accuracy of emotion recognition. The current classification methods can be divided into two categories: Classical machine learning methods and deep learning methods. This section will introduce several commonly used machine learning methods and deep learning methods.

5.1. Machine Learning Methods

5.1.1. SVM

Support vector machine (SVM) [133] aims to find the hyperplane with the largest interval in the sample space to produce more robust classification results, as shown in Figure 11. For more complex samples, it can be mapped from the original space to a higher dimensional space. Solving the corresponding kernel function [134] makes these samples linearly separable in the feature space. Soft margins [135] and regularization can be added to prevent overfitting of the trained model.

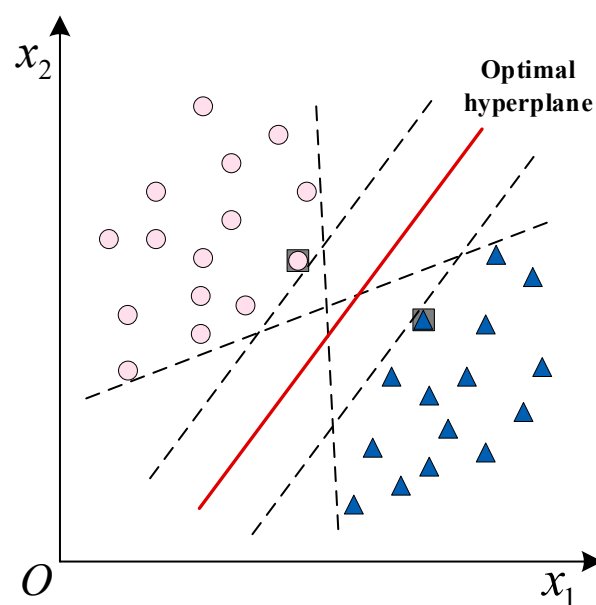


Figure 11. Support vectors in two-dimensional space.

The hyperplane can be described by $w^T x + b$. $w = (w_1; w_2; \dots; w_d)$ is the normal vector, which determines the direction of the hyperplane. b is the displacement term, which determines the distance between the hyperplane and the origin. The distance from any point x in the sample space to the hyperplane (w, b) is

$$r = \frac{|w^T x + b|}{\|w\|} \quad (1)$$

In order to find the optimal plane, the sum of the distances from each support vector to the hyperplane needs to be minimum, so it is only necessary to maximize $\|w\|^{-1}$.

In [136], the author used SVM to achieve a classification accuracy of 93.75% on the Berlin Emotion speech dataset. In [137], the author used the SVM model trained by the LDC dataset and the Emo-DB dataset to achieve an accuracy rate of 83.1% in SER based on seven emotions. In [76], the author used SVM to classify EEG signals, and achieved 85% classification accuracy. In [138], the author used SVM to perform FER on the CK+ dataset and reached 91.95% accuracy. In [139], the author used binary-SVM to realize text sentiment classification based on 15 types of emotions, and the F-score was as high as 68.86%.

5.1.2. GMM

GMM aims to classify data by superimposing Gaussian distributions in a linear combination and formalize them into a probability model, as shown in Figure 12.

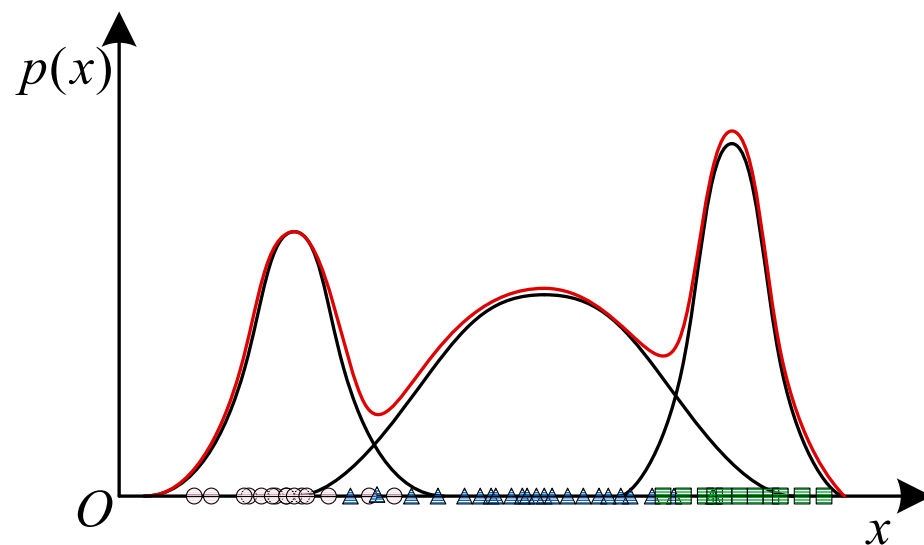


Figure 12. GMM classification.

GMM is an unsupervised learning method, which usually uses expectation maximization (EM) [140] to determine the parameters of GMM, the main processes of EM are:

1. Expectation: Infer optimal latent variables from the training set;
2. Maximization: Use maximum likelihood estimation of parameters based on observed variables. It can obtain a mixture model of probabilities of all sub-distribution contained in the overall distribution. In this way, a better classification effect can be achieved without pre-determining the label of the data.

In [141], the authors achieved 82.5% accuracy on mixed gender SER by SVM based on GMM super vectors. In [142], the author used the GMM-DNN model to achieve a classification accuracy of 83.97% for six emotions. In [143], the author proposed a GMM-based federated learning framework and fully considered the privacy issues in face monitoring data, and achieved 84.1% and 74.39% accuracy for the EmotionNet and SFEW datasets.

5.1.3. HMM

The Hidden Markov model (HMM) is a dynamic Bayesian network with a simple structure, which can estimate and predict unknown variables based on some observed data, as shown in Figure 13. HMM can efficiently improve the matching degree between the evaluation model and the observation sequence. HMM is able to infer hidden model states from observation sequences and better describe observed data.

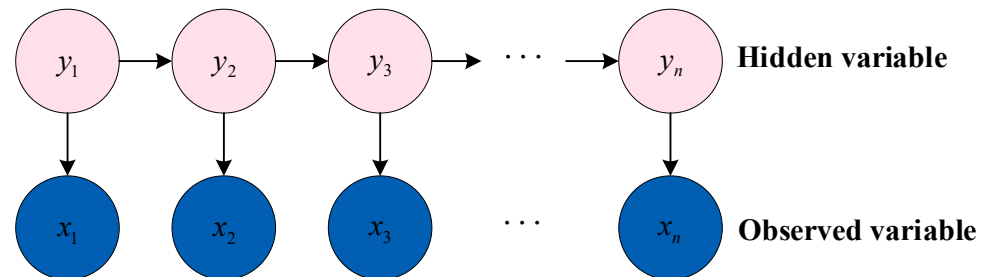


Figure 13. Structure of HMM.

The hidden variables (state variables) of the HMM can be expressed as $\{y_1, y_2, \dots, y_n\}$, so the state space of hidden variables contains N possible values. Observed variables can be described as $\{x_1, x_2, \dots, x_n\}$, and it is usually assumed that the value range of the observed variable is $\{o_1, o_2, \dots, o_M\}$. The system usually transitions between multiple states $\{s_1, s_2, \dots, s_N\}$.

The state transition probability of the model between each state is:

$$a_{ij} = P(y_{t+1} = s_j | y_t = s_i) \quad 1 \leq i, j \leq N \quad (2)$$

The observed probability is:

$$b_{ij} = P(x_t = o_j | y_t = s_i) \quad 1 \leq i \leq N, 1 \leq j \leq M \quad (3)$$

The initial state probability is:

$$\pi_i = P(y_1 = s_i) \quad 1 \leq i \leq N \quad (4)$$

At any moment, the value of the observed variable only depends on the state variable. The state variable y_t at time t is unrelated to y_{t-2} and only depends on the state variable y_{t-1} at time $t - 1$. Based on this dependence, the joint probability distribution of all variables is:

$$P(x_1, y_1, x_2, y_2, \dots, x_n, y_n) = \pi_1 b_{11} \prod_{i=2}^n a_{ij} b_{ij} \quad (5)$$

According to the above parameters, an HMM can be determined.

In [144], the authors used the HMM to classify six types of emotions for person-dependent and person-independent facial expressions, and achieved 82.46% and 58% accuracy. In [145], the authors used continuous HMMs to fully utilize low-level temporal features, and, in the SER of seven emotions, the accuracy rate was 86%. In [146], the authors developed an HMM-based audiovisual model that improved emotion recognition performance for visual and auditory signals in noisy environments. The accuracy rate of multi-modal emotion recognition in four emotions was 91.55%. In [147], the authors used the HMM for hidden sentiment detection in continuous text, and achieved 61.83% ACC and 66% AP.

5.1.4. RF

Random forest (RF) [148] is a type of parallel ensemble learning. RF uses the decision tree as the base learner to construct Bagging, and further introduces random attribute

selection in the training process of the decision tree. RF has a simple structure and a small amount of calculation. It can be used for both classification and regression problems. Even if the dataset is not complete, RF can maintain high classification accuracy. The increase in the classification tree does not affect the generalization performance of the classifier.

RF is an extended variant of the Bagging algorithm, as shown in Figure 14.

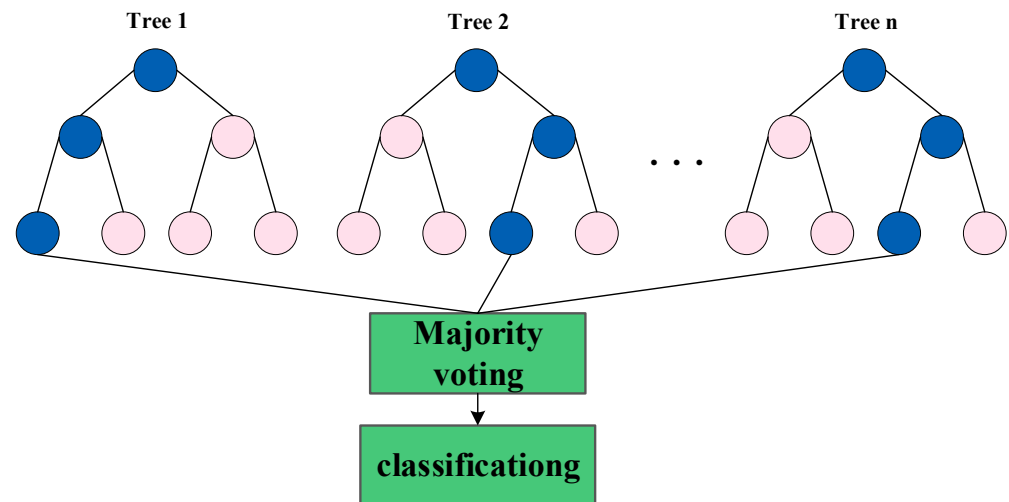


Figure 14. Structure of RF.

It randomly selects a subset containing k attributes from each node attribute set of the base decision tree, and then selects an optimal attribute from this subset for division. The parameter k controls the degree of randomness introduced. Finally, the base learners that have been trained are combined, and the majority voting method is usually used for classification prediction tasks.

In [149], the authors proposed two-layer fuzzy multiple random forest and achieved SER accuracy rates of 81.75% and 77.94% in CASIA and EmoDB datasets. In [78], the authors used RF to classify five emotions represented by HR and GSR physiological signals with an accuracy of 74%. In [150], the authors utilized multi-modal physiological signals and RF for anxiety state assessment. The classification accuracy for the five anxiety intensities reached 80.83%.

5.2. Deep Learning Methods

Compared to traditional machine learning methods, deep learning methods combine a feature extraction step and a classification step. With the support of large datasets, deep learning methods can learn higher-level semantic features. They have better discrimination ability for different emotions. Moreover, their generalization ability is stronger.

5.2.1. CNN

As a typical deep neural network, the convolutional neural network (CNN) plays an important role in the field of emotion recognition. The convolutional neural network is mainly composed of convolutional layers, a pooling layer, a fully connected layer, and a classification layer. The convolutional layer acts as a filter to extract features of the input signal. The introduction of nonlinear factors through activation functions can enhance the expressive ability of the model. The number of parameters and calculation consumption are reduced through the pooling layer. Finally, the classification layer is used to complete the classification of the input data.

One of the earliest convolutional neural networks [151] is shown in Figure 15 (adapted from [151]). The convolutional neural network has the characteristics of parameter sharing and local connection, which makes the training of the model more efficient.

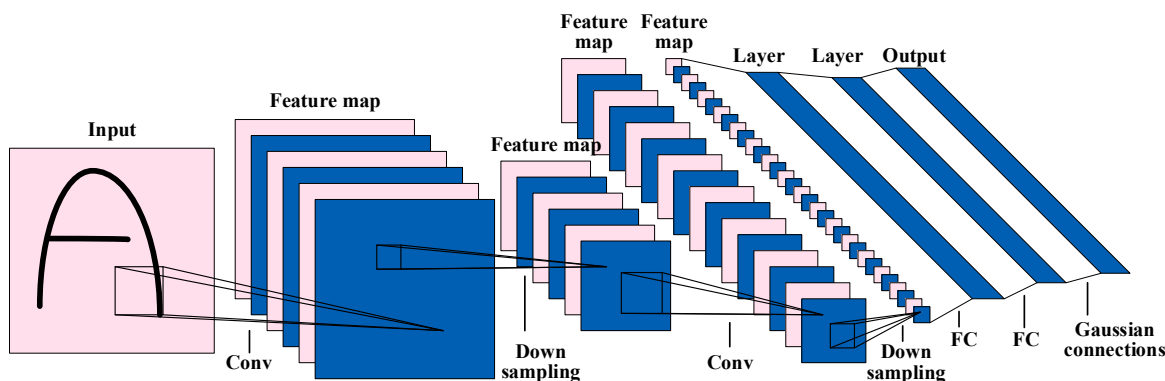


Figure 15. The network structure of LeNet.

In [152], the authors built Att-Net based on CNN, and the average recall of SER in three datasets was 78.01%, 80%, and 93%. In [153], the authors proposed a CNN-RNN-based approach for dimensional emotion recognition; in FER on the gradient emotion dataset, the average concordance correlation coefficient (CCC) of the valence dimension and the arousal dimension reached up to 0.450. In [154], the authors proposed the DCNN method and achieved the best accuracies of 87.31%, 75.34%, 79.25%, and 44.61% on four FER datasets. In [155], the authors proposed a dynamical graph convolutional neural network (DGCNN) for emotion recognition on multi-channel EEG signals; the average accuracy rate in the SEED dataset and the DREAMER dataset was 90.4%. In [156], the authors proposed a 3D-CNN network framework for multimodal emotion recognition from EEG signals and face video data; the accuracy of valence dimension and arousal dimension was 96.13% and 96.79%.

5.2.2. LSTM

Long Short-Term Memory (LSTM) [157] is an excellent recurrent neural network that can learn long-term dependencies from input data. At the same time, it can overcome problems such as exploding gradients and vanishing gradients. The classic LSTM framework is shown in Figure 16, which mainly includes three kinds of gate units: Input gate i_t , output gate o_t , and forget gate f_t . These gate units are used to control the information transfer of hidden state h_t , candidate state c_t , and candidate internal state \tilde{c}_t .

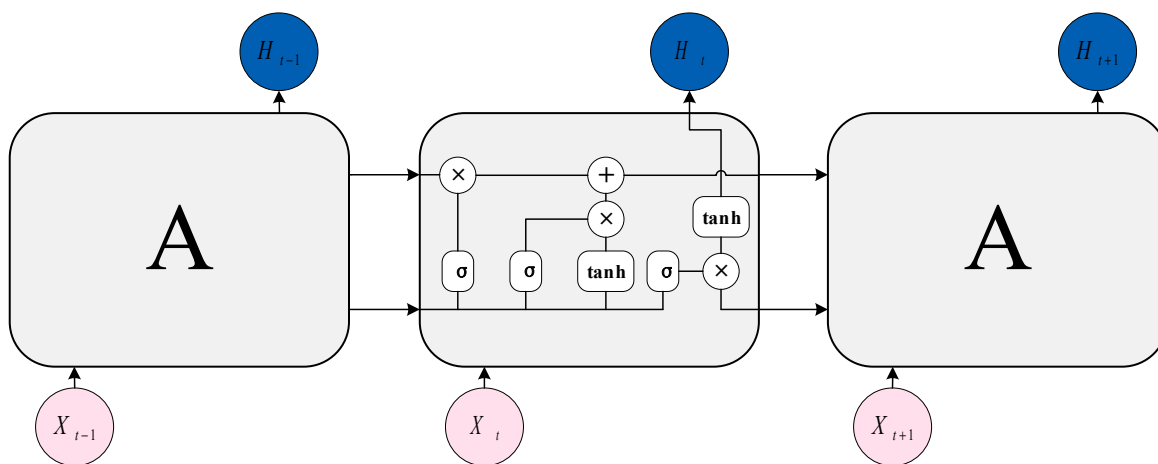


Figure 16. The structure of LSTM.

Each control gate and control state are calculated by the following formula:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{6}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (7)$$

$$\tilde{c}_t = f(W_{ct} x_t + W_{ch} h_{t-1} + b_c) \quad (8)$$

$$o_t = \sigma(W_o x_t + W_o h_{t-1} + W_{oc} c_t + b_o) \quad (9)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (10)$$

$$h_t = o_t \odot f(c_t) \quad (11)$$

Among them, W and U represent the weight, b is the bias, the matrix σ represents the logistic function, f is the activation function, and \odot represents the product of vector elements.

In [158], the authors used LSTM to achieve a FER accuracy of 73.5% for six emotions based on MFCC and spectrograms features. In [159], the authors proposed a CNN-LSTM model for emotion recognition based on EEG signals. For RAW data and STD data, the accuracy rates were 90.12% and 94.17%, and the loss rates were 30.12% and 42.43%. In [160], the authors proposed the Bi-direction Long-Short Term Memory with Direction Self-Attention (BLSTM-DSA) model for SER. For the IEMOCAP dataset and the EMO-DB dataset, the overall accuracy rates were 61.20% and 85.95%, and the average accuracy rates for each category were 54.99% and 82.06%.

5.2.3. DBN

The Deep Belief Network (DBN) generally consists of multiple restricted Boltzmann machines (RBM), as shown in Figure 17. RBM can avoid falling into local optimum. Each layer of the RBM is updated based on the previous layer. A DBN uses unsupervised learning and joint probability distributions to produce outputs. Hidden layer units are used to extract the correlation of high-order data in the display layer. The training of DBN mainly includes pre-training and fine-tuning.

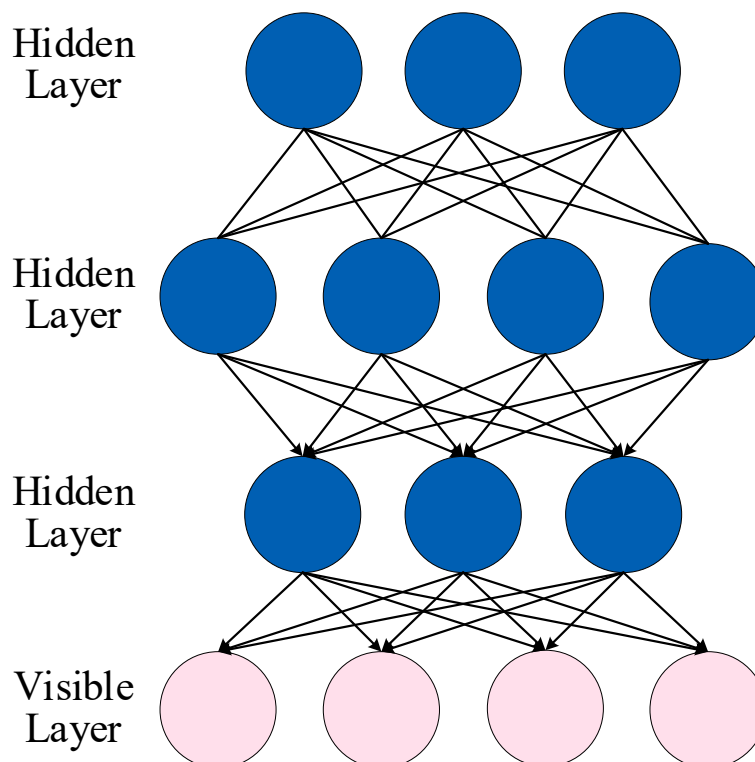


Figure 17. The structure of DBN.

In [161], the authors demonstrated the effectiveness of DBN in multimodal emotion recognition and achieved the best classification accuracy of 73.78% on the IEMOCAP audio-

visual dataset. In [162], the authors used DBN to extract deep features from EDA, PPG, and EMG signals and then classified them. The final overall accuracy rate was 89.53%. In [163], the authors proposed a bimodal deep belief network (BDBN) to fuse speech features and expression features for multimodal emotion recognition. The classification accuracy rate on the Friends dataset was 90.89%. In [164], the author proposed a method combining PCA, LDA, and DBN; the average recognition rate in the self-built facial expression recognition dataset was 92.50%.

5.2.4. Other Classification Methods

With the advancement in hardware and the improvement of computer processing power, many modern models have been proposed. They tend to have stronger classification performance and more complex structures. In order to make the classification method described in this article more comprehensive, we collected some excellent classification methods on large-scale datasets (including SER dataset, FER dataset, physiological signal dataset, and multimodal dataset). The classification method and details used by these articles are shown in Table 2 (details adapted from the cited article). Additionally, these datasets are introduced in Section 6 of this paper.

Table 2. Other classification methods.

Model Name	Dataset Used	Classification Method	Details
T5-3B [165]	SST (NLP)	Transformer and self-attention	The authors used transfer learning and self-attention to convert all text-based language problems into a text-to-text format. The authors compared the pre-training objectives, architectures, unlabeled datasets, and transfer methods of NLP. The classification accuracy on the SST dataset is 97.4%.
MT-DNN-SMART [166]	SST (NLP)	Transformer and smoothness inducing regularization	The authors proposed smoothness-induced regularization based on transfer learning to manage the complexity of the model. At the same time, a new optimization method was proposed to prevent over-updating. The classification accuracy on the SST dataset is 97.5%.
GRU [167]	CREMA-D (SER)	Self-supervised representation learning	The authors proposed a framework for learning audio representations guided by the visual modality in the context of audiovisual speech. The authors demonstrated the potential of visual supervision for learning audio representations; and achieved 55.01% SER accuracy on the CREMA-D dataset.
EmoAffectNet [168]	CREMA-D and AffectNet (FER)	CNN-LSTM	The authors proposed a flexible FER system using CNN and LSTM. This system consists of a backbone model and several temporal models. Every component of the system can be replaced by other models. The backbone model achieved an accuracy of 66.4% on the AffectNet dataset. The overall model achieved an accuracy of 79% on the CERMA-D dataset.
M2FNet [169]	IEMOCAP and MELD (multimodal)	Multi-task CNN and multi-head attention-based fusion	The multimodal fusion network proposed by the authors can extract emotional features from visual, audio, and textual modalities. The feature extractor was trained by an adaptive margin-based triplet loss function. The model achieved 67.85% accuracy and a 66.71 weighted average F1 score on the MELD dataset. Meanwhile, it achieved 69.69% accuracy and a 69.86 weighted average F1 score on the MELD dataset.
CH Fusion [170]	IEMOCAP (multimodal)	RNN and feature fusion strategy	The authors used RNN to extract the unimodal features of the three modalities of audio, video, and text. These unimodal features were then fused through a fully connected layer to form trimodal features. Finally, feature vectors for sentiment classification were obtained. The model achieved an F1 score of 0.768 and an accuracy rate of 0.765 on the IEMOCAP dataset.

Table 2. Cont.

Model Name	Dataset Used	Classification Method	Details
EmotionFlow-large [171]	MELD (multimodal)	BERT model and Conditional random field (CRF)	The authors researched the propagation of emotions in dialogue emotion recognition. The authors utilized an encoder-decoder structure to learn user-specific features. Conditional random fields (CRF) were then applied to capture sequence information at the sentiment level. The weighted F1 score on the MELD dataset was 66.50.
FN2EN [172]	CK+ (FER)	DCNN	The authors proposed a two-stage training algorithm. In the first stage, high-level neuronal responses were modeled using probability distribution functions based on the fine-tuned face network. In the second stage, the authors conducted label supervision to improve the discriminative ability. The model achieved 96.8% (eight emotions) and 98.6% (six emotions) accuracy on the CK+ dataset.
Multi-task EfficientNet-B2 [173]	AffectNet (FER)	MTCNN and Adam optimization	In the article, the authors analyzed the behavior of students in the e-learning environment. The facial features obtained by the model could be used to quickly predict student engagement, individual emotions, and group-level influence. The model could even be used for real-time video processing on each student's mobile device without sending the video to a remote server or the teacher's PC. The model achieved 63.03% (eight emotions) and 66.29% (seven emotions) accuracy on the AffectNet dataset.
EAC [174]	RAF-DB (FER)	CNN and Class Activation Mapping (CAM)	The authors approached noisy label FER from the perspective of feature learning, and proposed Erase Attention Consistency (EAC). EAC does not require noise rate or label integration. It can generalize better to noisy label classification tasks with a large number of classes. The overall accuracy on the RAF-DB dataset was 90.35%.
BiHDM [175]	SEED (EEG signal)	RNNs	The authors proposed a model to learn the differential information of the left and right hemispheres of the human brain to improve EEG emotion recognition. The authors employed four directed recurrent neural networks based on two orientations to traverse electrode signals on two separate brain regions. This preserved its inherent spatial dependence. The accuracy on the SEED dataset reached 74.35%.
MMLatch [176]	CMU-MOSEI (multimodal)	LSTM, RNNs and Transformers	The neural architecture proposed by the authors could capture top-down cross-modal interactions. A forward propagation feedback mechanism was used during model training. The accuracy rate on the CMU-MOSEI dataset reached 82.4.

6. Datasets

Datasets play an important role in data-driven learning [177], which can improve the performance and robustness of models. Emotion recognition datasets are based on signal categories. According to the different signal categories, the emotion recognition datasets can be divided into: speech (textual, audio) datasets, visual (facial expression picture or video) datasets, physiological datasets, and multi-modal signal datasets.

Speech datasets for emotion recognition can be divided into performer-based [178], induced [179], and natural [180] datasets according to the method of acquisition. The performer-based datasets mainly consist of speech recordings of various emotions performed by performers with extensive experience [49]. Induced datasets are the emotions expressed by people in artificially created environments [181]. Induced datasets are relatively less expressive, but closer to reality. Natural datasets are the most realistic, usually taken from public conversations [182] or call center conversations [183], these data contain more emotional changes and background noise, but the amount is relatively limited. The commonly used speech emotion recognition datasets are shown in Table 3 (details adapted from the cited article).

Table 3. Dataset for speech emotion recognition.

Name	Type	Details	Number of Emotion Categories	Number of Samples
MDS [184]	Textual	Product reviews from the Amazon shopping site; consisting of different words, sentences, and documents	2 or 5	100,000
SST [185]	Textual	Semantic emotion recognition database established by Stanford University	2 or 5	11,855
IMDB [186]	Textual	Contains a large number of movie reviews	2	25,000
EMODB [187]	Performer-based	The dataset consists of ten German voices spoken by ten speakers (five males and five females)	7	800
SAVEE [188]	Performer-based	Performed by four female speakers; spoken in English	7	480
CREAM-D [189]	Performer-based	Spoken in English	6	7442
IEMOCAP * [190]	Performer-based	Conversation between two people (one male and one female); spoken in English	4	-
Chinese Emotion Speech Dataset [191]	Induced	Spoken in Chinese	5	3649
MELD * [192]	Induced	Data from TC-series Friends	3	13,000
RECOLA Speech Database [179]	Natural	Spoken by 46 speakers (19 male and 27 female); spoken in French	5	7 h
FAU Aibo emotion corpus [193]	Natural	Communications between 51 children and a robot dog; spoken in German	11	9 h
Semaine Database [194]	Natural	Spoken by 150 speakers; spoken in English, Greek, and Hebrew	5	959 conversations
CHEAVD [195]	Natural	Spoken by 238 speakers (from children to the elderly); spoken in Chinese	26	2322

* Can also be used for multimodal emotion recognition.

For facial expression datasets, different datasets vary in terms of the acquisition environment, the number of emotion categories, age, race, image quality, etc. [196]. The commonly used facial expression recognition datasets are shown in Table 4 (details adapted from the cited article).

Table 4. Datasets for facial expression recognition.

Name	Type	Details	Number of Emotion Categories	Number of Samples
BP4D [197]	Induced	41 participants; 4 ethnicities; 18–29 years old	8	368,036
CK+ [198]	Induced	123 participants; 23 facial displays; 21–53 years old	7	593 sequences
BU-4DEF [199]	Induced	101 participants; 5 ethnicities	6	606 sequences
SEWA [200]	Induced	96 participants; 6 ethnicities; 18–65 years old	7	1990 sequences
MMI-V [201]	Performer-based	25 participants; 3 ethnicities; 19–62 years old	6	1.5 h
JAFFE [202]	Performer-based	10 participants	6	213
BU-3DEF [203]	Performer-based	100 participants 18–70 years old	6	2500
AffectNet [204]	Natural	Average age is 33.01 years old; downloaded from the Internet	6	450,000
RAF-DB [205]	Natural	Collected from Flickr	compound	29,672
EmotioNet [206]	Natural	Downloaded from the Internet	compound	1,000,000

Physiological signals can represent more real emotions and will not be affected by people's hidden emotional behavior. Common datasets based on physiological signals are shown in Table 5 (details adapted from the cited article).

Table 5. Datasets of physiological signals.

Name	Type	Details	Number of Emotion Categories	Physiological Signals
DEAP * [207]	Induced	32 participants; average age is 26.9 years old	Dimensional emotion (arousal-valence- dominance)	EEG; EMG; RSP; GSR; EOG; plethysmograph; skin temperature

Table 5. Cont.

Name	Type	Details	Number of Emotion Categories	Physiological Signals
DECAF * [208]	Induced	30 participants	Dimensional emotion (arousal-valence)	EMG; NIR; hEOG; ECG; tEMG
AMIGOS * [209]	Induced	Individual participant and group participants	Dimensional emotion (arousal-valence)	EEG; GSR; ECG
SEED * [210]	Induced	15 participants; average age is 23.3	3	EEG; EOG
DREAMER * [77]	Induced	23 participants; collected by wireless low-cost off-the-self devices	Dimensional emotion (arousal-valence-dominance)	EEG; ECG

* Can also be used for multimodal emotion recognition.

The radar sensor can be used to obtain people's heartbeat signals or breathing signals without contact. These sensors mainly include continuous wave radar [47,211], continuous frequency modulated wave (FMCW) radar [212], millimeter wave radar [213], and RFID tag [34]. Datasets based on radar sensor signals are less widely used according to our survey, and most researchers tend to make their own datasets. Most radar data use clipped videos or pictures as emotional inducers. Radar sensors are used to collect physiological signals of volunteers to make datasets.

Commonly used multi-modal emotion recognition datasets are shown in Table 6 (details adapted from the cited article). The multimodal signal dataset contains at least two different signals and richer information. Multi-modal emotion recognition datasets often require a larger amount of data, and the data usually needs to be labeled. Therefore, making multimodal signal datasets becomes more difficult than normal datasets.

Table 6. Datasets for multi-modal emotion recognition.

Name	Type	Details	Number of Emotion Categories	Types of Signals
eINTERFACE [49]	Induced	42 participants; 14 different nationalities	6	Visual signals; audio signals;
RECOLA [179]	Natural	46 participants; 9.5 h	Dimensional emotion (arousal-valence)	Visual signals; audio signals; ECG signals; EDA signals

Table 6. Cont.

Name	Type	Details	Number of Emotion Categories	Types of Signals
CMU-MOSEI [214]	Natural	23,453 annotated video segments; 1000 speaker; 250 topics	6	Textual signals visual signals; audio signals;
MAHNOB-HCI [215]	Induced	27 participants	Dimensional emotion (arousal-valence-dominance)	Textual signals visual signals; audio signals; EEG signals; RSP signals; GSR signals; ECG signals; skin temperature signals

At the same time, we also need to consider the synchronization of multi-channel signals during the recording of different sensors, as some devices may record on different time scales. Multi-modal signal datasets can make the machine's analysis of emotion more comprehensive. At present, researchers are paying increasing attention to multimodal emotion recognition.

7. Conclusions and Discussions

In this survey, we reviewed more than 200 papers, including working processes, methods, and commonly used datasets of different sensors for emotion recognition. In this section, we summarize the main findings from this survey.

Facial expressions can intuitively reflect the subjective emotions in interpersonal communication, but they are affected by limited lighting, occlusion, small changes in facial expressions, and individual differences. The performances of existing vision-based emotion recognition systems will significantly drop in environments with changing lighting conditions. Self-occlusion due to head rotation or face contact, and occlusion by other people passing in front of the camera, are both common problems. Moreover, individual differences can affect the feature extraction and learning of the model. There are large differences between infants and adults, males and females, and different groups, which makes it challenging to train a FER classifier with strong generalization performance.

SER is also of great significance in emotion recognition. Due to the variability of emotions, a piece of speech often has multiple emotions, which is challenging for the accurate extraction of speech information features. For multiple languages, cross-cultural emotion recognition is the future development trend. People in different countries and regions have certain cultural differences, but, for humans, even if they cannot understand what foreigners are saying, they can roughly understand their tone and attitude.

Emotional changes are also reflected in the physiological changes of the human body. The most basic challenge of emotion recognition from physiological signals is the accurate emotional labeling of data. In real life, parties often do not realize that they have developed certain emotions, because the parties are caught in the emotion itself. Therefore, participants need to exactly record when a certain emotion occurs. Only in this way can the corresponding physiological signals be extracted. Some physiological signal recording devices are expensive and invasive, which greatly limits the number of subjects and the length of the experiment. Therefore, some non-contact physiological signal recording devices are more popular.

Multi-modal emotion recognition based on multi-sensors can make up for the deficiency of single sensor. It is more robust and is now receiving more attention. It uses different signals to extract features and perform feature-level or decision-level fusion to a

certain extent, which can improve the accuracy of discrete or dimensional sentiment classification. The main challenges include how to choose an appropriate feature representation method and feature selection method based on multi-modal signal input. Different modal signals may also have mutual dependencies in different time dimensions, and classifiers need to be designed according to the potential correlation of different modal datasets.

We introduced several commonly used classical machine learning and deep learning classification methods. Classical machine learning methods have faster speeds and simpler structures. However, for large and high-dimensional data, researchers prefer deep learning. With the improvement in computer technology, deeper and larger deep learning models have been proposed, which can extract high-dimensional features better. However, this does not mean that classic machine learning methods are abandoned. For limited training data, machine learning often achieves better results.

Based on the above conclusions, we think that single-modal emotion recognition cannot meet human needs in some specific application scenarios. Therefore, the current research on multimodal information processing is more popular. However, the research on multimodal emotion recognition has more challenges. They include experimental environment, sensors, signal acquisition, signal processing, information annotation, etc. At the same time, we believe that emotion recognition is an important part of the development of artificial intelligence. Accurate recognition of emotions can enable machines to better serve people and care about people's health and life in more detail.

8. Future Trends

Emotion recognition is of great significance to both human and social development. The current challenges and development trends of emotion recognition mainly include technical aspects and security aspects.

The first is to improve user acceptance. At present, many people are not familiar with various emotional computing sensors, and some sensors need to be worn by users. In order to improve the degree of cooperation of users, practitioners need to give detailed instructions to users. The detection system should also be user-centered, with the primary goal of protecting the user's physical and mental health.

The second point is security. The process of human emotion recognition involves highly private personal information, including health, location, and physiological characteristics. Emotion recognition should be used in socially beneficial research rather than being used to cause legal problems or discrimination. Therefore, protecting user privacy has also become a major challenge for emotion recognition. At present, decentralized AI technology can overcome the limitations of centralized information storage and improve data privacy and security.

The third point is robustness and accuracy. The current emotion recognition model cannot simulate all aspects of human emotions. In order to be more comprehensive, multimodal emotion recognition has become the first choice for most researchers. With larger models and datasets, multimodal approaches can achieve better results. Emotion recognition often requires more information, and short-term or transient features can only represent people's psychological state at a specific time. Studies on personality analysis, such as autism diagnosis and intelligence testing, require longer-term observation. Therefore, the extraction of long-term features is also challenging and of great research significance for emotion recognition.

In order to obtain a good emotion recognition model, there are more stringent requirements for datasets. With the continuous production of large-scale datasets, the advantages of unsupervised learning and reinforcement learning are more obvious. Unsupervised learning does not require pre-stored labels or specifications. Moreover, it can also complete classification without category information. Reinforcement learning enables the model to maximize rewards through the principle of trial and error, and can continuously optimize the performance of the system. These emotion recognition methods are also worthy of research.

Author Contributions: Conceptualization, Y.C. and X.L.; methodology, Y.C.; formal analysis, Y.C. and X.L.; investigation, Y.C.; resources, X.L.; writing—original draft preparation, Y.C.; writing—review and editing, Y.C., X.L. and J.L.; visualization, Y.C.; supervision, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Some of the datasets mentioned in the paper can be downloaded in: <https://paperswithcode.com>.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Picard, R.W. *Affective Computing*; MIT Press: Cambridge, MA, USA, 2000; pp. 4–5.
2. Nayak, S.; Nagesh, B.; Routray, A.; Sarma, M. A Human-Computer Interaction Framework for Emotion Recognition through Time-Series Thermal Video Sequences. *Comput. Electr. Eng.* **2021**, *93*, 107280. [[CrossRef](#)]
3. Colonnello, V.; Mattarozzi, K.; Russo, P.M. Emotion Recognition in Medical Students: Effects of Facial Appearance and Care Schema Activation. *Med. Educ.* **2019**, *53*, 195–205. [[CrossRef](#)] [[PubMed](#)]
4. Feng, X.; Wei, Y.J.; Pan, X.L.; Qiu, L.H.; Ma, Y.M. Academic Emotion Classification and Recognition Method for Large-Scale Online Learning Environment-Based on a-Cnn and Lstm-Att Deep Learning Pipeline Method. *Int. J. Environ. Res. Public Health* **2020**, *17*, 1941. [[CrossRef](#)] [[PubMed](#)]
5. Fu, E.; Li, X.; Yao, Z.; Ren, Y.X.; Wu, Y.H.; Fan, Q.Q. Personnel Emotion Recognition Model for Internet of Vehicles Security Monitoring in Community Public Space. *Eurasip J. Adv. Signal Process.* **2021**, *2021*, 81. [[CrossRef](#)]
6. Oh, G.; Ryu, J.; Jeong, E.; Yang, J.H.; Hwang, S.; Lee, S.; Lim, S. DRER: Deep Learning-Based Driver’s Real Emotion Recognizer. *Sensors* **2021**, *21*, 2166. [[CrossRef](#)] [[PubMed](#)]
7. Sun, X.; Song, Y.Z.; Wang, M. Toward Sensing Emotions With Deep Visual Analysis: A Long-Term Psychological Modeling Approach. *IEEE Multimed.* **2020**, *27*, 18–27. [[CrossRef](#)]
8. Mandryk, R.L.; Atkins, M.S.; Inkpen, K.M. A continuous and objective evaluation of emotional experience with interactive play environments. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Montréal, QC, Canada, 22–27 April 2006; pp. 1027–1036.
9. Ogata, T.; Sugano, S. Emotional communication between humans and the autonomous robot which has the emotion model. In Proceedings of the 1999 IEEE International Conference on Robotics and Automation (Cat. No. 99CH36288C), Detroit, MI, USA, 10–15 May 1999; pp. 3177–3182.
10. Malfaz, M.; Salichs, M.A. A new architecture for autonomous robots based on emotions. *IFAC Proc. Vol.* **2004**, *37*, 805–809. [[CrossRef](#)]
11. Rattanyu, K.; Ohkura, M.; Mizukawa, M. Emotion monitoring from physiological signals for service robots in the living space. In Proceedings of the ICCAS 2010, Gyeonggi-do, Republic of Korea, 27–30 October 2010; pp. 580–583.
12. Hasnul, M.A.; Aziz, N.A.A.; Alelyani, S.; Mohana, M.; Aziz, A.A.J.S. Electrocardiogram-based emotion recognition systems and their applications in healthcare—A review. *Sensors* **2021**, *21*, 5015. [[CrossRef](#)]
13. Feidakis, M.; Daradoumis, T.; Caballé, S. Emotion measurement in intelligent tutoring systems: What, when and how to measure. In Proceedings of the 2011 Third International Conference on Intelligent Networking and Collaborative Systems, Fukuoka, Japan, 30 November–2 December 2011; pp. 807–812.
14. Saste, S.T.; Jagdale, S. Emotion recognition from speech using MFCC and DWT for security system. In Proceedings of the 2017 international Conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 20–22 April 2017; pp. 701–704.
15. Zepf, S.; Hernandez, J.; Schmitt, A.; Minker, W.; Picard, R. Driver emotion recognition for intelligent vehicles: A survey. *ACM Comput. Surv. (CSUR)* **2020**, *53*, 1–30. [[CrossRef](#)]
16. Houben, M.; Van Den Noortgate, W.; Kuppens, P. The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychol. Bull.* **2015**, *141*, 901. [[CrossRef](#)]
17. Bal, E.; Harden, E.; Lamb, D.; Van Hecke, A.V.; Denver, J.W.; Porges, S.W. Emotion recognition in children with autism spectrum disorders: Relations to eye gaze and autonomic state. *J. Autism Dev. Disord.* **2010**, *40*, 358–370. [[CrossRef](#)]
18. Martínez, R.; Ipiña, K.; Irigoyen, E.; Asla, N.; Garay, N.; Ezeiza, A.; Fajardo, I. Emotion elicitation oriented to the development of a human emotion management system for people with intellectual disabilities. In *Trends in Practical Applications of Agents and Multiagent Systems: 8th International Conference on Practical Applications of Agents and Multiagent Systems*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 689–696.
19. Ekman, P. Universals and cultural differences in facial expressions of emotion. In *Nebraska Symposium on Motivation*; University of Nebraska Press: Lincoln, NE, USA, 1971.

20. Darwin, C.; Prodger, P. *The Expression of the Emotions in Man and Animals*; Oxford University Press: Oxford, UK, 1998.
21. Ekman, P.; Sorenson, E.R.; Friesen, W.V.J.S. Pan-cultural elements in facial displays of emotion. *Science* **1969**, *164*, 86–88. [[CrossRef](#)]
22. Plutchik, R. *Emotions and Life: Perspectives from Psychology, Biology, and Evolution*; American Psychological Association: Washington, DC, USA, 2003.
23. Bakker, I.; Van Der Voordt, T.; Vink, P.; De Boon, J. Pleasure, arousal, dominance: Mehrabian and Russell revisited. *Curr. Psychol.* **2014**, *33*, 405–421. [[CrossRef](#)]
24. Mehrabian, A.; Russell, J.A. *An Approach to Environmental Psychology*; the MIT Press: Cambridge, MA, USA, 1974.
25. Bain, A. *The senses and the intellect*; Longman, Green, Longman, Roberts, and Green: London, UK, 1864.
26. Hassan, M.U.; Rehmani, M.H.; Chen, J.; Computing, D. Differential privacy in blockchain technology: A futuristic approach. *J. Parallel Distrib. Comput.* **2020**, *145*, 50–74. [[CrossRef](#)]
27. Ray, T.R.; Choi, J.; Bandodkar, A.J.; Krishnan, S.; Gutruf, P.; Tian, L.; Ghaffari, R.; Rogers, J. Bio-integrated wearable systems: A comprehensive review. *Chem. Rev.* **2019**, *119*, 5461–5533. [[CrossRef](#)]
28. Li, S.; Deng, W. Deep facial expression recognition: A survey. *IEEE Trans. Affect. Comput.* **2020**, *13*, 1195–1215. [[CrossRef](#)]
29. Schmid, P.C.; Mast, M.S.; Bombardi, D.; Mast, F.W.; Lobmaier, J. How mood states affect information processing during facial emotion recognition: An eye tracking study. *Swiss J. Psychol.* **2011**. [[CrossRef](#)]
30. Sandbach, G.; Zafeiriou, S.; Pantic, M.; Yin, L.J.I.; Computing, V. Static and dynamic 3D facial expression recognition: A comprehensive survey. *Image Vis. Comput.* **2012**, *30*, 683–697. [[CrossRef](#)]
31. Wang, W.; Den Brinker, A.C.; Stuijk, S.; De Haan, G. Algorithmic principles of remote PPG. *IEEE Trans. Biomed. Eng.* **2016**, *64*, 1479–1491. [[CrossRef](#)] [[PubMed](#)]
32. Xie, K.; Fu, C.-H.; Liang, H.; Hong, H.; Zhu, X. Non-contact heart rate monitoring for intensive exercise based on singular spectrum analysis. In Proceedings of the 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), San Jose, CA, USA, 28–30 March 2019; pp. 228–233.
33. Ghimire, D.; Lee, J. Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. *Sensors* **2013**, *13*, 7714–7734. [[CrossRef](#)] [[PubMed](#)]
34. Zhao, M.; Adib, F.; Katabi, D. Emotion recognition using wireless signals. In Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking, New York, NY, USA, 3–7 October 2016; pp. 95–108.
35. Zhang, J.; Yin, Z.; Chen, P.; Nichele, S. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Inf. Fusion* **2020**, *59*, 103–126. [[CrossRef](#)]
36. Lopes, A.T.; De Aguiar, E.; De Souza, A.F.; Oliveira-Santos, T. Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order. *Pattern Recognit.* **2017**, *61*, 610–628. [[CrossRef](#)]
37. Kim, D.H.; Baddar, W.J.; Jang, J.; Ro, Y. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Trans. Affect. Comput.* **2017**, *10*, 223–236. [[CrossRef](#)]
38. Zhong, L.; Liu, Q.; Yang, P.; Huang, J.; Metaxas, D. Learning multiscale active facial patches for expression analysis. *IEEE Trans. Cybern.* **2014**, *45*, 1499–1510. [[CrossRef](#)] [[PubMed](#)]
39. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015; pp. 815–823.
40. Hertzman, A.B. Photoelectric plethysmography of the fingers and toes in man. *Proc. Soc. Exp. Biol. Med.* **1937**, *37*, 529–534. [[CrossRef](#)]
41. Ram, M.R.; Madhav, K.V.; Krishna, E.H.; Komalla, N.R.; Reddy, K.A. A novel approach for motion artifact reduction in PPG signals based on AS-LMS adaptive filter. *IEEE Trans. Instrum. Meas.* **2011**, *61*, 1445–1457. [[CrossRef](#)]
42. Temko, A. Accurate heart rate monitoring during physical exercises using PPG. *IEEE Trans. Biomed. Eng.* **2017**, *64*, 2016–2024. [[CrossRef](#)] [[PubMed](#)]
43. Poh, M.-Z.; McDuff, D.J.; Picard, R.W. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Trans. Biomed. Eng.* **2010**, *58*, 7–11. [[CrossRef](#)] [[PubMed](#)]
44. Li, X.; Chen, J.; Zhao, G.; Pietikainen, M. Remote heart rate measurement from face videos under realistic situations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 4264–4271.
45. Tarassenko, L.; Villarroel, M.; Guazzi, A.; Jorge, J.; Clifton, D.; Pugh, C. Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. *Physiol. Meas.* **2014**, *35*, 807. [[CrossRef](#)] [[PubMed](#)]
46. Jeong, I.C.; Finkelstein, J. Introducing contactless blood pressure assessment using a high speed video camera. *J. Med. Syst.* **2016**, *40*, 1–10. [[CrossRef](#)]
47. Zhang, L.; Fu, C.-H.; Hong, H.; Xue, B.; Gu, X.; Zhu, X.; Li, C. Non-contact Dual-modality emotion recognition system by CW radar and RGB camera. *IEEE Sens. J.* **2021**, *21*, 23198–23212. [[CrossRef](#)]
48. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.E.; Mohamed, A.R.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.N.; et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Process. Mag.* **2012**, *29*, 82–97. [[CrossRef](#)]
49. Martin, O.; Kotsia, I.; Macq, B.; Pitas, I. The eINTERFACE'05 audio-visual emotion database. In Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06), Atlanta, GA, USA, 3–7 April 2016; p. 8.
50. Li, J.; Deng, L.; Haeb-Umbach, R.; Gong, Y. *Robust Automatic Speech Recognition: A Bridge to Practical Applications*; Academic Press: Cambridge, MA, USA, 2015.

51. Williams, C.E.; Stevens, K. Vocal correlates of emotional states. *Speech Eval. Psychiatry* **1981**, *52*, 221–240.
52. Schuller, B.W. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM* **2018**, *61*, 90–99. [[CrossRef](#)]
53. France, D.J.; Shiavi, R.G.; Silverman, S.; Silverman, M.; Wilkes, M. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Trans. Biomed. Eng.* **2000**, *47*, 829–837. [[CrossRef](#)]
54. Hansen, J.H.; Cairns, D.A. Icarus: Source generator based real-time recognition of speech in noisy stressful and lombard effect environments. *Speech Commun.* **1995**, *16*, 391–422. [[CrossRef](#)]
55. Ang, J.; Dhillon, R.; Krupski, A.; Shriberg, E.; Stolcke, A. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In Proceedings of the INTERSPEECH, Denver, CO, USA, 16–20 September 2002; pp. 2037–2040.
56. Cohen, R. A computational theory of the function of clue words in argument understanding. In Proceedings of the 10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics, Stanford University, Stanford, CA, USA, 2–6 July 1984; pp. 251–258.
57. Deng, J.; Frühholz, S.; Zhang, Z.; Schuller, B. Recognizing emotions from whispered speech based on acoustic feature transfer learning. *IEEE Access* **2017**, *5*, 5235–5246. [[CrossRef](#)]
58. Guo, S.; Feng, L.; Feng, Z.-B.; Li, Y.-H.; Wang, Y.; Liu, S.-L.; Qiao, H. Multi-view laplacian least squares for human emotion recognition. *Neurocomputing* **2019**, *370*, 78–87. [[CrossRef](#)]
59. Grosz, B.J.; Sidner, C.L. Attention, intentions, and the structure of discourse. *Comput. Linguist.* **1986**, *12*, 175–204.
60. Dellaert, F.; Polzin, T.; Waibel, A. Recognizing emotion in speech. In Proceedings of the Fourth International Conference on Spoken Language Processing. ICSLP'96, Philadelphia, PA, USA, 3–6 October 1996; pp. 1970–1973.
61. Burmania, A.; Busso, C. A Stepwise Analysis of Aggregated Crowdsourced Labels Describing Multimodal Emotional Behaviors. In Proceedings of the INTERSPEECH, Stockholm, Sweden, 20–24 August 2017; pp. 152–156.
62. Lee, S.-W. The generalization effect for multilingual speech emotion recognition across heterogeneous languages. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 5881–5885.
63. Ashhad, S.; Kam, K.; Del Negro, C.A.; Feldman, J. Breathing rhythm and pattern and their influence on emotion. *Annu. Rev. Neurosci.* **2022**, *45*, 223–247. [[CrossRef](#)]
64. Du, G.; Long, S.; Yuan, H. Non-contact emotion recognition combining heart rate and facial expression for interactive gaming environments. *IEEE Access* **2020**, *8*, 11896–11906. [[CrossRef](#)]
65. Verkruyse, W.; Svaasand, L.O.; Nelson, J. Remote plethysmographic imaging using ambient light. *Opt. Express* **2008**, *16*, 21434–21445. [[CrossRef](#)] [[PubMed](#)]
66. Qing, C.; Qiao, R.; Xu, X.; Cheng, Y. Interpretable emotion recognition using EEG signals. *IEEE Access* **2019**, *7*, 94160–94170. [[CrossRef](#)]
67. Theorell, T.; Ahlberg-Hulten, G.; Jodko, M.; Sigala, F.; De La Torre, B. Influence of job strain and emotion on blood pressure in female hospital personnel during workhours. *Scand. J. Work Environ. Health* **1993**, *19*, 313–318. [[CrossRef](#)] [[PubMed](#)]
68. Nouman, M.; Khoo, S.Y.; Mahmud, M.P.; Kouzani, A. Recent Advances in Contactless Sensing Technologies for Mental Health Monitoring. *IEEE Internet Things J.* **2021**, *9*, 274–297. [[CrossRef](#)]
69. Boiten, F. The effects of emotional behaviour on components of the respiratory cycle. *Biol. Psychol.* **1998**, *49*, 29–51. [[CrossRef](#)]
70. Yasuma, F.; Hayano, J. Respiratory sinus arrhythmia: Why does the heartbeat synchronize with respiratory rhythm? *Chest* **2004**, *125*, 683–690. [[CrossRef](#)]
71. Li, C.; Cummings, J.; Lam, J.; Graves, E.; Wu, W. Radar remote monitoring of vital signs. *IEEE Microw. Mag.* **2009**, *10*, 47–56. [[CrossRef](#)]
72. Li, H.; Shrestha, A.; Heidari, H.; Le Kernec, J.; Fioranelli, F. Bi-LSTM network for multimodal continuous human activity recognition and fall detection. *IEEE Sens. J.* **2019**, *20*, 1191–1201. [[CrossRef](#)]
73. Ren, L.; Kong, L.; Foroughian, F.; Wang, H.; Theilmann, P.; Fathy, A. Comparison study of noncontact vital signs detection using a Doppler stepped-frequency continuous-wave radar and camera-based imaging photoplethysmography. *IEEE Trans. Microw. Theory Technol.* **2017**, *65*, 3519–3529. [[CrossRef](#)]
74. Gu, C.; Wang, G.; Li, Y.; Inoue, T.; Li, C. A hybrid radar-camera sensing system with phase compensation for random body movement cancellation in Doppler vital sign detection. *IEEE Trans. Microw. Theory Technol.* **2013**, *61*, 4678–4688. [[CrossRef](#)]
75. James, W. *The Principles of Psychology*; Cosimo, Inc.: New York, NY, USA, 2007; Volume 1.
76. Petrantonakis, P.C.; Hadjileontiadis, L. Emotion recognition from brain signals using hybrid adaptive filtering and higher order crossings analysis. *IEEE Trans. Affect. Comput.* **2010**, *1*, 81–97. [[CrossRef](#)]
77. Katsigiannis, S.; Ramzan, N. DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE J. Biomed. Health Inform.* **2017**, *22*, 98–107. [[CrossRef](#)] [[PubMed](#)]
78. Wen, W.; Liu, G.; Cheng, N.; Wei, J.; Shangguan, P.; Huang, W. Emotion recognition based on multi-variant correlation of physiological signals. *IEEE Trans. Affect. Comput.* **2014**, *5*, 126–140. [[CrossRef](#)]
79. Jerritta, S.; Murugappan, M.; Nagarajan, R.; Wan, K. Physiological signals based human emotion recognition: A review. In Proceedings of the 2011 IEEE 7th International Colloquium on Signal Processing and Its Applications, Penang, Malaysia, 4–6 March 2011; pp. 410–415.

80. Kushki, A.; Fairley, J.; Merja, S.; King, G.; Chau, T. Comparison of blood volume pulse and skin conductance responses to mental and affective stimuli at different anatomical sites. *Physiol. Meas.* **2011**, *32*, 1529. [[CrossRef](#)]
81. Lim, J.Z.; Mountstephens, J.; Teo, J. Emotion recognition using eye-tracking: Taxonomy, review and current challenges. *Sensors* **2020**, *20*, 2384. [[CrossRef](#)] [[PubMed](#)]
82. Ekman, P. The argument and evidence about universals in facial expressions. In *Handbook of Social Psychophysiology*; John Wiley & Sons: Hoboken, NJ, USA, 1989; Volume 143, p. 164.
83. Li, S.; Kang, X.; Fang, L.; Hu, J.; Yin, H. Pixel-level image fusion: A survey of the state of the art. *Inf. Fusion* **2017**, *33*, 100–112. [[CrossRef](#)]
84. Cai, H.; Qu, Z.; Li, Z.; Zhang, Y.; Hu, X.; Hu, B. Feature-level fusion approaches based on multimodal EEG data for depression recognition. *Inf. Fusion* **2020**, *59*, 127–138. [[CrossRef](#)]
85. Ho, T.K.; Hull, J.J.; Srihari, S. Decision combination in multiple classifier systems. *IEEE Trans. Pattern Anal. Mach. Intell.* **1994**, *16*, 66–75.
86. Aziz, A.M. A new adaptive decentralized soft decision combining rule for distributed sensor systems with data fusion. *Inf. Sci.* **2014**, *256*, 197–210. [[CrossRef](#)]
87. Canal, F.Z.; Müller, T.R.; Matias, J.C.; Scotton, G.G.; de Sa Junior, A.R.; Pozzebon, E.; Sobieranski, A. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Inf. Sci.* **2022**, *582*, 593–617. [[CrossRef](#)]
88. Kartali, A.; Roglič, M.; Barjaktarović, M.; Đurić-Jovičić, M.; Janković, M.M. Real-time algorithms for facial emotion recognition: A comparison of different approaches. In Proceedings of the 2018 14th Symposium on Neural Networks and Applications (NEUREL), Belgrade, Serbia, 20–21 November 2018; pp. 1–4.
89. Nema, B.M.; Abdul-Kareem, A.A. Preprocessing signal for speech emotion recognition. *J. Sci.* **2018**, *28*, 157–165. [[CrossRef](#)]
90. Beigi, H. *Fundamentals of Speaker Recognition*; Springer Science & Business Media: Berlin, Germany, 2011.
91. Jerritta, S.; Murugappan, M.; Wan, K.; Yaacob, S. Emotion recognition from facial EMG signals using higher order statistics and principal component analysis. *J. Chin. Inst. Eng.* **2014**, *37*, 385–394. [[CrossRef](#)]
92. Izard, C.E. Emotion theory and research: Highlights, unanswered questions, and emerging issues. *Annu. Rev. Psychol.* **2009**, *60*, 1–25. [[CrossRef](#)]
93. Subasi, A. EEG signal classification using wavelet feature extraction and a mixture of expert model. *Expert Syst. Appl.* **2007**, *32*, 1084–1093. [[CrossRef](#)]
94. Pincus, S.M. Approximate entropy as a measure of system complexity. *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 2297–2301. [[CrossRef](#)]
95. Richman, J.S.; Moorman, J.; Physiology, C. Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol.-Heart Circ. Physiol.* **2000**, *278*, H2039–H2049. [[CrossRef](#)]
96. Schreiber, T. Measuring information transfer. *Phys. Rev. Lett.* **2000**, *85*, 461. [[CrossRef](#)] [[PubMed](#)]
97. Zhang, C.; Wang, H.; Fu, R. Automated detection of driver fatigue based on entropy and complexity measures. *IEEE Trans. Intell. Transp. Syst.* **2013**, *15*, 168–177. [[CrossRef](#)]
98. Tenenbaum, J.B.; Silva, V.d.; Langford, J. A global geometric framework for nonlinear dimensionality reduction. *Science* **2000**, *290*, 2319–2323. [[CrossRef](#)]
99. Abdulrahman, M.; Gwadabe, T.R.; Abdu, F.J.; Eleyan, A. Gabor wavelet transform based facial expression recognition using PCA and LBP. In Proceedings of the 2014 22nd Signal Processing and Communications Applications Conference (SIU), Trabzon, Turkey, 23–25 April 2014; pp. 2265–2268.
100. Arora, M.; Kumar, M. AutoFER: PCA and PSO based automatic facial emotion recognition. *Multimed. Tools Appl.* **2021**, *80*, 3039–3049. [[CrossRef](#)]
101. Yang, J.; Zhang, D.; Frangi, A.F.; Yang, J. Two-dimensional PCA: A new approach to appearance-based face representation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2004**, *26*, 131–137. [[CrossRef](#)] [[PubMed](#)]
102. Seng, K.P.; Ang, L.-M.; Ooi, C. A combined rule-based & machine learning audio-visual emotion recognition approach. *IEEE Trans. Affect. Comput.* **2016**, *9*, 3–13.
103. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
104. Dahmane, M.; Meunier, J. Emotion recognition using dynamic grid-based HoG features. In Proceedings of the 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG), Santa Barbara, CA, USA, 21–25 March 2011; pp. 884–888.
105. Kumar, P.; Happy, S.; Routray, A. A real-time robust facial expression recognition system using HOG features. In Proceedings of the 2016 International Conference on Computing, Analytics and Security Trends (CAST), Pune, India, 19–21 December 2016; pp. 289–293.
106. Hussein, H.I.; Dino, H.I.; Mstafa, R.J.; Hassan, M. Person-independent facial expression recognition based on the fusion of HOG descriptor and cuttlefish algorithm. *Multimed. Tools Appl.* **2022**, *81*, 11563–11586. [[CrossRef](#)]
107. Ahonen, T.; Hadid, A.; Pietikäinen, M. Face recognition with local binary patterns. In Proceedings of the European conference on Computer Vision, Prague, Czech Republic, 11–14 May 2004; pp. 469–481.
108. Chintalapati, S.; Raghunadh, M. Automated attendance management system based on face recognition algorithms. In Proceedings of the 2013 IEEE International Conference on Computational Intelligence and Computing Research, Enathi, India, 26–28 December 2013; pp. 1–5.

109. Swain, M.; Routray, A.; Kabisatpathy, P. Databases, features and classifiers for speech emotion recognition: A review. *Int. J. Speech Technol.* **2018**, *21*, 93–120. [[CrossRef](#)]
110. Molau, S.; Pitz, M.; Schluter, R.; Ney, H. Computing mel-frequency cepstral coefficients on the power spectrum. In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (cat. No. 01CH37221), Salt Lake City, UT, USA, 7–11 May 2001; pp. 73–76.
111. Wong, E.; Sridharan, S. Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification. In Proceedings of the 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing. ISIMP 2001 (IEEE Cat. No. 01EX489), Hong Kong, China, 4 May 2001; pp. 95–98.
112. Bandela, S.R.; Kumar, T.K. Emotion recognition of stressed speech using teager energy and linear prediction features. In Proceedings of the 2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT), Mumbai, India, 9–13 July 2018; pp. 422–425.
113. Idris, I.; Salam, M.S. Improved speech emotion classification from spectral coefficient optimization. In Proceedings of the Advances in Machine Learning and Signal Processing: Proceedings of MALSIP 2015, Ho Chi Minh City, Vietnam, 15–17 December 2015; pp. 247–257.
114. Feraru, S.M.; Zbancioc, M.D. Emotion recognition in Romanian language using LPC features. In Proceedings of the 2013 E-Health and Bioengineering Conference (EHB), Iasi, Romania, 21–23 November 2013; pp. 1–4.
115. Dey, A.; Chattopadhyay, S.; Singh, P.K.; Ahmadian, A.; Ferrara, M.; Sarkar, R. A hybrid meta-heuristic feature selection method using golden ratio and equilibrium optimization algorithms for speech emotion recognition. *IEEE Access* **2020**, *8*, 200953–200970. [[CrossRef](#)]
116. Bahoura, M.; Rouat, J. Wavelet speech enhancement based on the teager energy operator. *IEEE Signal Process. Lett.* **2001**, *8*, 10–12. [[CrossRef](#)]
117. Aouani, H.; Ayed, Y. Speech emotion recognition with deep learning. *Procedia Comput. Sci.* **2020**, *176*, 251–260. [[CrossRef](#)]
118. Li, X.; Li, X.; Zheng, X.; Zhang, D. EMD-TEO based speech emotion recognition. In Proceedings of the Life System Modeling and Intelligent Computing: International Conference on Life System Modeling and Simulation, LSMS 2010, and International Conference on Intelligent Computing for Sustainable Energy and Environment, ICSEE 2010, Wuxi, China, 17–20 September 2010; pp. 180–189.
119. Bandela, S.R.; Kumar, T.K. Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC. In Proceedings of the 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, India, 3–5 July 2017; pp. 1–5.
120. You, M.; Chen, C.; Bu, J.; Liu, J.; Tao, J. Emotion recognition from noisy speech. In Proceedings of the 2006 IEEE International Conference on Multimedia and Expo, Toronto, ON, Canada, 9–12 July 2006; pp. 1653–1656.
121. Schafer, R.W.; Rabiner, L. Digital representations of speech signals. *Proc. IEEE* **1975**, *63*, 662–677. [[CrossRef](#)]
122. Cochran, W.T.; Cooley, J.W.; Favin, D.L.; Helms, H.D.; Kaenel, R.A.; Lang, W.W.; Maling, G.C.; Nelson, D.E.; Rader, C.M.; Welch, P. What is the fast Fourier transform? *Proc. IEEE* **1967**, *55*, 1664–1674. [[CrossRef](#)]
123. Murugappan, M.; Murugappan, S. Human emotion recognition through short time Electroencephalogram (EEG) signals using Fast Fourier Transform (FFT). In Proceedings of the 2013 IEEE 9th International Colloquium on Signal Processing and Its Applications, Kuala Lumpur, Malaysia, 8–10 March 2013; pp. 289–294.
124. Acharya, D.; Billimoria, A.; Srivastava, N.; Goel, S.; Bhardwaj, A. Emotion recognition using fourier transform and genetic programming. *Appl. Acoust.* **2020**, *164*, 107260. [[CrossRef](#)]
125. Khare, S.K.; Bajaj, V. Time–frequency representation and convolutional neural network-based emotion recognition. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 2901–2909. [[CrossRef](#)] [[PubMed](#)]
126. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [[CrossRef](#)] [[PubMed](#)]
127. Zheng, W.-L.; Zhu, J.-Y.; Lu, B. Identifying stable patterns over time for emotion recognition from EEG. *IEEE Trans. Affect. Comput.* **2017**, *10*, 417–429. [[CrossRef](#)]
128. Atkinson, J.; Campos, D. Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers. *Expert Syst. Appl.* **2016**, *47*, 35–41. [[CrossRef](#)]
129. Alazrai, R.; Homoud, R.; Alwanni, H.; Daoud, M. EEG-based emotion recognition using quadratic time-frequency distribution. *Sensors* **2018**, *18*, 2739. [[CrossRef](#)] [[PubMed](#)]
130. Liu, Z.-T.; Xie, Q.; Wu, M.; Cao, W.-H.; Mei, Y.; Mao, J.-W. Speech emotion recognition based on an improved brain emotion learning model. *Neurocomputing* **2018**, *309*, 145–156. [[CrossRef](#)]
131. He, X.; Niyogi, P. Locality preserving projections. *Adv. Neural Inf. Process. Syst.* **2003**, *16*, 2.
132. Kira, K.; Rendell, L.A. A practical approach to feature selection. In *Machine Learning Proceedings 1992*; Elsevier: Amsterdam, The Netherlands, 1992; pp. 249–256.
133. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [[CrossRef](#)]
134. Soentpiet, R. *Advances in Kernel Methods: Support Vector Learning*; MIT Press: Cambridge, MA, USA, 1999.
135. Chen, D.-R.; Wu, Q.; Ying, Y.; Zhou, D.-X. Support vector machine soft margin classifiers: Error analysis. *J. Mach. Learn. Res.* **2004**, *5*, 1143–1175.
136. Pan, Y.; Shen, P.; Shen, L. Speech Emotion Recognition Using Support Vector Machine. *Int. J. Smart Home* **2012**, *6*, 101–108.

137. Bitouk, D.; Verma, R.; Nenkova, A. Class-level spectral features for emotion recognition. *Speech Commun.* **2010**, *52*, 613–625. [[CrossRef](#)]
138. Ghimire, D.; Jeong, S.; Lee, J.; Park, S. Facial expression recognition based on local region specific features and support vector machines. *Multimed. Tools Appl.* **2017**, *76*, 7803–7821. [[CrossRef](#)]
139. Desmet, B.; Hoste, V. Emotion detection in suicide notes. *Expert Syst. Appl.* **2013**, *40*, 6351–6358. [[CrossRef](#)]
140. Dempster, A.P.; Laird, N.M.; Rubin, D. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **1977**, *39*, 1–22.
141. Hu, H.; Xu, M.-X.; Wu, W. GMM supervector based SVM with spectral features for speech emotion recognition. In Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, Honolulu, HI, USA, 15–20 April 2007; pp. IV-413–IV-416.
142. Shahin, I.; Nassif, A.B.; Hamsa, S. Emotion recognition using hybrid Gaussian mixture model and deep neural network. *IEEE Access* **2019**, *7*, 26777–26787. [[CrossRef](#)]
143. Zhang, C.; Li, M.; Wu, D. Federated Multidomain Learning With Graph Ensemble Autoencoder GMM for Emotion Recognition. *IEEE Trans. Intell. Transp. Syst. Early Access.* [[CrossRef](#)]
144. Cohen, I.; Garg, A.; Huang, T.S. Emotion recognition from facial expressions using multilevel HMM. In Proceedings of the Neural Information PROCESSING systems, Breckenridge, CO, USA, 1–2 December 2000.
145. Schuller, B.; Rigoll, G.; Lang, M. Hidden Markov model-based speech emotion recognition. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03), Hong Kong, China, 6–10 April 2003.
146. Wu, C.-H.; Lin, J.-C.; Wei, W. Two-level hierarchical alignment for semi-coupled HMM-based audiovisual emotion recognition with temporal course. *IEEE Trans. Multimed.* **2013**, *15*, 1880–1895. [[CrossRef](#)]
147. Tang, D.; Zhang, Z.; He, Y.; Lin, C.; Zhou, D. Hidden topic–emotion transition model for multi-level social emotion detection. *Knowl.-Based Syst.* **2019**, *164*, 426–435. [[CrossRef](#)]
148. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
149. 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. *Inf. Sci.* **2020**, *509*, 150–163. [[CrossRef](#)]
150. Katsis, C.D.; Katertsidis, N.S.; Fotiadis, D. An integrated system based on physiological signals for the assessment of affective states in patients with anxiety disorders. *Biomed. Signal Process. Control* **2011**, *6*, 261–268. [[CrossRef](#)]
151. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [[CrossRef](#)]
152. Kwon, S. Att-Net: Enhanced emotion recognition system using lightweight self-attention module. *Proc. IEEE Haffner* **2021**, *102*, 107101.
153. Kollias, D.; Zafeiriou, S. Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset. *IEEE Trans. Affect. Comput.* **2020**, *12*, 595–606. [[CrossRef](#)]
154. Zhang, S.; Zhang, S.; Huang, T.; Gao, W. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Trans. Multimed.* **2017**, *20*, 1576–1590. [[CrossRef](#)]
155. Song, T.; Zheng, W.; Song, P.; Cui, Z. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Trans. Affect. Comput.* **2018**, *11*, 532–541. [[CrossRef](#)]
156. Salama, E.S.; El-Khoribi, R.A.; Shoman, M.E.; Shalaby, M. A 3D-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition. *Egypt. Inform. J.* **2021**, *22*, 167–176. [[CrossRef](#)]
157. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)] [[PubMed](#)]
158. Araño, K.A.; Gloor, P.; Orsenigo, C.; Vercellis, C. When old meets new: Emotion recognition from speech signals. *Cogn. Comput.* **2021**, *13*, 771–783. [[CrossRef](#)]
159. Zhang, Y.; Chen, J.; Tan, J.H.; Chen, Y.; Chen, Y.; Li, D.; Yang, L.; Su, J.; Huang, X.; Che, W. An investigation of deep learning models for EEG-based emotion recognition. *Front. Neurosci.* **2020**, *14*, 622759. [[CrossRef](#)] [[PubMed](#)]
160. Li, D.; Liu, J.; Yang, Z.; Sun, L.; Wang, Z. Speech emotion recognition using recurrent neural networks with directional self-attention. *Expert Syst. Appl.* **2021**, *173*, 114683. [[CrossRef](#)]
161. Kim, Y.; Lee, H.; Provost, E.M. Deep learning for robust feature generation in audiovisual emotion recognition. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 21 October 2013; pp. 3687–3691.
162. Hassan, M.M.; Alam, M.G.R.; Uddin, M.Z.; Huda, S.; Almogren, A.; Fortino, G. Human emotion recognition using deep belief network architecture. *Inf. Fusion* **2019**, *51*, 10–18. [[CrossRef](#)]
163. Liu, D.; Chen, L.; Wang, Z.; Diao, G. Speech expression multimodal emotion recognition based on deep belief network. *J. Grid Comput.* **2021**, *19*, 1–13. [[CrossRef](#)]
164. Uddin, M.Z.; Hassan, M.M.; Almogren, A.; Alamri, A.; Alrubaian, M.; Fortino, G. Facial expression recognition utilizing local direction-based robust features and deep belief network. *IEEE Access* **2017**, *5*, 4525–4536. [[CrossRef](#)]
165. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 5485–5551.

166. Jiang, H.; He, P.; Chen, W.; Liu, X.; Gao, J.; Zhao, T. SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Washington, DC, USA, 5–10 July 2020; pp. 2177–2190.
167. Shukla, A.; Vougioukas, K.; Ma, P.; Petridis, S.; Pantic, M. Visually guided self supervised learning of speech representations. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6299–6303.
168. Ryumina, E.; Dresvyanskiy, D.; Karpov, A. In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study. *Neurocomputing* **2022**, *514*, 435–450. [[CrossRef](#)]
169. Chudasama, V.; Kar, P.; Gudmalwar, A.; Shah, N.; Wasnik, P.; Onoe, N. M2FNet: Multi-modal fusion network for emotion recognition in conversation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 21–24 June 2022; pp. 4652–4661.
170. Majumder, N.; Hazarika, D.; Gelbukh, A.; Cambria, E.; Poria, S. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowl.-Based Syst.* **2018**, *161*, 124–133. [[CrossRef](#)]
171. Song, X.; Zang, L.; Zhang, R.; Hu, S.; Huang, L. Emotionflow: Capture the dialogue level emotion transitions. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 23–27 May 2022; pp. 8542–8546.
172. Ding, H.; Zhou, S.K.; Chellappa, R. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 118–126.
173. Savchenko, A.V.; Savchenko, L.V.; Makarov, I. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Trans. Affect. Comput.* **2022**, *13*, 2132–2143. [[CrossRef](#)]
174. Zhang, Y.; Wang, C.; Ling, X.; Deng, W. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In Proceedings of the Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; pp. 418–434.
175. Li, Y.; Wang, L.; Zheng, W.; Zong, Y.; Qi, L.; Cui, Z.; Zhang, T.; Song, T.; Systems, D. A novel bi-hemispheric discrepancy model for EEG emotion recognition. *IEEE Trans. Cogn. Dev. Syst.* **2020**, *13*, 354–367. [[CrossRef](#)]
176. Paraskevopoulos, G.; Georgiou, E.; Potamianos, A. Mmlatch: Bottom-up top-down fusion for multimodal sentiment analysis. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 23–27 May 2020; pp. 4573–4577.
177. Shin, H.-C.; Roth, H.R.; Gao, M.; Lu, L.; Xu, Z.; Nogue, I.; Yao, J.; Mollura, D.; Summers, R. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans. Med. Imaging* **2016**, *35*, 1285–1298. [[CrossRef](#)]
178. El Ayadi, M.; Kamel, M.S.; Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognit.* **2011**, *44*, 572–587. [[CrossRef](#)]
179. Ringeval, F.; Sonderegger, A.; Sauer, J.; Lalanne, D. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Shanghai, China, 22–26 April 2013; pp. 1–8.
180. Coan, J.A.; Allen, J.J. *Handbook of Emotion Elicitation and Assessment*; Oxford University Press: Oxford, UK, 2007.
181. Douglas-Cowie, E.; Cowie, R.; Schröder, M. A new emotion database: Considerations, sources and scope. In Proceedings of the ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, Newcastle, Northern Ireland, UK, 5–7 September 2000.
182. Grimm, M.; Kroschel, K.; Narayanan, S. The Vera am Mittag German audio-visual emotional speech database. In Proceedings of the 2008 IEEE International Conference on Multimedia and Expo, Hannover, Germany, 26 August 2008; pp. 865–868.
183. Lee, C.M.; Narayanan, S.S. Toward detecting emotions in spoken dialogs. *IEEE Trans. Speech Audio Process.* **2005**, *13*, 293–303.
184. Dredze, M.; Crammer, K.; Pereira, F. Confidence-weighted linear classification. In Proceedings of the 25th International Conference on Machine Learning, New York, NY, USA, 5–9 July 2008; pp. 264–271.
185. Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C.D.; Ng, A.Y.; Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 1631–1642.
186. Maas, A.; Daly, R.E.; Pham, P.T.; Huang, D.; Ng, A.Y.; Potts, C. Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 142–150.
187. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. In Proceedings of the Interspeech, Lisbon, Portugal, 4–8 September 2005; pp. 1517–1520.
188. Jackson, P. *Surrey Audio-Visual Expressed Emotion (Savee) Database*; University of Surrey: Guildford, UK, 2014.
189. Cao, H.; Cooper, D.G.; Keutmann, M.K.; Gur, R.C.; Nenkova, A.; Verma, R. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Trans. Affect. Comput.* **2014**, *5*, 377–390. [[CrossRef](#)]
190. Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [[CrossRef](#)]
191. Tao, J.; Kang, Y.; Li, A. Prosody conversion from neutral speech to emotional speech. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 1145–1154.

192. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 527–536.
193. Batliner, A.; Steidl, S.; Nöth, E. Releasing a Thoroughly Annotated and Processed Spontaneous Emotional Database: The FAU Aibo Emotion Corpus. In Proceedings of the Satellite Workshop of LREC 2008 on Corpora for Research on Emotion and Affect, Marrakech, Morocco, 26–27&31 May–1 June 2008; pp. 28–31.
194. McKeown, G.; Valstar, M.; Cowie, R.; Pantic, M.; Schroder, M. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Trans. Affect. Comput.* **2011**, *3*, 5–17. [[CrossRef](#)]
195. Li, Y.; Tao, J.; Chao, L.; Bao, W.; Liu, Y. CHEAVD: A Chinese natural emotional audio–visual database. *J. Ambient Intell. Humaniz. Comput.* **2017**, *8*, 913–924. [[CrossRef](#)]
196. Pantic, M.; Valstar, M.; Rademaker, R.; Maat, L. Web-based database for facial expression analysis. In Proceedings of the 2005 IEEE international conference on multimedia and Expo, Amsterdam, The Netherlands, 6 July 2005.
197. Zhang, X.; Yin, L.; Cohn, J.F.; Canavan, S.; Reale, M.; Horowitz, A.; Liu, P.; Girard, J.M.J.I.; Computing, V. Bp4d-spontaneous: A high-resolution spontaneous 3d dynamic facial expression database. *Image Vis. Comput.* **2014**, *32*, 692–706. [[CrossRef](#)]
198. 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 expression. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 94–101.
199. Zhang, X.; Yin, L.; Cohn, J.F.; Canavan, S.; Reale, M.; Horowitz, A.; Liu, P. A high-resolution spontaneous 3d dynamic facial expression database. In Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Shanghai, China, 22–26 April 2013; pp. 1–6.
200. Kossaifi, J.; Walecki, R.; Panagakis, Y.; Shen, J.; Schmitt, M.; Ringeval, F.; Han, J.; Pandit, V.; Toisoul, A.; Schuller, B.; et al. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1022–1040. [[CrossRef](#)] [[PubMed](#)]
201. Valstar, M.; Pantic, M. Induced disgust, happiness and surprise: An addition to the mmi facial expression database. In Proceedings of the 3rd International Workshop on EMOTION (Satellite of LREC): Corpora for Research on Emotion and Affect, Valletta, Malta, 23 May 2010; p. 65.
202. Lyons, M.; Akamatsu, S.; Kamachi, M.; Gyoba, J. Coding facial expressions with gabor wavelets. In Proceedings of the Proceedings Third IEEE international Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998; pp. 200–205.
203. 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 (FGR06), Southampton, UK, 10–12 April 2006; pp. 211–216.
204. Mollahosseini, A.; Hasani, B.; Mahoor, M. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **2017**, *10*, 18–31. [[CrossRef](#)]
205. Li, S.; Deng, W.; Du, J. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 22–25 July 2017; pp. 2852–2861.
206. Fabian Benitez-Quiroz, C.; Srinivasan, R.; Martinez, A.M. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5562–5570.
207. Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.-S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; Patras, I. Deap: A database for emotion analysis; using physiological signals. *IEEE Trans. Affect. Comput.* **2011**, *3*, 18–31. [[CrossRef](#)]
208. Abadi, M.K.; Subramanian, R.; Kia, S.M.; Avesani, P.; Patras, I.; Sebe, N. DECAF: MEG-based multimodal database for decoding affective physiological responses. *IEEE Trans. Affect. Comput.* **2015**, *6*, 209–222. [[CrossRef](#)]
209. Miranda-Correa, J.A.; Abadi, M.K.; Sebe, N.; Patras, I. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Trans. Affect. Comput.* **2018**, *12*, 479–493. [[CrossRef](#)]
210. Zheng, W.-L.; Lu, B.-L. A multimodal approach to estimating vigilance using EEG and forehead EOG. *J. Neural Eng.* **2017**, *14*, 026017. [[CrossRef](#)]
211. Gouveia, C.; Tomé, A.; Barros, F.; Soares, S.C.; Vieira, J.; Pinho, P. Study on the usage feasibility of continuous-wave radar for emotion recognition. *Biomed. Signal Process. Control* **2020**, *58*, 101835. [[CrossRef](#)]
212. Mercuri, M.; Lorato, I.R.; Liu, Y.-H.; Wieringa, F.; Hoof, C.V.; Torfs, T. Vital-sign monitoring and spatial tracking of multiple people using a contactless radar-based sensor. *Nat. Electron.* **2019**, *2*, 252–262. [[CrossRef](#)]
213. Dang, X.; Chen, Z.; Hao, Z. Emotion recognition method using millimetre wave radar based on deep learning. *IET Radar Sonar Navig.* **2022**, *16*, 1796–1808. [[CrossRef](#)]

214. Zadeh, A.B.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.-P. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 2236–2246.
215. Soleymani, M.; Lichtenauer, J.; Pun, T.; Pantic, M. A multimodal database for affect recognition and implicit tagging. *IEEE Trans. Affect. Comput.* **2011**, *3*, 42–55. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.