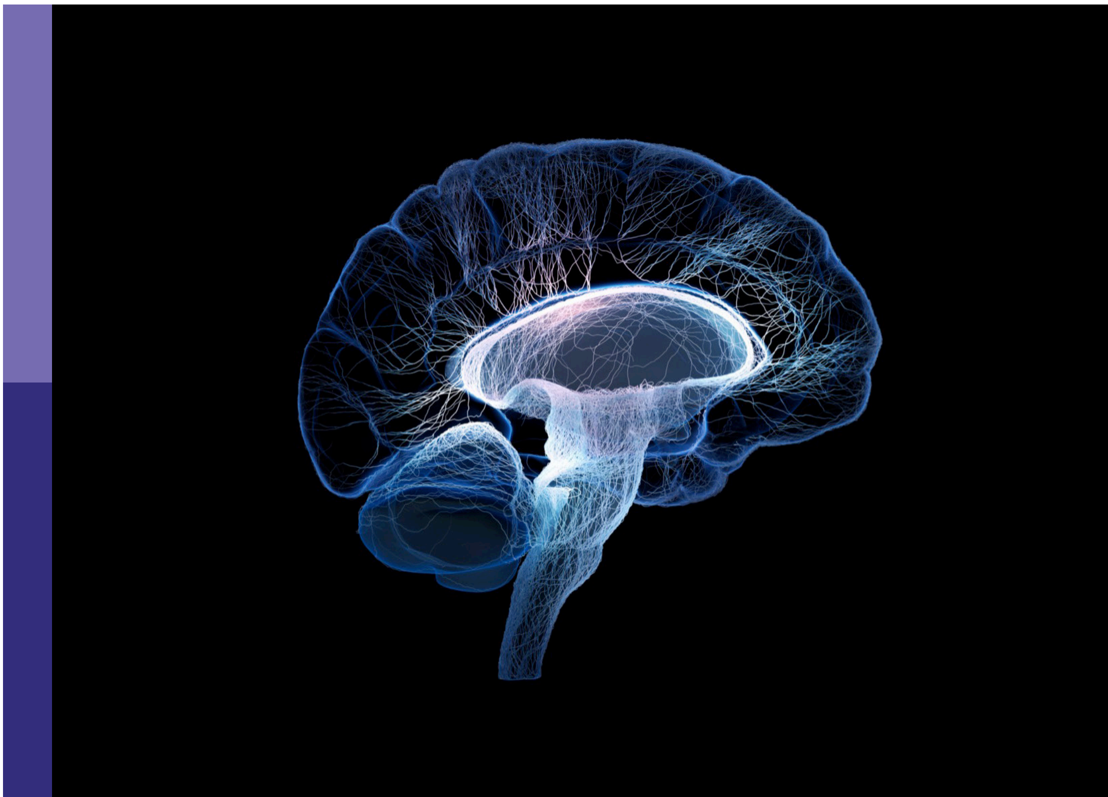
frontiers | Research Topics

Deep learning techniques applied to affective computing

Edited by
Shiqing Zhang, Xiaopeng Hong, Xiaobai Li, Zhen Cui
and Wenming Zheng

Published in
Frontiers in Neuroscience
Frontiers in Psychology
Frontiers in Neurorobotics



ISSN 1664-8714
ISBN 978-2-8325-2636-1
DOI 10.3389/978-2-8325-2636-1

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Deep learning techniques applied to affective computing

Topic editors

Shiqing Zhang — Taizhou University, China

Xiaopeng Hong — Harbin Institute of Technology, China

Xiaobai Li — University of Oulu, Finland

Zhen Cui — Nanjing University of Science and Technology, China

Wenming Zheng — Southeast University, China

Citation

Zhang, S., Hong, X., Li, X., Cui, Z., Zheng, W., eds. (2023). *Deep learning techniques applied to affective computing*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-2636-1

Table of contents

frontiers | Frontiers in Neuroscience

TYPE Review
PUBLISHED 26 August 2022
DOI 10.3389/fnins.2022.984404

Affective video recommender systems: A survey

Dandan Wang and Xiaoming Zhao*

Department of Computer Science, Taizhou University, Taizhou, China

OPEN ACCESS

EDITED BY
Zhen Cui,
Nanjing University of Science and Technology, China

REVIEWED BY
Yingqian Zhang,
Xiamen University Tan Kah Kee College, China
Kun Lang,
Dalian Maritime University, China

*CORRESPONDENCE
Xiaoming Zhao
tzxyzxm@163.com

SPECIALTY SECTION
This article was submitted to Neural Technology, a section of the journal Frontiers in Neuroscience

RECEIVED 02 July 2022
ACCEPTED 10 August 2022
PUBLISHED 26 August 2022

CITATION
Wang D and Zhao X (2022) Affective video recommender systems: A survey. *Front. Neurosci.* 16:984404. doi: 10.3389/fnins.2022.984404

Traditional video recommendation provides the viewers with customized media content according to their historical records (e.g., ratings, reviews). However, such systems tend to generate terrible results if the data is insufficient, which leads to a cold-start problem. An affective video recommender system (AVRS) is a multidiscipline and multimodal human-robot interaction (HRI) system, and it incorporates physical, physiological, neuroscience, and computer science subjects and multimedia resources, including text, audio, and video. As a promising research domain, AVRS employs advanced affective analysis technologies in video resources; therefore, it can solve the cold-start problem. In AVRS, the viewers' emotional responses can be obtained from various techniques, including physical signals (e.g., facial expression, gestures, and speech) and internal signals (e.g., physiological signals). The changes in these signals can be detected when the viewers face specific situations. The physiological signals are a response to central and autonomic nervous systems and are mostly involuntarily activated, which cannot be easily controlled. Therefore, it is suitable for reliable emotion analysis. The physical signals can be recorded by a webcam or recorder. In contrast, the physiological signals can be collected by various equipment, e.g., psychophysiological heart rate (HR) signals calculated by echocardiogram (ECG), electro-dermal activity (EDA), and brain activity (GA) from electroencephalography (EEG) signals, skin conductance response (SCR) by a galvanic skin response (GSR), and photoplethysmography (PPG) estimating users' pulse. This survey aims to provide a comprehensive overview of the AVRS domain. To analyze the recent efforts in the field of affective video recommendation, we collected 92 relevant published articles from Google Scholar and summarized the articles and their key findings. In this survey, we feature these articles concerning AVRS from different perspectives, including various traditional recommendation algorithms and advanced deep learning-based algorithms, the commonly used affective video recommendation databases, audience response categories, and evaluation methods. Finally, we conclude the challenge of AVRS and provide the potential future research directions.

KEYWORDS

video recommendation, affective video recommender system, multidiscipline, multimodal, neuroscience, affective analysis, physiological signals, deep learning

Introduction

Emotion or affection is a mental state which is unconscious and spontaneously arises accompanied by physiological and psychological status changes in human organs and tissues, e.g., heart rate, facial expression, brain, etc. (Shu et al., 2018b). Emotions are universal and have proved to be a highly multidisciplinary research field, from psychology, sociology, and neuroscience to computer science (Baveye et al., 2018). The emotional state of a consumer determines his behavior and decision-making process, i.e., click, purchase, or close. However, the viewer's emotional state is ignored in the recommendation process because of the complexity of the mutual interaction of physiological signals with human emotions. The subtle emotional expression is straightforward to be misunderstood. Previous studies have mainly focused on users' affection by ratings (Roy and Guntuku, 2016), comments (Orellana-Rodriguez et al., 2015), helpfulness votes, etc. However, acquiring this feedback requires users' cooperation, and some require plenty of time. Therefore, the amount of such feedback data is limited and faced with a cold-start problem.

Recent research employs techniques closely related to neuroscience and human-robot interaction (HRI). The viewers' emotional states are obtained from analyzing their physical and internal signal parameters with the help of various equipment. For example, researchers apply photoplethysmography (PPG) to estimate users' pulse by using the fluctuations in skin color related to blood volume and the proportion of reflected light (Bohlin et al., 2019). Dabas et al. (2018) studied human emotions with the help of electroencephalogram (EEG) signals. De Pessemier et al. (2019) facilitated HRI for users to watch videos by an automated procedure based on facial recognition. The automatic feedback is gathered when users play the videos using a front-facing camera. The viewer's physiological data is easy to get and can be obtained by several methods without the user's active cooperation in the viewing process. The physiological data can be achieved by measuring body parameters, including skin estimated pulse, heart rate, mood, motion, shot change rate, and sound energy. The viewers' psychophysiological signals of heart rate (HR) were calculated from an echocardiogram (ECG), while electro-dermal activity (EDA) and brain activity (BA) in EEG signals (Đorđević Čegar et al., 2020). Facial expressions or features can be obtained by a camera (Tkalčič et al., 2013a).

The affective computing technology promotes the rapid development of the affective video recommender systems (AVRSs). An AVRS is a new trending research direction of recommender families in recent years. Unlike text, image, and speech emotion recognition (Zhang et al., 2022), AVRS mainly analyzes the emotional states in videos and detects emotional reactions according to different scenes. An AVRS recommends video resources that viewers may be interested in based on the recognized emotional states. As a new branch of affective analysis and recommender systems, it is necessary to define AVRS according to previous literature research.

Definition 1: AVRS: is a multidiscipline and multimodal HRI system that videos are recommended based on the reviewers' emotional responses (implicit or explicit), e.g., physical, physiological signals, comments, etc.

The physical data reflect communicative signals, e.g., facial expressions, speech detection, and eye-tracking while viewing the video (Lim et al., 2020). In contrast, the physiological signals record body variations, e.g., heart rate, temperature, and blood pressure changes. These physical and physiological signals and comments are recognized and interpreted into emotional states. The AVRS recommends the videos based on emotion models according to the viewers' emotional states.

The differences between this survey and former studies

An AVRS is a relatively new recommender family branch that has begun to develop in recent years. At present, there are few comprehensive reviews related to affective video recommendations. Most works mainly focus on different domains of recommender systems, including recommender systems (Singh et al., 2021), the application of deep learning in recommender systems (Guo et al., 2017), tourism recommendation systems based on emotion recognition (Santamaria-Granados et al., 2021), affective recommender system techniques (Raheem and Ali, 2020), etc.

As shown in **Table 1**, we compare different aspects of our survey and recently existing related reviews, i.e., multimodal feature, multimodal data sources, deep learning methods, affective computing, multidiscipline knowledge, and video contents. Singh et al. (2021) mainly focused on different recommendation methods and existing problems without involving multimodal features, multimodal data sources, and multidiscipline knowledge. Zhang et al. (2019) provided a review of deep learning-based recommendations. However, they failed to supply multimodal data sources, affective computing, and multidiscipline knowledge. In Santamaria-Granados et al. (2021), they explored the emotional recognition of recommender systems in the tourist scenario. They provided guidelines for establishing emotion-sensitive tourist recommender systems. Unfortunately, they only cover a few publications related to multimodal data sources and video content. The contribution of Raheem and Ali (2020) is one of very few research works in the field of affective recommendation; they introduced the application of recommendation technology based on affective computing. However, Raheem and Ali (2020) haven't explored multimodal

TABLE 1 Comparisons between this survey and existing reviews.

Main concerns	Singh et al. (2021)	Zhang et al. (2019)	Santamaria-Granados et al. (2021)	Raheem and Ali (2020)	Our survey
Multimodal feature	×	✓	✓	✓	✓
Multimodal data sources	×	×	Few	×	✓
Deep learning methods	✓	✓	✓	✓	✓
Affective computing	✓	×	✓	✓	✓
Multidiscipline knowledge	×	×	✓	×	✓
Video content	✓	✓	Few	✓	✓

data sources and multidiscipline knowledge. This survey aims to provide a comprehensive review of current research on AVRS, to discuss the open problems and limitations, and point out future possible directions.

The method of collecting relevant publications and the distribution

The relevant publications in this survey are obtained from Google scholar and published by Science Direct, Springer, IEEE, ACM, etc. The collected publications are from 2009 to 2022; filters are applied to the search engine by subject (affection, emotion, sentiment, affective computing, video recommendation, recommender systems). Table 2 illustrates the number of publications and the percentage from different sources.

We collected 92 non-repeated publications related to AVRS. Most of the articles are from IEEE, accounting for 38.04%, more than three times that of ACM and Elsevier. The distribution of publications from ACM, Elsevier, and Springer is similar, accounting for about 11–17%. The remaining publications are from various published websites. It can be seen from Table 2 that the number of publications related to AVRS is relatively limited compared with other fields of recommender systems, and it is thus in its infancy, which requires a large number of researchers and their outstanding work.

The distribution of AVRS publications is shown in Figure 1. The x-axis represents the year of publication, and the y-axis represents the total number of publications in the corresponding year. As we can see from Figure 1, the number of research works on AVRS is scarce. Since the relevant articles were published in 2009, there have been no more than ten published articles every year except in 2018, reaching the peak of 14 in 2018 and showing an apparent downward trend afterward. The publication distribution in Figure 1 also indicates that the prosperity of AVRS currently requires a great deal of academic dedication.

TABLE 2 Publications from different sources.

Databases	Number of publications	Percentage
ACM	11	11.96%
IEEE	35	38.04%
Elsevier	11	11.96%
Springer	15	16.30%
Others	20	21.74%
Total	92	100%

Contributions of this survey

This survey provides a concise, comprehensive understanding of the latest AVRS research and gives dynamic guidelines in AVRS for scientific researchers, practitioners, and developers interested in video recommendations. We define the internal logic and operating mechanism of various models and algorithms, the classification of existing technologies and their characteristics, the databases for affective computing, the types of audience responses, and the evaluation metrics. The main contributions of this survey are summarized in the following three aspects:

(1) We systematically summarize and overview the current techniques in the affective video recommendation field.
(2) We classified the works of literature related to different models and algorithms, the possible database resources for video recommendation, the types of audience responses, and the evaluation metrics.
(3) We show the current challenges in the video recommendation field and envision possible future research directions.

The structure of this survey is arranged in the following: Section 2 introduces currently-used algorithms and models of video recommender systems; Section 3 shows the database resources commonly used in the research of AVRS; Section 4 classifies the ways to obtain user responses in publications; Section 5 summarizes the evaluation metrics of recommendation effect in different publications; Section 6 analyzes the challenges in the current research and discusses future research directions.

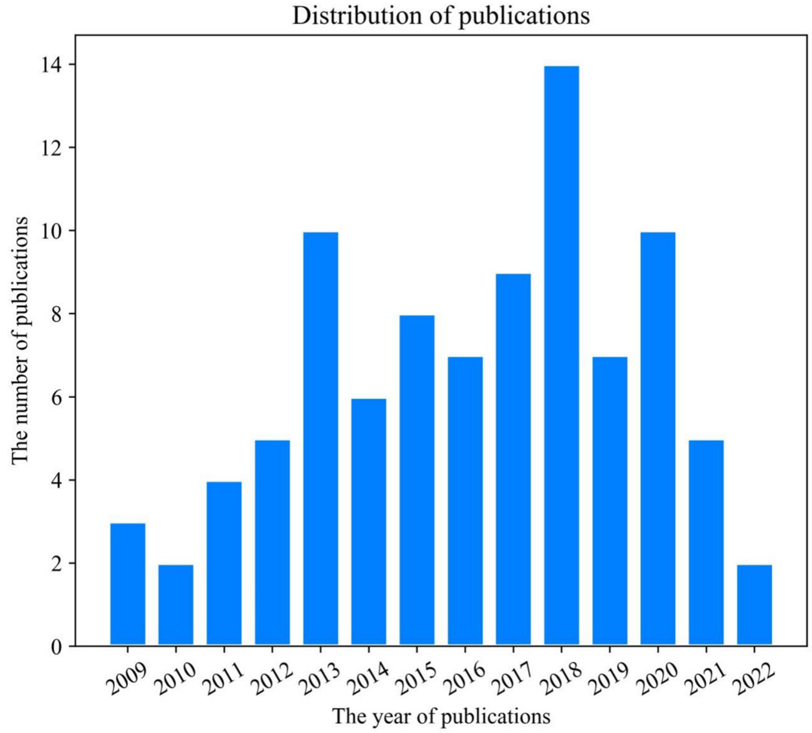


FIGURE 1
Distribution of publications in AVRS.

The state-of-the-art affective video recommendation algorithms and models

Video recommendation is based on video features and the viewers' profiles. According to video clips, the viewers' emotions are challenging to be captured simultaneously. Therefore, an AVRS is a more complex domain in recommender systems. Several researchers tend to solve the AVRS problem by various methods, traditional models, or algorithms, including support vector machine/support vector regression (SVM/SVR) (Arapakis et al., 2009a), clustering (Song and Yang, 2022), AdaBoost (Zhao et al., 2013), matrix-based algorithm (MA) (Kaklauskas et al., 2018), collaborative filtering (CF) (Diaz et al., 2018), content-based filtering (CBF) (Deldjoo et al., 2018), knowledge graph (KG) (Breitfuss et al., 2021), genetic algorithms (GA) (Wang and Chen, 2020), hybrid recommendation systems (HRS) (Wakil et al., 2015), the combination of several traditional recommendation algorithms, etc. Deep learning (DL) has gradually penetrated the field of affective computing and promoted the development of video recommendations. Deep learning-based models applied in AVRS in recent years include reinforcement learning (RL) (Leite et al., 2022), convolutional neural network (CNN) (Zhu et al., 2019), long short-term memory (LSTM) (Cao et al., 2022), multilayer perception (Đorđević Čegar et al., 2020) (MLP), deep hybrid models (DHM) (Mishra et al., 2020), etc. The evolution of AVRS with different algorithms and databases is illustrated in **Figure 2**.

Table 3 shows the publications of this survey based on different techniques. From **Table 3**, we can infer that the most general approaches used for video recommendation are SVM/SVR, MA, CF, and CBF methods. Other models adopted in AVRS are relatively rare, especially the work of deep learning-based algorithms.

In this section, we classify the publications according to the adopted algorithms or models. We first introduce several commonly-used traditional video recommendation algorithms, then describe the application of prevalent deep learning algorithms in AVRS, and analyze the advantages and disadvantages of both conventional recommendation algorithms and deep learning algorithms.

Traditional methods

Support vector machine (SVM) or support vector regression (SVR)

The fundamental idea of implementing SVM and SVR is classifying the mixed input features to predict the users' emotional states during their interaction with the robots.

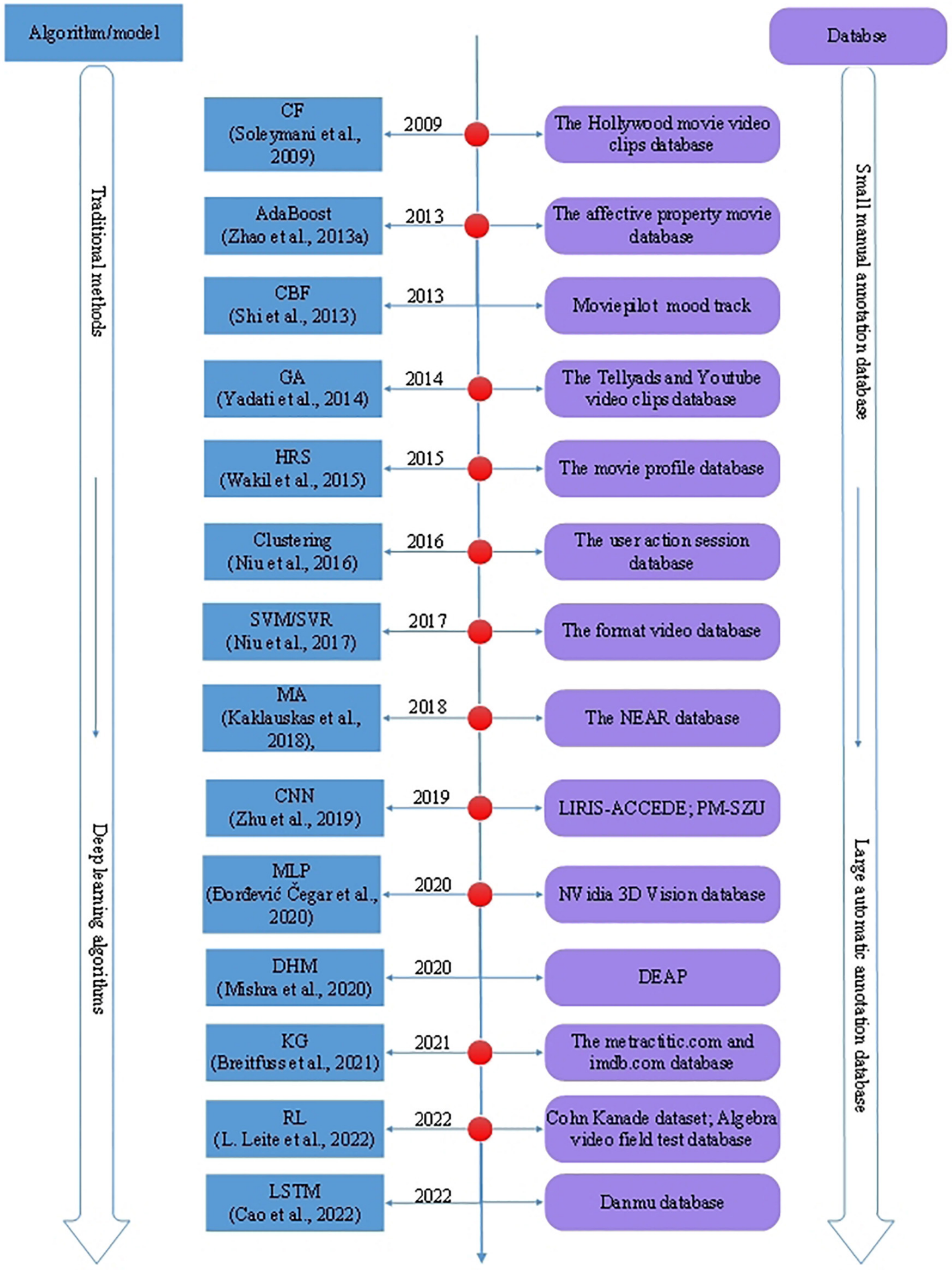

Algorithm/model
Databse
Traditional methods
Deep learning algorithms
Small manual annotation database
Large automatic annotation database
CF (Soleymani et al., 2009)
2009
The Hollywood movie video clips database
AdaBoost (Zhao et al., 2013a)
2013
The affective property movie database
CBF (Shi et al., 2013)
2013
Moviepilot mood track
GA (Yadati et al., 2014)
2014
The Tellyads and Youtube video clips database
HRS (Wakil et al., 2015)
2015
The movie profile database
Clustering (Niu et al., 2016)
2016
The user action session database
SVM/SVR (Niu et al., 2017)
2017
The format video database
MA (Kaklauskas et al., 2018),
2018
The NEAR database
CNN (Zhu et al., 2019)
2019
LIRIS-ACCEDE; PM-SZU
MLP (Đorđević Čegar et al., 2020)
2020
NVidia 3D Vision database
DHM (Mishra et al., 2020)
2020
DEAP
KG (Breitfuss et al., 2021)
2021
The metractitic.com and imdb.com database
RL (L. Leite et al., 2022)
2022
Cohn Kanade dataset; Algebra video field test database
LSTM (Cao et al., 2022)
2022
Danmu database

FIGURE 2
The evolution of AVRS with different algorithms and databases.

An SVM/SVR is one of the most widely-used techniques in the affective video recommendation domain. Researchers devoted valuable efforts to promoting the performances of video recommendations based on SVM/SVR. In Arapakis et al. (2009a), they trained a two-layer hierarchical SVM model by using interactive data, context information, and user response

TABLE 3 Publications based on different techniques.

Categories	Algorithm/Model	Publications
Traditional methods	SVM/SVR	Arapakis et al., 2009a,b; Soleymani and Pantic, 2012; Soleymani et al., 2012; Srivastava and Roy, 2014; Sivakumar et al., 2015; Niu et al., 2017; Dabas et al., 2018; Bohlin et al., 2019
	Clustering	Niu et al., 2013; Niu et al., 2016
	AdaBoost	Zhao et al., 2013; Shu et al., 2018a
	MA	Tkalčič et al., 2013b; Dnodxvndv et al., 2018; Kaklauskas et al., 2018
	CF	Soleymani et al., 2009; Winoto and Tang, 2010; Tkalčič et al., 2013b; Choi et al., 2016; Diaz et al., 2018
	CBF	Shi et al., 2013; Tkalčič et al., 2013b; Deldjoo et al., 2018
	KG	Breitfuss et al., 2021; Qi et al., 2021
	GA	Yadati et al., 2014
	HRS	Mugellini et al., 2014; Wakil et al., 2015
Deep learning-based methods	RL	Tripathi et al., 2018; Leite et al., 2022
	CNN	Hewitt and Gunes, 2018; Kwon et al., 2018; Yang et al., 2019; Zhu et al., 2019
	LSTM	Alhagry, 2017; Zhang and Zhang, 2017; Ogawa et al., 2018; Nie et al., 2020; Wang et al., 2020; Cao et al., 2022
	MLP	Boughrara et al., 2016; Đorđević Čegar et al., 2020; Krishnamurthy, 2020
	DHM	Fan et al., 2016; Liu et al., 2017; Yenter, 2017; Zhang et al., 2018; Tripathi et al., 2019; Mishra et al., 2020

to determine whether the user's unknown video is relevant or not. In Bohlin et al. (2019), they a support vector classifier was used to predict the ratings of video viewers and whether they will watch similar videos.

In Arapakis et al. (2009b), they leveraged a two-layer hierarchical SVM model to discriminate whether the video is relevant to a user. The real-time facial expressions were adopted for constructing a face model and classified into seven emotion categories. The classification results were forwarded to an SVM model to determine whether the videos were relevant or not. In Dabas et al. (2018), the authors classified users' emotions when watching musical videos by constructing a 3D emotional model consisting of several octants including eight emotional states, i.e., relaxed, peaceful, bored, disgusted, nervous, sad, surprised, and excited. The human emotions were studied using EEG signals on the DEAP database (Soleymani et al., 2012). In Soleymani et al. (2012), they proposed a facial expression recognition algorithm. In particular, they first extracted frames from video sequences. Then, the structures were used to locate the faces, and a feature extractor was employed to extract face features. Finally, the extracted face features were normalized to obtain a higher level feature set, followed by training the SVM classifier to recognize facial expressions in real-time. A modality fusion strategy with an SVM (Soleymani and Pantic, 2012) was used to classify arousal and valence into three categories, respectively. The SVM with RBF kernel was utilized to identify the samples by discriminative features from two modalities. However, the problem with employing an SVM in a fusion scheme is that the output of SVM classifiers is uncalibrated; it is not directly usable, being a confidence value when combining results of different classifiers. Therefore, in Soleymani and Pantic (2012) they used two methods to tackle the problem, i.e., to model the probability of two classes determining the output values of SVM and adopting a solution to the extent of multiple courses.

Although these SVM-based algorithms have made significant progress in affective video recommendation, they are facing the problem of ignoring the temporal video factor and seriously affecting the recommendation quality. To solve this problem, Niu et al. (2017) studied the temporal element of emotion, i.e., the characteristics of emotional fluctuation. They proposed a method based on Grey Relational Analysis (GRA) to solve the above-mentioned problems. First, video features were extracted and mapped to Lovheim emotion space through an SVM. Then, GRA calculated the relationship between videos based on emotional features. Finally, the Fisher model was used for video recommendation, and their method proved effective when recommending temporal video sources.

In Srivastava and Roy (2014), they used an SVR to extract the connotative features of the movie's audio to represent user reaction impressions. The SVR ranked the film according to the connotative features and then compared the ranking results with the user preferences and recommended movies to the users. An affective recommender framework was proposed to provide personalized movie recommendations (Sivakumar et al., 2015) using audio-visual descriptors and connotations to offer the viewers' emotional state. They adopted an SVR to predict the connotative values of each movie at the regression stage, and then the film nearing each other in the created connotative space were recommended to reviewers.

Clustering algorithms

The basic idea of video recommendations using a clustering algorithm is to cluster viewers or videos into groups based on the emotional similarity of viewers or the similarity of video features. The former recommends videos to users with similar emotional states, and the latter recommends unseen videos in the same cluster. In Niu et al. (2013), they presented a video browsing system called Affivir that dynamically adjusted session parameters according to viewers' current mood by modeling user-watching behavior. For a given user, Affivir first analyzed the user's emotional interest through an interactive process where user behavior of watching and skipping was recorded. When the user's preference was learned, the unseen videos with similar affective responses based on affective similarities were recommended. Four affective video features generated identical videos. To improve the efficiency of video retrieval, videos in the database were pre-clustered based on video similarities. Subsequently, Niu et al. (2016) proposed an improved similarity calculation method, normalized validity—approximate graphs (NVAG), and adopted the block-based color histogram for similarity measurement. NVAG significantly improved the recommendation effect in video sharing compared with the Affivir algorithm.

AdaBoost learning algorithms

The core idea of adopting AdaBoost learning algorithms is selecting discriminative features to construct a facial expression classifier. Unlike the original AdaBoost algorithms selecting the best features in several rounds and generating a weak classifier, the AdaBoost algorithms used in facial expression tend to develop a mid-strong classifier based on a compositional feature. In Shu et al. (2018a), an AdaBoost classifier was used based on ECG signals obtained by a wearable device to analyze the emotional state, whether positive or negative. In Zhao et al. (2013), they proposed an improved AdaBoost learning algorithm to classify and recommend videos. The proposed method was based on facial expression recognition fused with spatiotemporal features. The spatial features combined Haar-like elements with training a mid-classifier and then were embedded into the improved AdaBoost learning algorithm to achieve spatial characteristics. For the temporal feature combination process, a time dimension variable was employed by the hidden dynamic conditional random fields (HDCRFs), and then the spatial features were embedded into HDCRFs to recognize facial expressions. The affective curve reflected the process of emotional changes. The video affection was classified into affective sections by psychology-based rules and probability-based scores by segmenting different emotional states. Finally, the videos were recommended to the users according to their affection states. **Figure 3** illustrates the framework of the improved AdaBoost learning algorithm.

Matrix-based algorithms (MA)

The main idea of the matrix-based algorithms is to compile the multidimensional attributes in the data into a neural decision matrix (including the user's emotional state, physiological parameters, etc.) and then conduct multiple standard neural analyses based on the neural decision matrix. To solve the recommendation problem of real estate advertising, video (Kaklauskas et al., 2018) considered the emotional state of buyers and proposed a neuro decision matrix based on house attributes, the emotional conditions of buyers, and physiological parameters. They selected the most personalized video alternatives according to the performance of a multiple criteria neuro analysis. They designed the neuro advertising property video recommendation system to provide effective video advertising for real estate buyers for a long time. In Dnodxvndv et al. (2018), they proposed a video neuro-advertising recommender model to analyze consumers' emotions, measure the engagement of relevant ads, and make advertisements more efficient. The video neuro-advertising recommender model contained two Video Neuro-advertising Models and Systems (VINERS) Sub-models. The first Sub-model was based on the compiled neuro-matrix for assessing the effectiveness of a recommended advertisement; another Sub-model was used to generate a large number of variants for every viewer of an already developed advertisement.

Collaborative filtering (CF)

The general idea of AVRS research based on CF is mainly realized by measuring similarity, either recommending videos with similar emotions according to users' emotional states or adding affective analysis factors when measuring users' similarity. The collaborative filtering-based algorithm was one of recommender families' most extensively used methods. In Soleymani et al. (2009), they proposed a collaborative, personalized affective video retrieval, which can retrieve videos according to emotional queries, arousal, and valence. Based on the traditional CF algorithm, Winoto and Tang (2010) considered the emotional factors and analyzed the impact between the user's mood and the ratings of different movies. For example, whether a user with positive mood scores higher on romantic comedies or whether the user will score higher on action movies when he is in a tense mood.

Traditional CF algorithms recommend users based on their historical behavior similarity. However, new users face the cold-start problem. Instead of using historical behavior records (Choi et al., 2016), the changes in users' facial scales were used to describe the dynamic preferences of usage. Through this method, they provided accurate, personalized recommendations for new and existing users, thus solving the users' cold-start problem. In Diaz et al. (2018), they designed a recommender entirely based on the impression data of viewers. When a user views a video, the recommender system retrieves the metric data from user information. The video impression metric was used to determine which video resembled the metric of the current video. They tested three categories, i.e., the joy impression, the fear impression, and the sad impression. This impression-based recommender system was proved to break the lack of feature-based recommender systems.

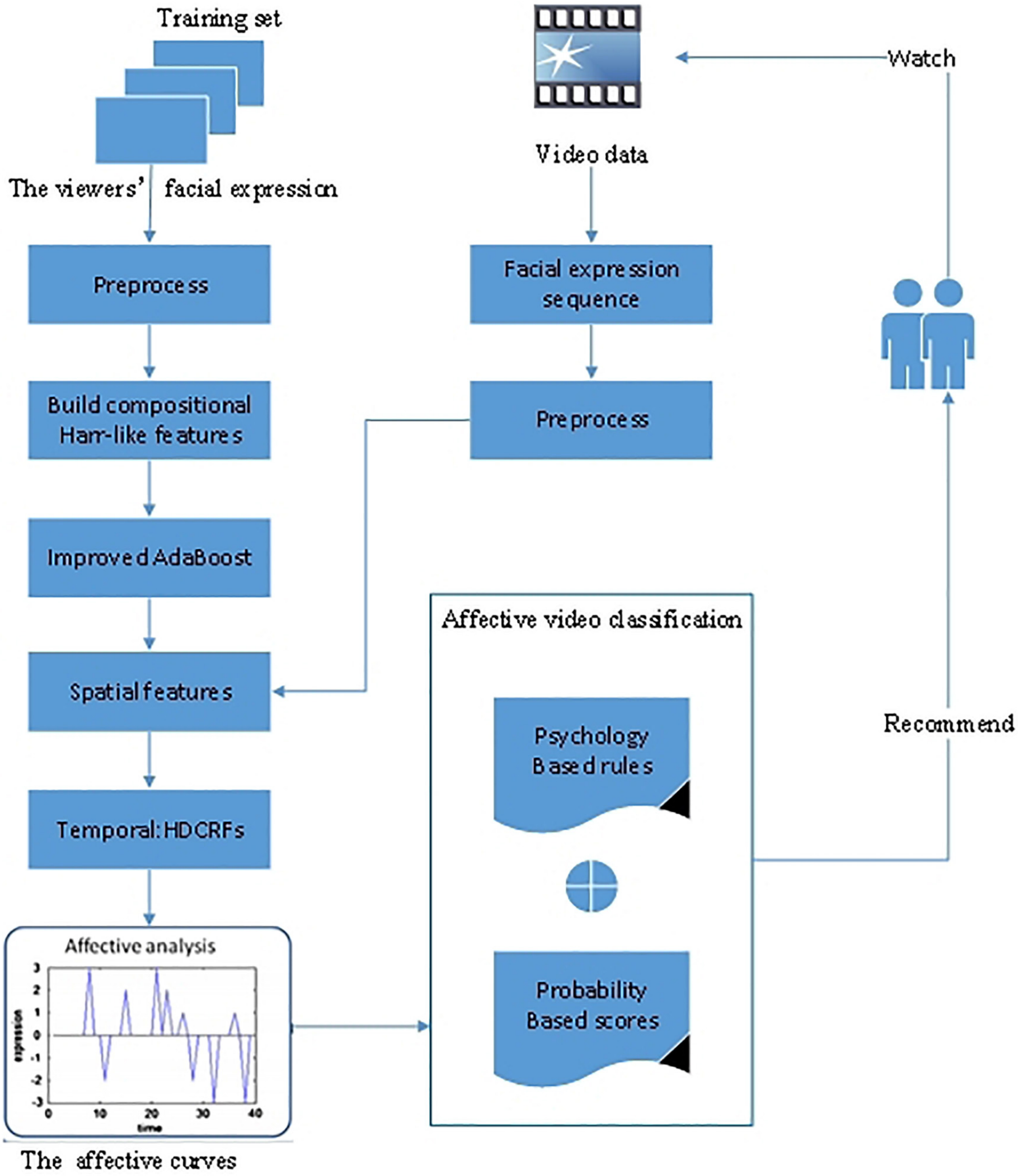


FIGURE 3
The framework of the improved AdaBoost learning algorithm (Zhao et al., 2013).

Content-based filtering (CBF)

The dominant thought behind the CBF of AVRS is to incorporate affective video metadata, explicit feedback information, and user mood as part of an item or user attributes. In Canini et al. (2013), they believed that emotional content recommendations could better meet users' tastes and preferences, so they extracted video tags and audio-visual features to combine semantic and affective video information. This method solved the problem of insufficient individual user preference space characteristics by processing user logs and boosting strategies. In Tkalčič et al. (2013b), a new database named LDOS-PerAff-1 Corpus was collected. To confirm the value of the new database consisting of emotion tags and the users' ratings, they used four recommendation algorithms for verification: a fusion content-based algorithm, a collaborative filtering algorithm, an emotion detection algorithm, and matrix factorization. These four algorithms were tested involving different values of the used corpus in the recommendation, including the effectiveness of affected data improving the content-based advice, personality information that improves the cold-start problem, the role of emotion detection methods in face recognition, and user preferences for items with different favorite attributes.

Video data on the Internet does not exist alone but co-exist. For example, multimedia resources can contain video, audio, images, and other forms of existence at the same time (Soleymani et al., 2015). The affective analysis of multimedia content focuses on estimating users' expected emotional state. In Deldjoo et al. (2018), they developed a content-based multimedia recommendation system (CB-MMRS) model based on CBF according to distinct resources. For video recommendation, items came from videos, movies, movie clips, trailers, etc. Items were used to match the user's emotional state and obtain clear feedback by stimulating the user's emotional state or by analyzing multimedia data.

Knowledge graph (KG)

The central idea behind KG of AVRS is to look for a particular emotion by KG, which has a similar emotion state extracted from user movie reviews. In Qi et al. (2021), they aimed to choose a small set of video frames based on the viewers' personalized interest for video highlight detection. Specifically, they extracted the concept representation video clips by a front-end network, the concepts were used to build an emotion-related KG, and the relationships in the graph were related to the external public KGs. The emotional state influences decision-making when users consume movies. Therefore, a knowledge graph-based method (Breitfuss et al., 2021) was proposed to include the emotional state factor in movie recommendations. They extracted emotions from pre-existing movie reviews to construct the knowledge graph. To test the efficiency of the proposed method, they developed a chatbot with a reasoning mechanism combing users' emotions analyzed from chat messages. **Figure 4** shows the recommendation process based on KG. The chat messages of movie reviews between an AI chatbot and a user was extracted and categorized by a Bayesian classifier based on emotions. Natural language processing technology was used to remove emotions. To promote the speed of data retrieval, a graph database named graph DB API was employed to store the processing emotions.

Genetic algorithms (GA)

The GA is often used to solve the optimization problem of multiple objectives with conflicts. In AVRS, the critical idea of GAs is to balance the imbalance between users' emotional preferences and actual business objectives. In Yadati et al. (2014), they studied the application of emotion analysis in in-stream video advertising as one of few excellent video recommendation works based on affective analysis and considering multiple objectives. They explained that emotion played a vital role in users' purchasing behavior, and the consideration of emotional influence should be added to video advertising. Therefore, they proposed a method of Computational Affective Video-in-Video Advertising (CAVVA) strategy, which mainly considered two factors: identifying candidate advertising insertion points and the most appropriate advertisement. They modeled the problem as non-linear integer programming. Due to the conflict between these two objectives, minimizing the impact of advertising insertion on users and maximizing users' participation in advertising, they adopted a genetic algorithm to solve the above conflict problems.

Hybrid recommender systems (HRS)

The dominant thought of employing HRS in AVRS is that combining multiple algorithms involving the viewers' emotional states can promote recommendation efficiency. The effect of video recommendation by a single algorithm is limited, so researchers turned to HRS. In Wakil et al. (2015), they provided a hybrid model combining CF, CBF, and emotion detection algorithms. The CF and CBF algorithm was used to capture users' preferences, and the emotion detection algorithm considered the influence of users' emotion, which the traditional recommendation algorithms did not consider. An exciting research direction on video recommendation is temporary saliency, i.e., detecting the most critical video events, which may be the most attractive parts for users. A time series of arousal model (Mugellini et al., 2014) was designed based on audio-visual features to analyze users' emotions. The multimodal system helps extract the parts that users may be interested in and can combine with various recommendation algorithms.

To summarize, in the last few years, researchers have made great efforts to video retrieve and recommendation domains by various traditional recommendation algorithms, including SVM/SVR, clustering, AdaBoost, MA, CF, CBF, KG, GA, and HRS. Some of their research works have achieved remarkable success, promoted the progress of AVRS, and improved the efficiency and quality of viewers' access to video information. However, these algorithms still face the following problems:

1) Although the algorithm is simple and easy to implement, it cannot make accurate judgments on complex scenarios, and the recommendation effect is minimal. For example, Niu et al. (2013) recommended videos by clustering viewers' moods, which was not a personalized recommendation strategy, and thus the recommendations may not work well.
2) The experiment databases are relatively small and not diverse. The portability of the recommendation strategy generated based on such a database is low, significant-good results on one database, while probably inferior on other databases. For example, Zhao et al. (2013) relied heavily on exaggerated and unnatural facial emotion expressions and lacked direct and intuitive expression, making the recommendation model unsuitable for the actual situation.

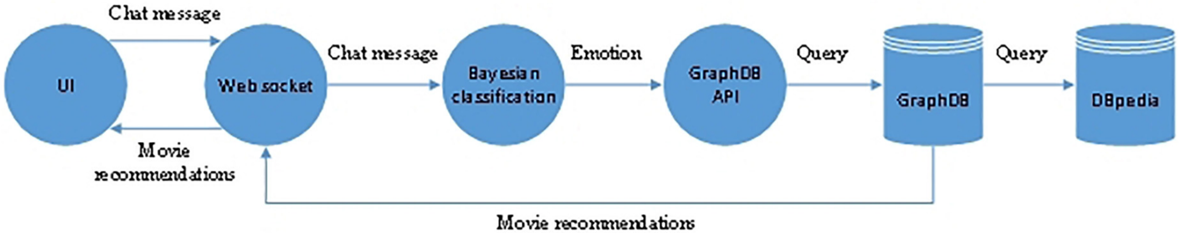


FIGURE 4
The recommendation process based on KG (Breitfuss et al., 2021).

Deep learning-based methods

Traditional recommendation algorithms, such as matrix factorization algorithms, are linear models, and the recommended effect is limited. Compared with conventional linear recommendation models, deep learning (DL) (Zhang et al., 2019) can obtain the non-linear characteristics of user interaction data, thereby capturing more complex information about user interaction patterns (Dai, 2021). The sequential modeling of DL also shows promising aspects in processing speech recognition, text analysis, etc. Therefore, the recommendation effectiveness of deep learning in recommender systems has been superior. Deep learning has penetrated a series of fields; the publication of deep learning algorithms has grown exponentially in industry and academia. Although DL has proved its essential role in the recommendation system, the exploration of the recommendation system in video recommendation is still limited, which needs to be paid attention by more scholars and supported by works in more fields. This subsection introduces several state-of-the-art DL models for solving affective video recommendations.

Reinforcement learning (RL)

The core idea of adopting RL in AVRS is that DL can continuously and dynamically learn strategies through the real-time state changes caused by the impact of users on the surrounding environment to maximize the cumulative reward. In Leite et al. (2022), they discussed the role of deep reinforcement learning (DRL) in video recommendation when used in a virtual learning environment. They also considered two different student groups, i.e., common effect and high effect. They designed a recommender system including five categories, i.e., the new videos to watch, the students communicating the current topic with a new tutor, the students displaying the segment with the current tutor, the corresponding piece with a new tutor, and the following video to watch. The type of recommender system was determined by the scores of students' tests and the sensor-free participation detection model. The recommended strategy was based on a DRL algorithm. It was evaluated by a large field experiment, which showed the effectiveness of video recommendations during the regular school period. In Tripathi et al. (2018), they believe that the cognitive preferences of viewers are dynamic and should track the behavior of viewers and their cognitive preferences for different emotions in real-time. Therefore, they proposed an RL method to learn video recommendation decisions and monitor the interaction between users and recommended videos in real-time through the created user interface and webcam. **Figure 5** illustrates the RL sequence of states and actions. The S_t, a_t, and r_t demonstrate the state, action, and the reward of time t, whereas r_{t+1} represents the reward gained by performing action in the state of s_t. The learning process continued until state s_{t+n}.

Convolutional neural network (CNN)

The basic idea of CNN in affective video analysis is that the CNNs can be employed for feature extraction from various types of signals and information. In Hewitt and Gunes (2018), they deployed a CNN model for facial affective analysis used on mobile devices. The proposed CNN model incorporates three variants of CNN architectures (i.e., AlexNet Variant, VGGNet Variant, and MobileNet Variant), which consider both the high performance and the low storage requirements. In Kwon et al. (2018), they designed a CNN architecture for accurate emotional classification. The CNN model extracts both temporal and frequency characteristic features from electroencephalogram signals and the pre-processed galvanic skin response (GSR) signals. The electroencephalogram signals reflect temporal characteristics as human emotions are time sequence data. A wavelet transform represents the frequency feature through the frequency axis. In Yang et al. (2019), they presented a multi-column CNN model using EEG signals for emotion recognition. The decision of the proposed CNN model is generated by a weighted sum of multiple individual recognizing modules.

Unlike the above method of detecting the viewer's emotion change through the device, Zhu et al. (2019) automatically recognized the viewer's emotion by acquiring the information about the protagonist. They used a protagonist-based key frame selection strategy to extract features from video clips to alleviate the considerable workload of analyzing a large amount of video information. Then, the characteristics of keywords

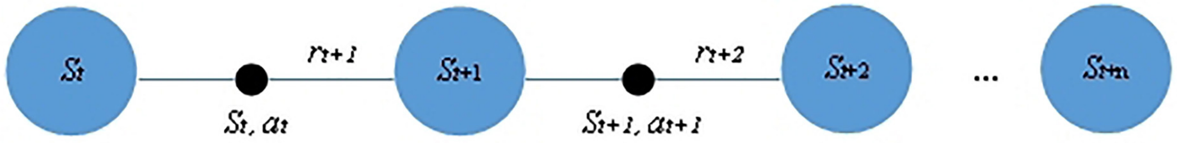


FIGURE 5
The RL sequence of states and actions (Tripathi et al., 2018).

were fed into a CNN model based on optical flow images, and the CNN model incorporated temporal information from video clips. Then all of the features were fused as inputs of an SVM and SVR model for affective video recognition. The framework of the proposed method (Zhu et al., 2019) is shown in **Figure 6**. The framework is composed of two parts: feature extraction and feature concatenation. In the first process, they employed two CNN models to extract features related to hand-crafted visual and audio elements. The protagonists' keyframes (PKFs) were selected from video clips. Then, two parallel extraction strategies were adopted to collect the matrix and optical flow images through two CNN models. These features were finally concatenated to map the affective dimension by an SVM/SVR model.

Long short-term memory (LSTM)

The dominant thought of adopting LSTM models in emotional video classification is that LSTMs can consider temporal, spatial, and frequency characteristics of various signals and information. In Alhagry (2017), an LSTM is adopted to learn the EEG features for emotional video recognition. The LSTM model takes the dense layer to classify the raw EEG features into low and high arousal, valence, and predicting the continuous scale between 1 and 9. In Wang et al. (2020), they established a Bi-LSTM model to extract emotional features for analyzing danmaku video data and users' affective characteristics. The Bi-LSTM model classifies the users' emotions into four dimensions, i.e., pleasure, anger, sorrow, and joy. In Zhang and Zhang (2017), they studied the inherent correlations between video content and the viewers' affective states by presenting an LSTM model, which simultaneously predicts the arousal and valance dimensions. The LSTM model extracts a collection of low-level multimodal features from videos and projects these features into arousal and valence value pairs. In Nie et al. (2020), they considered the relations between the utterances and handled the multimodal feature fusion problem in the feature learning process with an LSTM-based model. In Ogawa et al. (2018), they introduced a Bi-LSTM network, which collaboratively adopts video features and EEG signals. They first used transfer learning for video classification as the limited number of video labels which difficult to classify. Then, a user study was conducted to verify the effective representation of EEG signals calculated by Bi-LSTM.

In Cao et al. (2022), they proposed the Visual Enhanced Comments Emotion Recognition Model (VECERM) to analyze users' emotions, thereby overcoming the problem of user-generated comments related to plots. The VECERM model was composed of four layers.

Input embedding layer

In the input embedding layer, two significant parts are included: users' text data comments and the images of video frames. This layer reduces the dimension of the input information, VGG processes the video information, and the Transformer processes the text information. The Transformer then converts the text representation into embedding vectors.

Context enhancement layer

Since text information and comments are synchronized, the Context Enhancement Layer mixes video information and text data through the attention mechanism.

Emotion attention layer

The purpose of the Emotion Attention Layer is to mine the emotional semantics of the comment text to obtain a good text representation. Due to the short length of the text, Bi-directional Long Short-Term Memory (BiLSTM) is adopted for mining the text data.

Classification layer

The Classification Layer realizes the classification of users' emotions throughout the whole connection layer. This is a multi-classification classification problem, including glad, dismissed, sad, amazed, and afraid.

The VECERM architecture is shown in **Figure 7**.

Multilayer perception (MLP)

The central idea behind the MLP of AVRS is to extract features from multimodal data to classify emotional expressions, e.g., visual, audio, and textual information. Krishnamurthy (2020) utilized an MLP network to classify user sentiments. The MLP model analyzes the users' emotions based on web recordings from multimodal resources. They employed a feature-level fusion method to fuse the extracted features from various modalities, i.e., video, posts, and pictures. An oppositional grass bee algorithm then chooses the extracted features to generate the best optimal feature set.

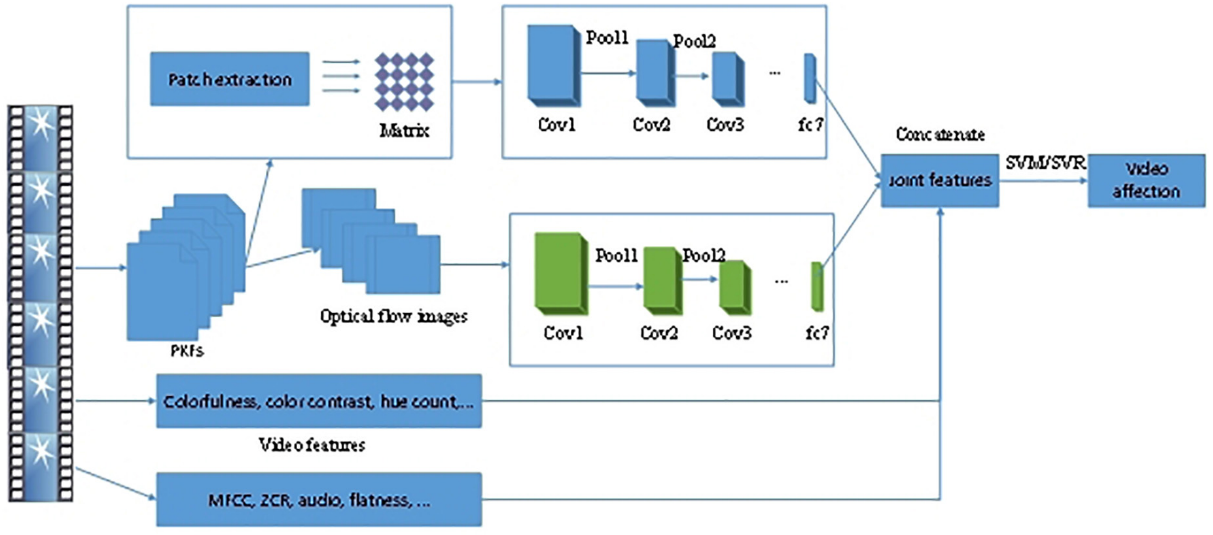


FIGURE 6
The framework of the proposed method (Zhu et al., 2019).

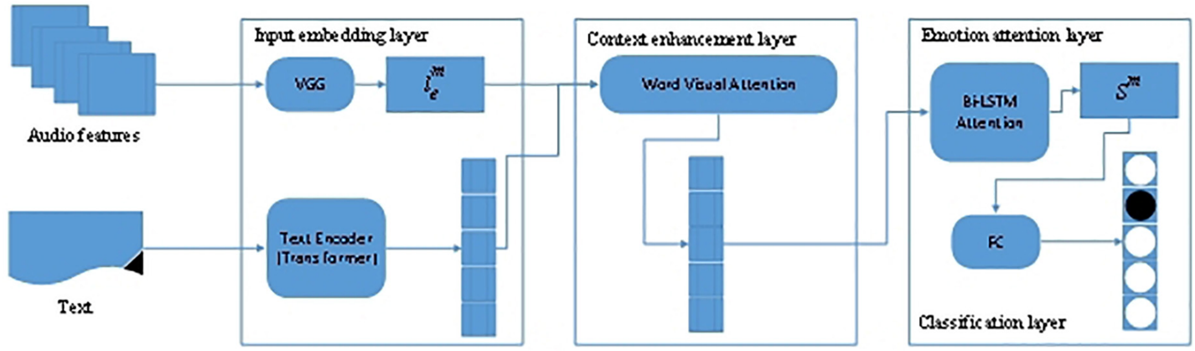


FIGURE 7
The architecture of VECERM (Cao et al., 2022).

In Boughrara et al. (2016), they proposed an MLP for facial expression classification. The established MLP model consists of a single hidden layer, which seeks to find synthesis parameters in the training stage. They adopted a biological vision-based facial description in the feature extraction step to extract face image features.

To predict the emotional state of users when watching a stereoscopic 3D video, Đorđević Čegar et al. (2020) extracted features from the volunteers' psychological data of ECG, EDA, and EEG signals and then used an emotional state estimator based on feedforward multilayer perception artificial neural network to predict the state of viewers when they were viewing different kinds of stereoscopic 3D video content. The MLP model is shown in **Figure 8**. The configuration of MLP based on HR and EDA selected features were as the input features, including IIR Median, HR Moving STD, HR Moving PCA, EDA Median, EDA STD, EDA PCA, and SCR Mean. They adopted the Levenberg-Marquardt back-propagation algorithm for training the network. The output of MLP was a linear activation function, which generated the estimated scores.

Deep hybrid models (DHM)

The fundamental idea of implementing DHM is combining different DL models (e.g., CNN, RNN, LSTM, RL, etc.). The fusing mode of multiple DL models can be either the output of one or several models is used as the input of another model, or several models simultaneously extract the features of video or multimodal data or signals. The combination of several models improves the limited non-linear performance of a single model (e.g., LSTM has a memory for long-time data processing). In Zhang et al. (2018), they established an audio-visual emotion recognition model, which is fused with a CNN, 3D-CNN, and

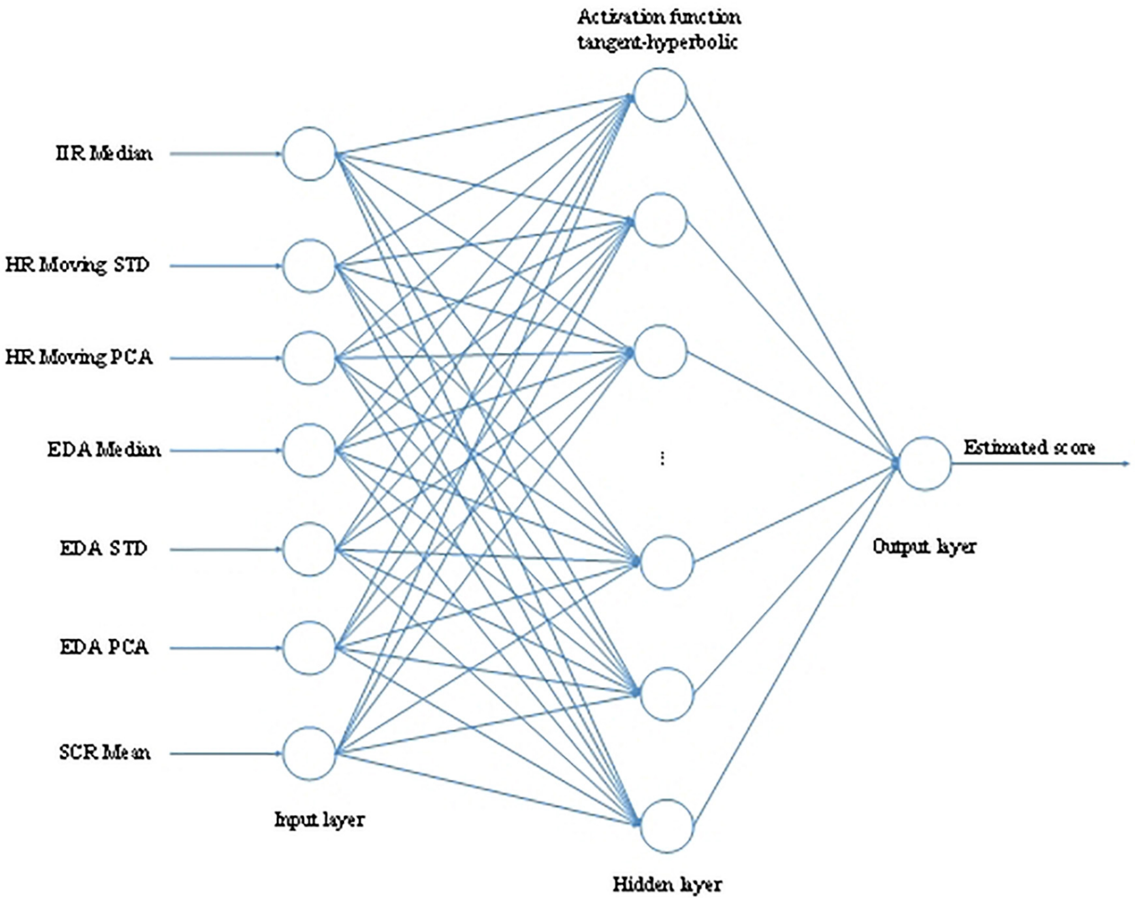


FIGURE 8
The MLP model (Đorđević Čegar et al., 2020).

a Deep Believe Networks (DBNs). The designed model is a two-step procedure. The CNN and 3D-CNN are firstly pre-trained according to a large-scale of both image and video tasks, which are fine-tuned to learn audio and visual segment features. Then, the output of the former step is combined into a fusion network to build a DBN model, and a linear SVM obtains the final results of emotional classification. In Fan et al. (2016), they proposed a hybrid DL model for video-based emotional recognition. The model is the combination of a recurrent neural network (RNN) and a 3D CNN. The 3D CNN models the video appearance and motion concurrently, while the RNN model processes the appearance features obtained by the CNN model over individual video frames, which are used for the input features, then RNN encodes the motion. In Yenter (2017), they produced an architecture that combined CNN and LSTM models for textual sentiment analysis. The CNN model is consisted of multiple branches, whereas the LSTM model is a word-level classification. The output of CNN branches is transferred to the LSTM and then concatenated to a fully-connected layer to generate a single output for sentiment polarity classification.

Mishra et al. (2020) established a fascinating empirical analysis. Firstly, they used two CNN models (AlexNet and GoogLeNet) and an LSTM model to classify EEG data into different emotion categories. The purpose was to recognize the emotional state of EEG data through the deep learning model. Using the pre-trained CNN and LSTM models can reduce the computing cost of the training network through simple parameter adjustment. Then, these models were used to verify whether the trained models were universal and effective in different fields. In Liu et al. (2017), they presented two attention mechanisms, i.e., LSTM and RNN, for emotion recognition. These two models integrate temporal attention and band attention, which are based on untrimmed visual signals and EEG signals. The LSTM and RNN models take all the signal data as inputs and then generate representations of each signal, which are transferred to a multimodal fusion unit for predicting the emotional labels. Tripathi et al. (2019) designed a personalized and emotional intelligence video recommendation engine named EmoWare, which employed reinforcement learning (RL) and deep-bidirectional recurrent neural networks

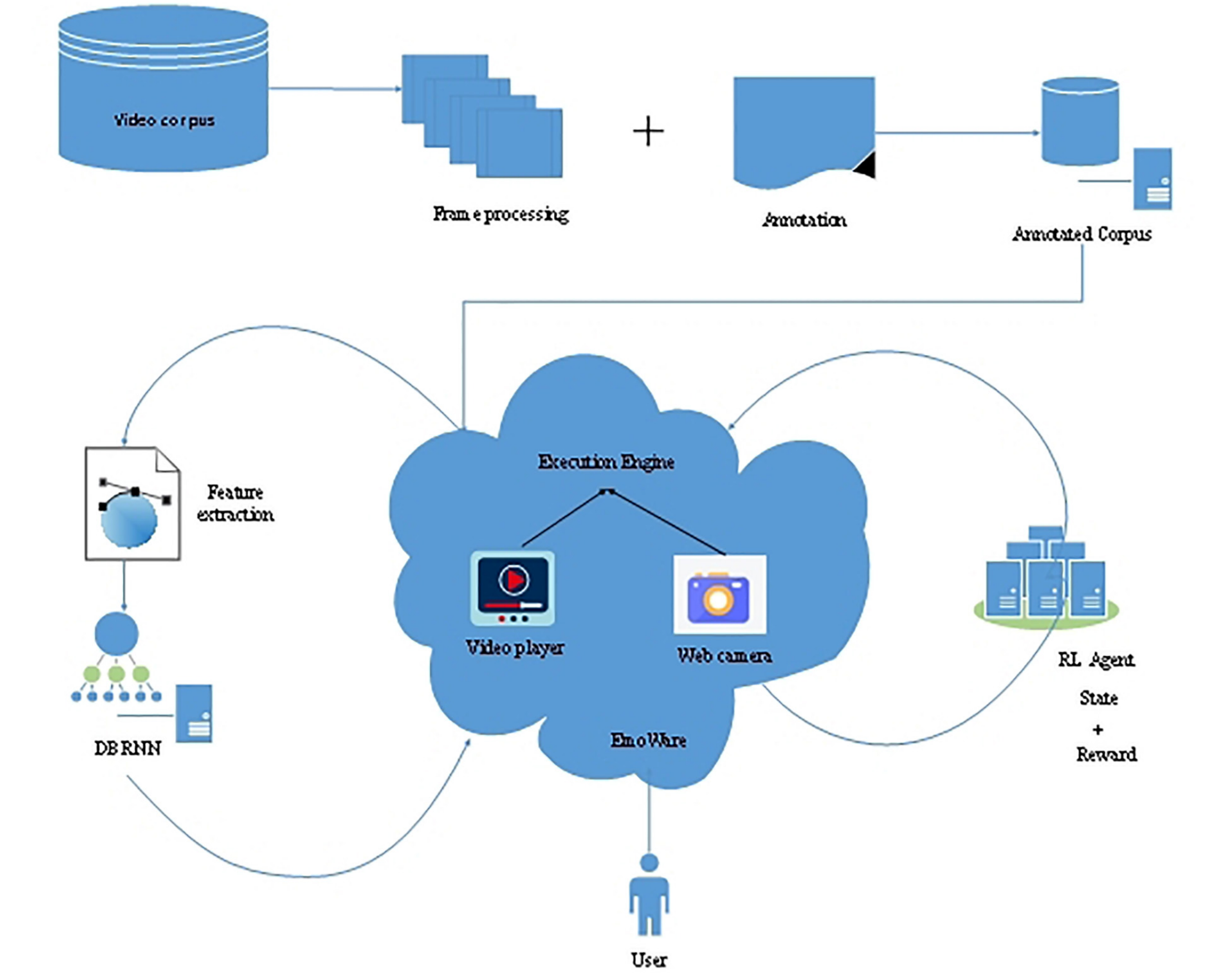


FIGURE 9
The framework of EmoWare (Tripathi et al., 2019).

(DBRNN) models. The framework of EmoWare is shown in Figure 9.

To summarize, deep learning-based AVRS algorithms can learn the potential characteristics of audio, text, video, and other multimedia and obtain representations and abstraction from multiple levels, resulting in its significant advantages in dealing with emotional analysis. For example, the CNN can capture the global and local features and analyze spatial information changes during short time periods of video clips, remarkably enhancing efficiency (Fonnegra, 2018). The RNN architecture is good at processing sequential data by remembering former computations in loops. Each deep learning algorithm has its personalized advantages and disadvantages. Therefore, researchers combine several deep learning models to solve complex problems. Especially, Tripathi et al. (2019) adopted RNN and LSTM algorithms concurrently. However, deep learning is still in its infancy in affective video recommendation. The research work of exploration is scarce, and the available databases are also very precious. It still needs a large amount of research support.

Affective video recommendation databases

In this section, we introduce the existing 31 valuable databases which play a vital role in AVRS research. These databases are composed of multiple modes, including comments, ratings, videos, films, audio, images, etc. There are various methods to obtain these data, such as capturing the changes in the viewer's facial expressions through webcams, getting the user's physiological signals through EEG, questionnaires, or a combination of these methods. Most of these databases are manually collected by researchers, which is time-consuming and error-prone. In Lucey (2012), they provided an effective way to construct two databases without manually scanning the full movies, and the movie labelers only reviewed video clips recommended by an RS. These video clips are the most representative. This method can quickly collect and obtain much-annotated video information. The various databases and their details are listed in **Table 4**.

The audience responses

The audience response to a video can be obtained in various ways, mainly including two categories: explicit acquisition and implicit acquisition. Standard methods for explicit acquisition include user interactions (i.e., watching videos, skipping videos.), questionnaires, surveys, and quizzes. The questionnaires can be achieved through self-assessment manikin (SAM) (Dnodxvndv et al., 2018). There is a wide range of ways to implicitly obtain the emotional characteristics of viewers, including facial expressions or features, measuring skin estimated pulse, heart rate, body gestures, reviews, or comments. The viewers' psychophysiological signals of heart rate (HR) are calculated from an echocardiogram (ECG) (Baveye et al., 2015), while electro-dermal activity (EDA) and brain activity (BA) are from electroencephalography (EEG) signals (Đorđević Čegar et al., 2020). The facial expressions or features [e.g., gaze distance (Soleymani and Pantic, 2012)] can be obtained by a camera (Tkalčič et al., 2013a). The questionnaire can accurately convey the emotional state of users. However, it is also faced with the problem that the amount of data is limited, affecting the viewing experience, costly for organizations to conduct, and volunteers sacrifice much time (Mulholland et al., 2017). Therefore, an implicit acquisition that obtains affective states from face recognition, heart rate, mood, EDA, BA, and body gestures plays a significant role and provides more ways for affective video recommendation. The method of implicit acquisition is more flexible. Only by recording the physical signs of the viewer can we obtain the emotional state through the algorithm. Martha and Larson (2013) provide a unique perspective to analyze the emotional states, that is, perceived connotative properties, which prove to be more intersubjectively shared.

Table 5 shows the audience responses in different publications. It can be inferred that facial expressions/features, skin-estimated pulse/heart rate, movie reviews/comments, and questionnaire/survey/quizzes are the most frequently used user responses in affective video computing. Some researchers also get users' emotional feedback on videos from other different perspectives, such as mood (Winoto and Tang, 2010), EDA (Đorđević Čegar et al., 2020), BA (Đorđević Čegar et al., 2020), body gestures (Hassib et al., 2017), and perceived connotative properties (Martha and Larson, 2013). Some experimental studies use one of these methods to obtain emotional expression, but most of the research work uses a combination of multiple user feedback methods. For example, Bohlin et al. (2019) and Soni et al. (2019) evaluate the emotional state by facial expressions/features and skin-estimated pulse/heart rate (Diaz et al., 2018) adopt the method of combination of skin-estimated pulse/heart rate and questionnaire.

Evaluation methods

The commonly used performance indicators include mean accuracy, precision/recall/F1, mean absolute error (MAE), mean square error (MSE)/root mean square error (RMSE), confusion matrix, and valence, arousal, and dominance. However, viewers do not need perfect prediction accuracy but need wise recommendation strategies. Therefore, in addition to the former metrics, several researchers also began to pay attention to the quality of perceived recommendations to evaluate their models and algorithms. For example, Arapakis et al. (2009a) adopted Pearson's ChiSquare test and the Dependent *t*-test to analyze the emotion variance and the recommender system's performance. Niu et al. (2013, 2016) used CTR, session length, and points test to evaluate the recommendation performance. The higher the CTR, the longer the session length, and the better the recommendation quality. The compiler average causal effect (CACE) evaluator was employed by Leite et al. (2022) to test the impact of recommendations offered to the treatment group. Breitfuss et al. (2021) tested their knowledge graph-based recommendation strategy by various metrics, including Sparsity impact, the granularity of emotions, extensibility, recommendation quality, and additional characteristics. **Table 6** lists the evaluation metrics used in different publications.

Challenges and opportunities

In this survey, an overview of traditional recommendation methods (e.g., SVM, SVR, CF, CBF, AdaBoost, GA, Clustering, MA, KG, HRS) and deep learning-based technologies (e.g., CNN, MLP, RL, RNN, LSTM, DHM) adopted in AVRS has been depicted. The research of AVRS is challenging since a tremendous effort involving a multidisciplinary understanding of human behavior and perception and multimodal approaches integrating different modalities are required, such as text, audio, image, and video. Although many scholars have begun to pay attention to the field of AVRS in recent years and have made valuable contributions from the perspective of data, models, and algorithms, AVRS is still in its infancy. The challenges of the AVRS domain mainly come from the following three aspects:

(1) Insufficient data and data analysis is highly sophisticated.

Much of the existing facial data exists a lot of unnatural and exaggerated expressions (Zhang et al., 2022). More intuitive, natural, scalable, and transportable facial expressions are needed. In addition, the research on emotion analysis in the field

TABLE 4 The databases for affective computing.

Name	Details	Publication
The affective feedback database	Questionnaires of 24 participants on tasks, search process, and emotional experience of the information-seeking process	Arapakis et al., 2009a
Cohn–Kanada expression database	The database has 2105 digitized image sequences of 182 adult subjects, suitable for comparative studies by multiple tokens of most primary FACS action units.	Zhao et al., 2011
Moviepilot mood track	It consists of 4.5M ratings assigned by 105K users on 25K movies. Various contextual information is provided, i.e., gender, age, production year, the audience of each movie, movie-mood tag, etc.	Shi et al., 2013
The Hollywood movie video clips database	Contains 155 video clips from Hollywood movies, annotated by 40 participants with more than 1,300 annotations.	Soleymani et al., 2009
The Tellyads and YouTube video clips database	Contains 15 videos of 165 min duration from various genres, e.g., TV shows, movie clips, and news broadcasts.	Yadati et al., 2014
The affective property movie database	The database contains more than 2,000 videos; movie affective properties are measured by arousal and valence.	Niu et al., 2013
Nvidia 3D Vision database	The database contains nine stereoscopic sequences of nearly 2 min duration.	Đorđević Čegar et al., 2020
The movie profile database	It contains an item profile of various attributes describing the movie content.	Wakil et al., 2015
The five emotional reactions database	Two standard webcams are operating in real-time used to capture the users' facial expressions and estimate the pulse. The users' reactions can be classified into five categories: happiness, sadness, anger, fear, and surprise.	Bohlin et al., 2019; Soni et al., 2019
Cohn–Kanada database	Consists of 100 students of different races, i.e., African–American, Asian, and Latino. Each subject performs a series of 23 facial displays. The selected sequences are labeled with six emotions: anger, disgust, fear, happiness, sadness, and surprise.	Zhao et al., 2013
The clicker and emotional reaction database	It consists of 30 subjects from the age of 18–35. Each subject watches five videos, and two webcams monitor the behavior. The issues must also be surveyed according to their watching and rating.	Diaz et al., 2018
DEAP	The database is a multimodal database using EEG and physiological signals for emotion analysis. The database obtains 32 subjects' 1-min musical physiological video signals.	Soleymani et al., 2012; Dabas et al., 2018; Mishra et al., 2020
Algebra video field test database	The data are collected by a field experiment of 18,925 school students and 152 teachers in 149 schools.	Leite et al., 2022
Cohn Kanade database	It contains photos of different emotions, from a neutral state to an explicit one.	Leite et al., 2022
The 0-MOOD, 7-MOOD, 16-MOOD	It contains 0, 7, and 16 mood states, respectively.	Winoto and Tang, 2010
The user action session database	Affivir constantly crawls video data from the Internet, and user preference features are extracted.	Niu et al., 2013; Niu et al., 2016
The format video database	It contains 1,000 format mp4 videos ranging from 30 s to 10 min. The videos are from various websites, i.e., Youku.com, YouTube.com, etc.	Niu et al., 2017
The footwear advertising videos database	The user facial features and ratings of 52 subjects record the movement of vital facial points continuously.	Choi et al., 2016
The NEAR database	The NEAR database consists of a wide range of databases, i.e., the Property Video Clip Ads Database, a text database of video clips.	Kaklauskas et al., 2018
LIRIS-ACCEDE	It contains 160 feature films and short films from 9,800 video clips. It is the largest video database with emotional labels and can be used for video indexing, summarization, and browsing.	Baveye et al., 2013; Baveye et al., 2015; Zhu et al., 2019
PM-SZU	It is a new database for affective video analysis. It consists of 386 video clips extracted from 8 films.	Zhu et al., 2019
The metractitic.com and imdb.com database	It consists of 2,627,476 movie reviews.	Breitfuss et al., 2021
Danmu database	It contains a large amount of user-generated comments from Bilibili.	Cao et al., 2022
LDOS-PerAff-1 Corpus	It consists of subjects' affective responses to video clips, answers are annotated in the continuous valence-arousal-dominance space, and topics are annotated with personality information.	Tkalčič et al., 2011a, 2013b
Mechanical Turk setup	It contains affective annotations for the corpus to evaluate viewers' reported boredom.	Martha and Larson, 2013; Soleymani et al., 2014
Multidimensional sentiment dictionary from Ren CE	It includes 1,487 blogs and many emotional words and is labeled as a vector of 8 dimensions.	Pan et al., 2020
YouTube video clips	Containing f 600 videos, 480 had transcripts.	Pan et al., 2020
LDOS-CoMoDa	It consists of contextual information and ratings on the users' consumed movies and personality profiles.	Odic et al., 2014
The IMDB movie scenes	Some 240 users are viewing videos on 25 movie scenes on IMDB. The duration is recorded.	Benini et al., 2011
The AFEW database	A dynamic, temporal facial-expression data corpus contains short video clips of facial expressions close to the real world.	Lucey, 2012
The SFEW database	It is a static, harsh conditions database consisting of seven facial expression classes.	Lucey, 2012

TABLE 5 The audience responses in different publications.

Audience responses	Publications
Facial expressions/features	Soleymani and Pantic, 2012; Zhao et al., 2013; Boughrara et al., 2016; Choi et al., 2016; Kaklauskas et al., 2016; Mahata et al., 2017; Diaz et al., 2018; Fonnegra, 2018; Hewitt and Gunes, 2018; Kaklauskas et al., 2018; Bohlin et al., 2019; Soni et al., 2019; De Pessemier et al., 2020; Mishra et al., 2020; Leite et al., 2022
Skin-estimated pulse/heart rate	Dabas et al., 2018; Diaz et al., 2018; Shu et al., 2018a; Bohlin et al., 2019; Soni et al., 2019; Đorđević Čegar et al., 2020
Mood	Winoto and Tang, 2010
EDA	Đorđević Čegar et al., 2020
BA	Alhagry, 2017; Liu et al., 2017; Kwon et al., 2018; Ogawa et al., 2018; Yang et al., 2019; Đorđević Čegar et al., 2020
User interactions	Niu et al., 2013; Niu et al., 2016
GSR	Kwon et al., 2018
Body gestures	Hassib et al., 2017
Perceived connotative properties	Martha and Larson, 2013; Zhang and Zhang, 2017
Movie reviews/comments/web recordings	Mulholland et al., 2017; Yenter, 2017; Tripathi et al., 2019; Krishnamurthy, 2020; Pan et al., 2020; Wang et al., 2020; Breitfuss et al., 2021; Cao et al., 2022
Questionnaire/survey/quiz	Arapakis et al., 2009a; Soleymani and Pantic, 2012; Tkalčič et al., 2013b, 2014; Polignano, 2015; Hassib et al., 2017; Diaz et al., 2018; Dnodxvndv et al., 2018; Kaklauskas et al., 2018; Bohlin et al., 2019; Zhu et al., 2019; Mishra et al., 2020; Kaklauskas et al., 2020; Leite et al., 2022

TABLE 6 The evaluation metrics of different publications.

Metrics	Related research papers
Pearson’s chi-square test and the dependent *t*-test	Arapakis et al., 2009a
Mean accuracy	Zhao et al., 2011, 2013; Tkalčič et al., 2013b; Fan et al., 2016; Alhagry, 2017; Liu et al., 2017; Yenter, 2017; Zhang and Zhang, 2017; Dabas et al., 2018; Fonnegra, 2018; Hewitt and Gunes, 2018; Kwon et al., 2018; Shu et al., 2018a; Zhang et al., 2018; Bohlin et al., 2019; Soni et al., 2019; Yang et al., 2019; De Pessemier et al., 2020; Krishnamurthy, 2020; Mishra et al., 2020; Nie et al., 2020; Wang et al., 2020; Qi et al., 2021; Leite et al., 2022
Precision/recall/F1	Niu et al., 2013; Shi et al., 2013; Liu et al., 2017; Ogawa et al., 2018; Zhang et al., 2018; Tripathi et al., 2019; Yang et al., 2019; Krishnamurthy, 2020; Mishra et al., 2020; Wang et al., 2020; Cao et al., 2022
MAE	Winoto and Tang, 2010; Choi et al., 2016
MSE/RMSE	Boughrara et al., 2016; Hewitt and Gunes, 2018; Tripathi et al., 2019; Zhu et al., 2019; Đorđević Čegar et al., 2020
ROC	Winoto and Tang, 2010
CTR	Niu et al., 2013; Niu et al., 2016
Session length	Niu et al., 2013; Niu et al., 2016
Confusion matrix	Tkalčič et al., 2011b, 2013b; Zhao et al., 2013; Boughrara et al., 2016; Fan et al., 2016
CACE	Leite et al., 2022
Sparsity impact, the granularity of emotions, extensibility, recommendation quality, additional characteristics	Breitfuss et al., 2021
Valence, arousal	Wang and Cheong, 2006; Soleymani et al., 2009; Soleymani and Pantic, 2012; Oliveira et al., 2013; Tkalčič et al., 2013b; Liu et al., 2017; Kwon et al., 2018; Yang et al., 2019

of recommender systems is not comprehensive. More complex expressions that are not easily exposed should be paid attention to, for example, micro-expression recognition (Ben et al., 2021). Additionally, the EEG signals are difficult to analyze from which part of the brain the electrical activity originates (Dabas et al., 2018). This undoubtedly makes it more challenging to accurately diagnose users’ emotional states on video.

(2) Combining existing models and algorithms with deep learning-based techniques is insufficient.

The exploration of affective video recommendation algorithms based on deep learning is currently limited. It only involves several deep models, such as RL, CNN, RNN, LSTM, MLP, and hybrid algorithms of several models. More advanced works and better performance are needed based on emotional analysis recommendations. The state-of-the-art technologies emerging in recent years may also be combined with the AVRS domain, e.g., the self-attention-based transformer model in sentiment changes detection (Wu et al., 2020), and the generative adversarial network (GAN) may provide data augmentation for small-scale video or multimodal databases (Ma et al., 2022).

(3) The research direction is monotonous.

The current focus is limited to the accuracy of prediction on video recommendations, and the main problem to be solved is the cold-start or long-tail effect (Roy and Guntuku, 2016). However, other research directions of recommendation systems are not involved, such as multiobjective recommender systems (MORS) (Wang and Chen, 2021) or multi-task recommender systems (MTRS) (Ma et al., 2018) and explainable recommender systems (ERS) (Zhang and Chen, 2020). The MORS or MTRS can incorporate more objectives or tasks into the video recommendation based on affective computing; these models focus on more extensive aspects of recommendation quality, such as diversity, novelty, etc. The ERS is a promising research direction, which provides the viewers with the recommendation reasoning according to their facial expressions, body gestures, or other kinds of emotional responses.

Author contributions

DW: conceptualization, methodology, formal analysis, resources, data curation, writing—original draft preparation, and visualization. XZ: validation and supervision. Both authors contributed to the writing—review and editing and approved the submitted version.

Funding

This work was supported by the Zhejiang Provincial National Science Foundation of China and the National Science Foundation of China (NSFC) under Grant Nos. LZ20F020002, LQ21F020002, and 61976149.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alhagry, S. (2017). Emotion recognition based on EEG using LSTM recurrent neural network. *Int. J. Adv. Comput. Sci. Appl.* 8, 8–11. doi: 10.14569/IJACSA.2017.081046

Arapakis, I., Moshfeghi, Y., Joho, H., Ren, R., Hannah, D., and Jose, J. M. (2009a). "Enriching user profiling with affective features for the improvement of a multimodal recommender system," in *CIVR 2009 - Proceedings of the ACM International Conference on Image and Video Retrieval*, 216–223. doi: 10.1145/1646396.1646433

Arapakis, I., Moshfeghi, Y., Joho, H., Ren, R., Hannah, D., and Jose, J. M. (2009b). "Integrating facial expressions into user profiling for the improvement of a multimodal recommender system," in *2009 IEEE International Conference on Multimedia and Expo*, 1440–1443. doi: 10.1109/ICME.2009.5202773

Baveye, Y., Chamaret, C., Dellandrea, E., and Chen, L. (2018). Affective video content analysis: A multidisciplinary insight. *IEEE Trans. Affect. Comput.* 9, 396–409. doi: 10.1109/TAFFC.2017.2661284

Baveye, Y., Dellandr, E., Chen, L., Chamaret, C., and Lyon, D. (2013). A large video database for computational models of induced emotion. *Affect. Comput. Intelli. Int.* 2013, 1–6. doi: 10.1109/ACII.2013.9

Baveye, Y., Dellandréa, E., Chamaret, C., and Chen, L. (2015). LIRIS-ACCEDE: A video database for affective content analysis. *IEEE Trans. Affect. Comput.* 6, 43–55. doi: 10.1109/TAFFC.2015.2396531

Ben, X., Ren, Y., Zhang, J., Wang, S., Kpalma, K., Meng, W., et al. (2021). Video-based facial micro-expression analysis?: A survey of datasets, features and algorithms. *IEEE Trans. Pattern Anal. Mach. Intelli.* 8828, 1–20. doi: 10.1109/TPAMI.2021.3067464

Benini, S., Member, A., Canini, L., Leonardi, R., and Member, S. (2011). A connotative space for supporting movie affective recommendation. *IEEE Trans. Multi.* 13, 1356–1370. doi: 10.1109/TMM.2011.2163058

Bohlin, G., Linderman, K., Alm, C. O., and Bailey, R. (2019). "Considerations for face-based data estimates: Affect reactions to videos," in *VISIGRAPP 2019 - Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, doi: 10.5220/0007687301880194

Boughrara, H., Chtourou, M., Ben Amar, C., and Chen, L. (2016). Facial expression recognition based on a mlp neural network using constructive training algorithm. *Multi. Tools Appl.* 75, 709–731. doi: 10.1007/s11042-014-2322-6

Breitfuss, A., Errou, K., Kurteva, A., and Fensel, A. (2021). Representing emotions with knowledge graphs for movie recommendations. *Future Generat. Comput. Syst.* 125, 715–725. doi: 10.1016/j.future.2021.06.001

Canini, L., Benini, S., Member, A., Leonardi, R., and Member, S. (2013). Affective recommendation of movies based on selected connotative features. *IEEE Trans. Circ. Syst. Video* 23, 636–647. doi: 10.1109/TCSVT.2012.2211935

Cao, W., Zhang, K., Wu, H., Xu, T., Chen, E., Lv, G., et al. (2022). Video emotion analysis enhanced by recognizing emotion in video comments. *Int. J. Data Sci. Anal.* 14, 175–189.

doi: 10.1007/s41060-022-00317-0

Choi, I. Y., Oh, M. G., Kim, J. K., and Ryu, Y. U. (2016). Collaborative filtering with facial expressions for online video recommendation. *Int. J. Inform. Manage.* 36, 397–402. doi: 10.1016/j.ijinfomgt.2016.01.005

Dabas, H., Sethi, C., Dua, C., Dalawat, M., and Sethia, D. (2018). Emotion classification using EEG signals. *ACM Int. Confer. Proc. Seri.* 2018, 380–384. doi: 10.1145/3297156.3297177

Dai, S. (2021). Quantum cryptanalysis on a multivariate cryptosystem based on clipped hopfield neural network. *IEEE Trans. Neural Net. Learn. Syst.* 2021, 1–5. doi: 10.1109/TNNLS.2021.3059434

De Pessemier, T., Coppens, I., and Martens, L. (2019). Using facial recognition services as implicit feedback for recommenders. *CEUR Workshop Proc.* 2450, 28–35.

De Pessemier, T., Coppens, I., and Martens, L. (2020). Evaluating facial recognition services as interaction technique for recommender systems. *Multi. Tools Appl.* 79, 23547–23570. doi: 10.1007/s11042-020-09061-8

Deldjoo, Y., Schedl, M., Cremonesi, P., and Pasi, G. (2018). Content-based multimedia recommendation systems: Definition and application domains. *CEUR Workshop Proc.* 2140, 1–12.

Diaz, Y., Alm, C. O., Nwogu, I., and Bailey, R. (2018). Towards an affective video recommendation system. 2018 IEEE international conference on pervasive computing and communications workshops. *PerCom Workshops* 2018, 137–142. doi: 10.1109/PERCOMW.2018.8480130

Dnodxvndv, U., Ndnodxvndv, D., and Ow, Y. (2018). "Video neuro-advertising recommender model for affective BIM," in *The Proceedings of the 7th International Conference on Computers Communications and Control*, 246–251.

Đorđević Čegar, D., Barreda-Ángeles, M., Kukolj, D., and Le Callet, P. (2020). Modelling effects of S3D visual discomfort in human emotional state using data mining techniques. *Multi. Tools Appl.* 79, 19803–19829. doi: 10.1007/s11042-020-08844-3

Fan, Y., Lu, X., Li, D., and Liu, Y. (2016). "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in *Proceeding of the International Conference on Multimodal*, 1–6. doi: 10.1145/2993148.2997632

Fonnegra, D. (2018). "Deep learning based video spatio-temporal modeling for emotion recognition," in *Proceeding of the International Conference on Human-Computer Interaction*, 397–408. doi: 10.1007/978-3-319-91238-7

Guo, H., Tang, R., Ye, Y., Li, Z., and He, X. (2017). DeepFM: A factorization-machine based neural network for CTR prediction. *IJCAI Int. Joint Confer. Artif. Intelli.* 2017, 1725–1731. doi: 10.24963/ijcai.2017/239

Hassib, M., Pfeiffer, M., Schneegass, S., Rohs, M., and Alt, F. (2017). "Emotion actuator : Embodied emotional feedback through electroencephalography and electrical muscle stimulation," in *Proceeding of the Conference on Human Factors in Computing Systems - Proceedings*, 1–14. doi: 10.1145/3025453.3025953

Hewitt, C., and Gunes, H. (2018). CNN-based facial affect analysis on mobile devices. *Arxiv* [preprint].

Kaklauskas, A., Gudauskas, R., Kozlovas, M., and Peciure, L. (2016). An affect-based multimodal video recommendation system. *Stud. Inform. Control* 25, 1–10. doi: 10.24846/v25i1y201601

Kaklauskas, A., Ubarte, I., Kozlovas, M., Cerkauskas, J., Raupys, D., Lill, I., et al. (2020). Video neuroadvertising recommender system. *J. Internat. Sci. Publi.* 14, 1–9.

Kaklauskas, A., Zavadskas, E. K., Banaitis, A., Liberman, A., Dzitac, S., Ubarte, I., et al. (2018). A neuro-advertising property video recommendation system. *Technol. Forecast. Soc. Change* 131, 78–93. doi: 10.1016/j.techfore.2017.07.011

Krishnamurthy, S. B. M. (2020). Novel OGBEE-based feature selection and feature-level fusion with MLP neural network for social media multimodal sentiment analysis. *Soft Comput.* 24, 18431–18445. doi: 10.1007/s00500-020-05049-6

Kwon, Y.-H., Shin, S.-B., and Kim, S.-D. (2018). Electroencephalography based fusion two-dimensional (2D)-convolution neural networks (CNN) model for emotion recognition system. *Sensors* 18:1383. doi: 10.3390/s18051383

Leite, W., Roy, S., Chakraborty, N., Michailidis, G., Huggins-Manley, A. C., D'Mello, S., et al. (2022). "A novel video recommendation system for algebra: An effectiveness evaluation study," in *Proceeding of the ACM International Conference Proceeding Series*, 294–303. doi: 10.1145/3506860.3506906

Lim, J. Z., Mountstephens, J., and Teo, J. (2020). Emotion recognition using eye-tracking: taxonomy, review and current challenges. *Sensors* 20:2384. doi: 10.3390/s20082384

Liu, J., Su, Y., and Liu, Y. (2017). Multi-modal emotion recognition with temporal-band attention based on LSTM-RNN. *Pacific Rim Conferen. Multi. Springer* 1, 194–204. doi: 10.1007/978-3-319-77380-3

Lucey, S. (2012). Collecting large, richly annotated facial-expression databases from movies. *IEEE Multi.* 19, 34–41. doi: 10.1109/MMUL.2012.26

Ma, F., Li, Y., Ni, S., Huang, S., and Zhang, L. (2022). applied sciences data augmentation for audio – visual emotion recognition with an efficient multimodal conditional GAN. *Appl. Sci.* 12, 527. doi: 10.3390/app12010527

Ma, X., Zhao, L., Huang, G., Wang, Z., Hu, Z., Zhu, X., et al. (2018). Entire space multi-task model: An effective approach for estimating post-click conversion rate. in *Proceeding of the 41st International ACM SIGIR conference on Research and Development in Information Retrieval*. 1137–1140. doi: 10.1145/3209978.3210104

Mahata, A., Saini, N., Saharawat, S., and Tiwari, R. (2016). "Intelligent movie recommender system using machine learning," in *International conference on intelligent human computer interaction* (Cham: Springer), 94–110.

Martha, C., and Larson, M. (2013). Crowdsourcing for affective annotation of video?: Development of a viewer-reported boredom corpus. *IEEE Trans. Circ. Syst. Video Technol.* 23, 1–6.

Mishra, A., Ranjan, P., and Ujlayan, A. (2020). Empirical analysis of deep learning networks for affective video tagging. *Multi. Tools Appl.* 79, 18611–18626. doi: 10.1007/s11042-020-08714-y

Mugellini, E., Khaled, O. A., Bertini, M., and Bimbo, A. (2014). "Towards temporal saliency detection?: Better video understanding for richer tv experiences," in *Proceedings of the ICDS 2014, the 8th international conference on digital society*, Barcelona, 199–202.

Mulholland, E., Mc Kevitt, P., Lunney, T., and Schneider, K. M. (2017). Analysing emotional sentiment in people's YouTube channel comments. *Lect. Notes Instit. Comput. Sci. Soc. Inform. Telecommun. Eng.* 196, 181–188. doi: 10.1007/978-3-319-55834-9_21

Nie, W., Yan, Y., Song, D., and Wang, K. (2020). Multi-modal feature fusion based on multi-layers LSTM for video emotion recognition. *Multi. Tools Appl.* 2020, 1–10. doi: 10.1007/s11042-020-08796-8

Niu, J., Wang, S., Su, Y., and Guo, S. (2017). Temporal factor-aware video affective analysis and recommendation for cyber-based social media. *IEEE Trans. Emerg. Top. Comput.* 5, 412–424. doi: 10.1109/TETC.2017.2705341

Niu, J., Zhao, X., and Aziz, M. A. A. (2016). A novel affect-based model of similarity measure of videos. *Neurocomputing* 173, 339–345. doi: 10.1016/j.neucom.2015.01.104

Niu, J., Zhao, X., Zhu, L., and Li, H. (2013). Affivir: An affect-based Internet video recommendation system. *Neurocomputing* 120, 422–433. doi: 10.1016/j.neucom.2012.07.050

Odic, A., Tkalcic, M., Tasic, J. F., and Kosir, A. (2014). "Personality and social context: Impact on emotion induction from movies," in *Proceedings of the CEUR Workshop*, Rome, 1–7.

Ogawa, T., Sasaka, Y., Maeda, K., and Haseyama, M. (2018). Favorite video classification based on multimodal bidirectional LSTM. *IEEE Access* 6, 61401–61409. doi: 10.1109/ACCESS.2018.2876710

Oliveira, E., Chambel, T., and Pessoa, U. F. (2013). Sharing video emotional information in the web. *Int. J. Web Portals (IJWP)* 5, 19–39. doi: 10.4018/ijwp.2013070102

Orellana-Rodriguez, C., Diaz-Aviles, E., and Nejdl, W. (2015). "Mining affective context in short films for emotion-aware recommendation," in *Proceedings of the 26th ACM Conference on Hypertext and Social Media*, 185–194. doi: 10.1145/2700171.2791042

Pan, Z., Li, X., Cui, L., and Zhang, Z. (2020). Video clip recommendation model by sentiment analysis of time-sync comments. *Multi. Tools Appl.* 79, 33449–33466. doi: 10.1007/s11042-019-7578-4

Polignano, M. (2015). The inuence of user's emotions in recommender systems for decision making processes. *CEUR Workshop Proc.* 1462, 58–66.

Qi, F., Yang, X., and Xu, C. (2021). Emotion knowledge driven video highlight detection. *IEEE Trans. Multi.* 23, 3999–4013. doi: 10.1109/TMM.2020.3035285

Raheem, K. R., and Ali, I. H. (2020). Survey: Affective recommender systems techniques. *IOP Confer. Seri. Mater. Sci. Eng.* 928, 1–11. doi: 10.1088/1757-899X/928/3/032042

Roy, S., and Guntuku, S. C. (2016). "Latent factor representations for cold-start video recommendation. RecSys 2016," in *Proceedings of the 10th ACM Conference on Recommender Systems*, 99–106. doi: 10.1145/2959100.2959172

Santamaria-Granados, L., Mendoza-Moreno, J. F., and Ramirez-Gonzalez, G. (2021). Tourist recommender systems based on emotion recognition—a scientometric review. *Future Int.* 13, 1–38. doi: 10.3390/fi13010002

Shi, Y., Larson, M., and Hanjalic, A. (2013). Mining contextual movie similarity with matrix factorization for context-aware recommendation. *ACM Trans. Intelli. Syst. Technol.* 4, 1–19. doi: 10.1145/2414425.2414441

Shu, J., Shen, X., Liu, H., Yi, B., and Zhang, Z. (2018a). A content-based recommendation algorithm for learning resources. *Multi. Syst.* 24, 163–173. doi: 10.1007/s00530-017-0539-8

Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., et al. (2018b). A review of emotion recognition using physiological signals. *Sensors* 18:2074. doi: 10.3390/s18072074

Singh, P. K., Pramanik, P. K. D., Dey, A. K., and Choudhury, P. (2021). Recommender systems : An overview, research trends, and future directions pradeep kumar singh *, pijush kanti dutta pramanik, avick kumar dey and prasenjit choudhury. *Int. J. Bus. Syst. Res.* 15, 14–52. doi: 10.1504/IJBSR.2021.111753

Sivakumar, N., Balaganesh, N., and Muneeswaran, K. (2015). Feature selection for recommendation of movies. in global conference on communication technologies. *GCCT* 2015, 250–255. doi: 10.1109/GCCT.2015.7342661

Soleymani, M., Davis, J., and Pun, T. (2009). "A collaborative personalized affective video retrieval system," in *Proceeding of the International Conference on Affective Computing & Intelligent Interaction & Workshops*, 1–3. doi: 10.1109/ACII.2009.5349526

Soleymani, M., Larson, M., Pun, T., and Hanjalic, A. (2014). Corpus development for affective video indexing. *IEEE Trans. Multi.* 16, 1075–1089. doi: 10.1109/TMM.2014.2305573

Soleymani, M., Member, S., and Lee, J. (2012). DEAP?: A database for emotion analysis using physiological signals. *IEEE Trans. Affect. Comput.* 3, 18–31. doi: 10.1109/T-AFFC.2011.15

Soleymani, M., and Pantic, M. (2012). Multimodal emotion recognition in response to videos. *IEEE Trans. Affect. Comput.* 3, 211–223. doi: 10.1109/T-AFFC.2011.37

Soleymani, M., Yang, Y. H., Irie, G., and Hanjalic, A. (2015). Guest editorial: Challenges and perspectives for affective analysis in multimedia. *IEEE Trans. Affect. Comput.* 6, 206–208. doi: 10.1109/TAFFC.2015.2445233

Song, H., and Yang, W. (2022). GSCCTL?: A general semi-supervised scene classification method for remote sensing images based on clustering and transfer learning. *Int. J. Remote Sen.* 2022, 1–25. doi: 10.1080/01431161.2021.2019851

Soni, Y., Alm, C. O., and Bailey, R. (2019). "Affective video recommender system," in *Proceeding of the 2019 IEEE Western New York Image and Signal Processing Workshop, WNYISPW 2019*, doi: 10.1109/WNYIPW.2019.8923087

Srivastava, S. K., and Roy, S. N. (2014). Connotative features based affective movie recommendation system. *ICICES* 2014, 111–127. doi: 10.4018/978-1-5225-2851-7.ch008

Tkalčič, M., Burnik, U., Odić, A., Košir, A., and Tasiè, J. (2013a). Emotion-aware recommender systems - A framework and a case study. *Adv. Intelli. Syst. Comput.* 207, 141–150. doi: 10.1007/978-3-642-37169-1_14

Tkalčič, M., Košir, A., and Tasič, J. (2013b). The LDOS-PerAff-1 corpus of facial-expression video clips with affective, personality and user-interaction metadata. *J. Multi. User Int.* 7, 143–155. doi: 10.1007/s12193-012-0107-7

Tkalčič, M., Košir, A., and Tasić, J. (2011a). "Affective recommender systems: The role of emotions in recommender systems," in *Proceedings of the RecSys 2011 workshop on human decision making in recommender systems* (Chicago, IL: ACM), 9–13.

Tkalčič, M., Kosir, A., and Tasič, J. (2011b). Usage of affective computing in recommender systems. *Elektrotehniski Vestnik/Electrotechnical Rev.* 78, 12–17.

Tkalčič1, M., de Gemmis2, M., and Semeraro, G. (2014). Personality and emotions in decision making and recommender systems. *Int. Workshop Dec. Mak. Recommender Syst.* 2014, 1–5.

Tripathi, A., Ashwin, T. S., and Guddeti, R. M. R. (2019). EmoWare: A context-aware framework for personalized video recommendation using affective video sequences. *IEEE Access* 7, 51185–51200. doi: 10.1109/ACCESS.2019.2911235

Tripathi, A., Manasa, D. G., Rakshitha, K., Ashwin, T. S., and Reddy, G. R. M. (2018). Role of intensity of emotions for effective personalized video recommendation?: A reinforcement learning approach. *Recent Find. Intelli. Comput. Techn.* 2018, 507–517. doi: 10.1007/978-981-10-8633-5

Wakil, K., Bakhtyar, R., Ali, K., and Alaadin, K. (2015). Improving web movie recommender system based on emotions. *Int. J. Adv. Comput. Sci. Appl.* 6:60232. doi: 10.14569/IJACSA.2015.060232

Wang, D., and Chen, Y. (2020). A novel many-objective recommendation algorithm for multistakeholders. *IEEE Access* 8, 196482–196499. doi: 10.1109/ACCESS.2020.3034716

Wang, D., and Chen, Y. (2021). A novel cascade hybrid many-objective recommendation algorithm incorporating multistakeholder concerns. *Inform. Sci.* 577, 105–127. doi: 10.1016/j.ins.2021.07.005

Wang, H. L., and Cheong, L. (2006). Affective understanding in film. *IEEE Trans. Circ. Syst. Video* 16, 689–704. doi: 10.1109/TCSVT.2006.873781

Wang, S., Chen, Y., Ming, H., Mi, L., and Shi, Z. (2020). Improved danmaku emotion analysis and its application based on bi-LSTM model. *IEEE Access* 99, 114123–114134. doi: 10.1109/ACCESS.2020.3001046

Winoto, P., and Tang, T. Y. (2010). The role of user mood in movie recommendations. *Exp. Syst. Appl.* 37, 6086–6092. doi: 10.1016/j.eswa.2010.02.117

Wu, Z., Huang, S., Zhang, R., and Li, L. (2020). "Video review analysis via transformer-based sentiment change detection," in *Proceeding of the 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 330–335. doi: 10.1109/MIPR49039.2020.00074

Yadati, K., Katti, H., and Kankanhalli, M. (2014). CAVVA: Computational affective video-in-video advertising. *IEEE Trans. Multi.* 16, 15–23. doi: 10.1109/TMM.2013.2282128

Yang, H., Han, J., and Min, K. (2019). A multi-column CNN model for emotion recognition from EEG signals. *Sensors* 19, 1–12. doi: 10.3390/s19214736

Yenter, A. (2017). "Deep CNN-LSTM with combined kernels from multiple branches for IMDb review sentiment analysis," in *Proceeding of the 2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, 540–546. doi: 10.1109/UEMCON.2017.8249013

Zhang, L., and Zhang, J. (2017). "Synchronous prediction of arousal and valence using LSTM network for affective video content analysis," in *Proceeding of the 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 727–732. doi: 10.1109/FSKD.2017.8393364

Zhang, S., Yao, L., Sun, A., and Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. *ACM Comput. Surv.* 52:5029. doi: 10.1145/3285029

Zhang, S., Zhang, S., Huang, T., and Member, S. (2018). Learning affective features with a hybrid deep model for audio – visual emotion recognition. *IEEE Trans. Circ. Syst. Video Technol.* 28, 3030–3043. doi: 10.1109/TCSVT.2017.2719043

Zhang, S., Zhao, X., and Tian, Q. (2022). Spontaneous speech emotion recognition using multiscale deep convolutional LSTM. *IEEE Trans. Affect. Comput.* 13, 680–688. doi: 10.1109/TAFFC.2019.2947464

Zhang, Y., and Chen, X. (2020). Explainable recommendation: A survey and new perspectives. *Found. Trends Inform. Retrieval* 14, 1–101. doi: 10.1561/1500000066

Zhao, S., Yao, H., and Sun, X. (2013). Video classification and recommendation based on affective analysis of viewers. *Neurocomputing* 119, 101–110. doi: 10.1016/j.neucom.2012.04.042

Zhao, S., Yao, H., Sun, X., Xu, P., Liu, X., and Ji, R. (2011). "Video indexing and recommendation based on affective analysis of viewers BT," in *Proceeding of the 19th ACM International Conference on Multimedia ACM Multimedia 2011, MM'*. doi: 10.1145/2072298.2072043

Zhu, Y., Tong, M., Jiang, Z., Zhong, S., and Tian, Q. (2019). Hybrid feature-based analysis of video ' s affective content using protagonist detection. *Exp. Syst. Appl.* 128, 316–326. doi: 10.1016/j.eswa.2019.03.017

TYPE Original Research
PUBLISHED 15 September 2022
DOI 10.3389/fnbot.2022.987146

OPEN ACCESS

EDITED BY
Shiqing Zhang,
Taizhou University, China

REVIEWED BY
Yun Jin,
Jiangsu Normal University, China
Peng Song,
Yantai University, China

*CORRESPONDENCE
Yuan Zong
xhzongyuan@seu.edu.cn

RECEIVED 05 July 2022
ACCEPTED 25 July 2022
PUBLISHED 15 September 2022

CITATION
Zong Y, Lian H, Zhang J, Feng E, Lu C, Chang H and Tang C (2022) Progressive distribution adapted neural networks for cross-corpus speech emotion recognition.
Front. Neurorobot. 16:987146.
doi: 10.3389/fnbot.2022.987146

Progressive distribution adapted neural networks for cross-corpus speech emotion recognition

Yuan Zong[1,2]*, Hailun Lian[1], Jiacheng Zhang[1,3], Ercui Feng[4], Cheng Lu[1], Hongli Chang[1] and Chuangao Tang[1,2]

[1]Key Laboratory of Child Development and Learning Science of Ministry of Education, Southeast University, Nanjing, China, [2]School of Biological Science and Medical Engineering, Southeast University, Nanjing, China, [3]School of Cyber Science and Engineering, Southeast University, Nanjing, China, [4]Affiliated Jiangning Hospital, Nanjing Medical University, Nanjing, China

In this paper, we investigate a challenging but interesting task in the research of speech emotion recognition (SER), i.e., cross-corpus SER. Unlike the conventional SER, the training (source) and testing (target) samples in cross-corpus SER come from different speech corpora, which results in a feature distribution mismatch between them. Hence, the performance of most existing SER methods may sharply decrease. To cope with this problem, we propose a simple yet effective deep transfer learning method called progressive distribution adapted neural networks (PDAN). PDAN employs convolutional neural networks (CNN) as the backbone and the speech spectrum as the inputs to achieve an end-to-end learning framework. More importantly, its basic idea for solving cross-corpus SER is very straightforward, i.e., enhancing the backbone's corpus invariant feature learning ability by incorporating a progressive distribution adapted regularization term into the original loss function to guide the network training. To evaluate the proposed PDAN, extensive cross-corpus SER experiments on speech emotion corpora including EmoDB, eNTERFACE, and CASIA are conducted. Experimental results showed that the proposed PDAN outperforms most well-performing deep and subspace transfer learning methods in dealing with the cross-corpus SER tasks.

KEYWORDS

cross-corpus speech emotion recognition, speech emotion recognition, deep transfer learning, domain adaptation, deep learning

1. Introduction

Speech is one major way human beings communicate in daily life, which carries abundant emotional information. Consider that if computers were able to understand the emotional states of human beings' speech signals, human-computer interaction would undoubtedly be more natural. Consequently, the research of automatically recognizing emotional states from speech signals, a. k. a., speech emotion recognition (SER) has attracted wide attention among the affective computing, human-computer interaction, and speech signal processing communities (El Ayadi et al., 2011; Schuller, 2018). Over the past several decades, many well-performing SER methods have been proposed and achieved promising performance on widely-used publicly available speech emotion corpora (Zong et al., 2016; Zhang et al., 2017, 2022; Kwon, 2021; Lu et al., 2022). However, it is noted that most of them did not consider the realistic scenario where the training and testing speech signals are possibly recorded by different microphones or in different environments. In this case, a feature distribution mismatch may exist between the training and testing speech samples, and hence the performance of these originally well-performing SER methods may decrease sharply. This brings us a meaningful and more challenging task in SER, i.e., cross-corpus SER. Unlike the conventional SER, the labeled training and unlabeled testing samples in cross-corpus SER come from different speech corpora. Following the naming conventions in cross-corpus SER, we will refer to the training and testing samples/corpora/feature sets as the source and target ones throughout this paper in what follows.

In recent years, researchers have been devoted to the research of cross-corpus SER and proposed many promising methods. Schuller et al. (2010b) may be the first to have investigated this problem, and designed three different normalization methods including speaker normalization (SN), corpus normalization (CN), and speaker-corpus normalization (SCN) to alleviate the feature distribution mismatch between the source and target speech samples. Since that, lots of transfer learning and domain adaptation methods have been successively designed to deal with cross-corpus SER tasks. For example, Hassan et al. (2013) proposed to compensate for the corpus shift by reweighting the source speech samples to deal with cross-corpus SER tasks. A new version of the modified support vector machine (SVM) called importance-weighted SVM (IW-SVM) was designed by incorporating three typical transfer learning methods including kernel mean matching (KMM) (Gretton et al., 2009), unconstrained least-squares importance fitting (uLSIF) (Kanamori et al., 2009), and Kullback-Leibler importance estimation procedure (KLIEP) (Tsuboi et al., 2009) to learn the source sample weights. In the work of Song et al. (2016), Song et al. presented a transfer non-negative matrix factorization (TNMF) for the cross-corpus SER problem. The basic idea of TNMF is to decompose the source and target speech feature sets into different non-negative feature matrices under the guidance of maximum mean discrepancy (MMD) (Borgwardt et al., 2006) and hence the gap between the source and target speech signals described by the non-negative matrices can be alleviated (Liu et al., 2018). Moreover, Liu et al. proposed a domain-adaptive subspace learning (DoSL) model to handle the cross-corpus SER problem. This method measures the distribution gap between the source and the target speech samples through a one-order moment, i.e., the mean value of speech feature vectors. Then a subspace learning model enhanced by the one-order moment regularization term is built to learn a projection matrix to transform the source and target speech sample from the original feature space to the labeled one. The transformed source and target speech samples in such label space would share similar feature distributions. More recently, Zhang et al. (2021) further proposed an extended version of DoSL called joint distribution adaptive regression (JDAR) to align the source and target speech feature distributions to remove their mismatch by considering the marginal distribution gap together with the emotion class aware conditional one. By jointly minimizing both feature distribution gaps, the JDAR model can achieve a better performance than DoSL in dealing with the cross-corpus SER tasks.

On the other hand, deep transfer learning techniques have also been used to cope with the cross-corpus SER tasks. Unlike the transfer subspace learning methods, most deep transfer learning ones try to learn a robust deep neural network to learn corpus invariant features to describe the speech signals. For example, Deng et al. (2014, 2017) proposed a series of unsupervised domain adaptation methods based on autoencoder (AE) to bridge the gap between the source and target speech emotion corpora. The basic idea of these methods is to learn a common subspace through AE instead of widely used subspace learning such that the source and target speech signals have the same or similar feature distributions in the learned subspace. Different from the work of Deng et al. (2014, 2017), Abdelwahab and Busso (2018) proposed to use another deep neural network, i.e., deep belief network (DBN), to investigate the cross-language and cross-corpus SER problem on five speech emotion corpora and the experimental results demonstrated more promising performance than sparse AE and SVM based baseline systems. Recently, adversarial learning-based methods have also been applied to coping with cross-corpus SER tasks. Abdelwahab and Busso (2018) made use of adversarial multi-task training to learn a common representation for training and testing speech feature sets. Two tasks were designed to enable the networks to be robust to the corpus variance. Specifically, one task is to build the relationship between the emotion classes and acoustic descriptors of speech signals. The other is to learn the common representation by enforcing the source and target speech features cannot be distinguished. More recently, Gideon et al. (2019) presented an adversarial discriminative domain generalization (ADDoG) model with the help of domain generalization. Unlike most deep transfer learning methods, the ADDoG model used the speech spectrums as the inputs instead of the handcrafted speech features and simultaneously improved its corpus robustness in multiple speech corpora. Following the work of Gideon et al. (2019), Zhao et al. (2022) also used the speech spectrums as the inputs of the networks to achieve the end-to-end learning manner for cross-corpus SER tasks and proposed a deep transductive transfer regression neural network (DTTRN) with an emotion knowledge guided MMD loss to remove the feature distribution mismatch between the source and target speech corpora.

Inspired by the success of the above deep transfer learning methods, in this paper we also focus on the research of designing deep transfer learning methods to deal with the cross-corpus SER tasks. We propose a novel method called progressive distribution adapted neural networks (PDAN). The basic idea of PDAN is very straightforward, i.e., enabling the deep neural networks to directly learn an emotion discriminative and corpus invariant representations for both source and target original speech signals by leveraging the powerful nonlinear mapping ability and hierarchical structure of deep neural networks. Specifically, we first make use of convolutional neural networks to build the relationship between the source emotion label information and speech spectrums to endow the emotion discriminant ability to PDAN. Then, three feature distribution adapted regularization terms are imposed on different fully connected layers to respectively guide the network to learn the corpus invariant common representations for both speech corpora. To evaluate the effectiveness of the PDAN, we conduct extensive cross-corpus SER experiments on three widely-used speech emotion corpora, i.e., EmoDB (Burkhardt et al., 2005), eNTERFACE (Martin et al., 2006), and CASIA (Zhang and Jia, 2008). Experimental results demonstrate the effectiveness and superior performance of PDAN over recent state-of-the-art transfer learning methods in dealing with cross-corpus SER tasks. In summary, the main contributions of this paper include three folds:

1. We proposed a novel end-to-end deep transfer learning model called PDAN to cope with cross-corpus SER tasks. Unlike most existing methods, PDAN can directly learn the corpus invariant and emotion discriminative speech features from the original speech spectrums by resorting to the nonlinear mapping ability of deep neural networks.
2. We presented a new idea of progressively adapting the feature distributions between the source and target speech samples for the proposed PDAN by designing three different derived MMD loss functions.
3. Extensive cross-corpus SER tasks are designed to evaluate the proposed PDAN method. By deeply analyzing the experimental results, several interesting findings and discussions are given in our paper.

2. Proposed method

2.1. Overall picture and notations

In this section, we address the proposed PDAN model in detail and also show how to use PDAN to deal with cross-corpus SER tasks. To this end, we draw a picture shown in Figure 1 to illustrate the basic idea and overall structure of the proposed PDAN. To make the readers better understand this paper, we first introduce some necessary notations which are used in Figure 1 for formulating PDAN. The speech spectrums of source and target speech samples are denoted by $\mathcal{D}_s = \{\mathcal{X}_1^s, \cdots, \mathcal{X}_{N_s}^s\}$ and $\mathcal{D}_t = \{\mathcal{X}_1^t, \cdots, \mathcal{X}_{N_t}^t\}$, respectively, where N_s and N_t are the source and target sample numbers. According to the task setting of cross-corpus SER, the source emotion labels are given, while the target ones are entirely unknown. Hence, we denote the source emotion labels by $\mathcal{Y}^s = \{\mathbf{y}_1^s, \cdots, \mathbf{y}_{N_s}^s\}$. Note that the i^{th} sample's emotion label $\mathbf{y}_i^s \in \mathbb{R}^{C\times 1}$ is a one-hot vector whose k^{th} entry would be 1 while the others are all 0 if its corresponding label was k^{th} of C emotions.

2.2. Formulating PDAN

As described in Sect. Introduction, the basic idea of PDAN is very straightforward, i.e., building an **emotion discriminative** and **corpus invariant** end-to-end neural network for cross-corpus SER. To achieve this goal, we first construct a convolutional neural network (CNN) consisting of a set of convolutional layers and three fully connected (FC) layers to serve as the basic structure of PDAN. Then, to achieve the goal of end-to-end learning, we transform the original speech signals into spectrums to serve as the inputs of the PDAN. Note that in PDAN, the source and target speech spectrums will be simultaneously fed to train the PDAN, which can also be interpreted as inputting them into two weight-shared CNNs shown in Figure 1. Subsequently, it is clear to see from Figure 1 that our PDAN has four major loss functions to guide the network training, i.e., $\mathcal{L}_s$, $\mathcal{L}_m$, $\mathcal{L}_{rc}$, and $\mathcal{L}_{fc}$, respectively, which correspond to the basic idea of the proposed PDAN. The first loss function is called emotion discriminative loss denoted by $\mathcal{L}_s$, which is designed for enabling the network to be **emotion discriminative** and can be formulated as

$$\mathcal{L}_s = \frac{1}{N_s}\sum_{i=1}^{N_s} \mathcal{J}_{CE}(g_3(g_2(g_1(f(\mathcal{X}_i^s)))), \mathbf{y}_i^s), \quad (1)$$

where $\mathcal{J}_{CE}$ is the cross-entropy loss bridging the source speech spectrums and their corresponding emotion labels, g_1, g_2, and g_3 are the parameters of fully connected layers, and f denotes the parameters of the convolutional layers, respectively.

As for the resting loss functions, they aim to improve the robustness of the speech features learned by PDAN to the **corpus invariance**. To this end, based on the MMD criterion (Borgwardt et al., 2006), we first design marginal distribution adapted loss $\mathcal{L}_m$ and impose it on the first FC layer in PDAN, which is formulated as follows:

$$\mathcal{L}_m = \|\frac{1}{N_s}\sum_{i=1}^{N_s} \Phi(g_1(f(\mathcal{X}_i^s))) - \frac{1}{N_t}\sum_{i=1}^{N_t} \Phi(g_1(f(\mathcal{X}_i^t)))\|_{\mathcal{H}}^2, \quad (2)$$

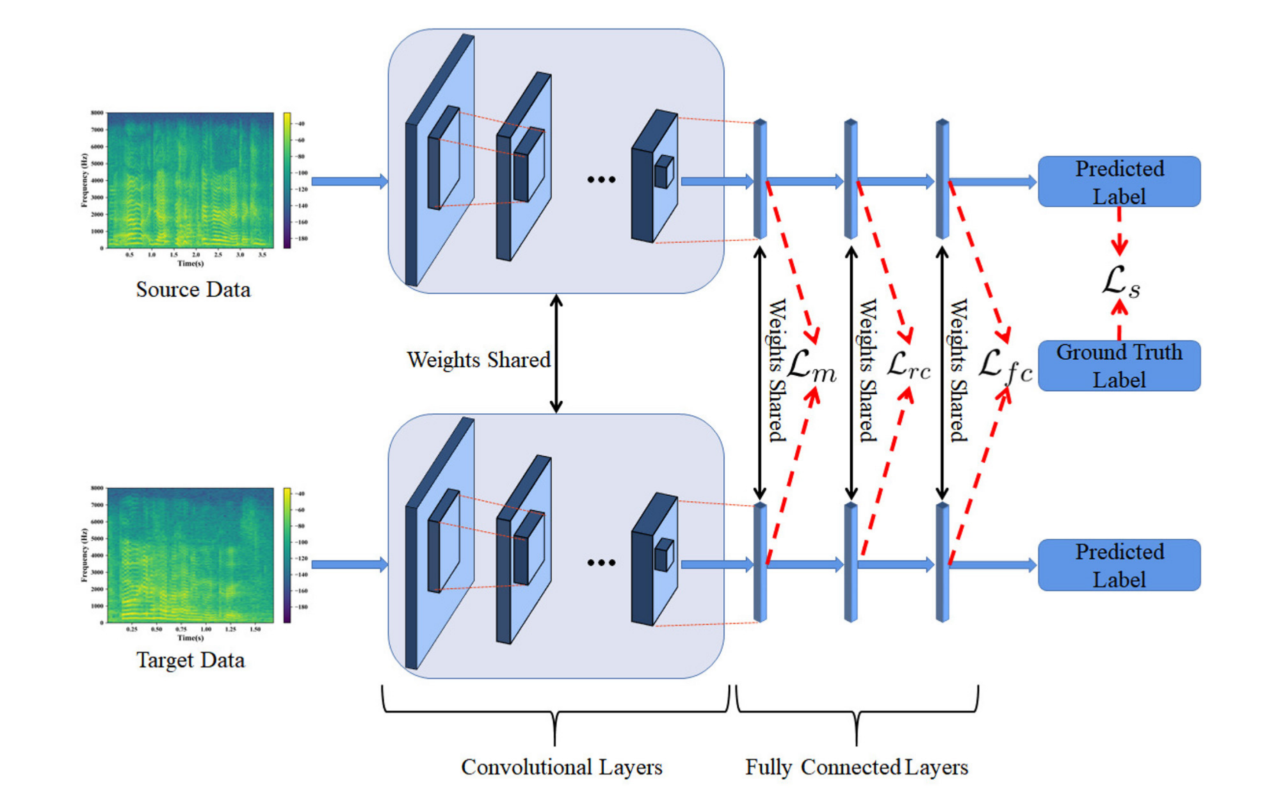


FIGURE 1
The overview of progressive distribution adapted neural networks (PDAN). The PDAN uses the speech spectrums as the inputs and directly builds the relationship between the emotion labels and speech signals. It consists of several convolutional layers and three fully connected (FC) layers and is trained under the guidance of the combination of four loss functions, i.e., emotion discriminative loss $\mathcal{L}_s$, marginal distribution adapted loss $\mathcal{L}_m$, rough emotion class aware conditional distribution adapted loss $\mathcal{L}_{rc}$, and fine emotion class aware conditional distribution adapted loss $\mathcal{L}_{fc}$.

where $\mathcal{L}_m$ is the square of the original MMD function and can be used to measure the marginal distribution difference between the source and target feature sets, $\Phi(\cdot)$ is the kernel mapping operator, and $\|\cdot\|_{\mathcal{H}}$ means the inner product in such reproduced kernel Hilbert space (RKHS).

Secondly, we design a fine emotion class aware conditional distribution adapted loss $\mathcal{L}_{fc}$, which is added to regularize the last FC layer and can be expressed as follows:

$$\mathcal{L}_{fc} = \frac{1}{C}\sum_{j=1}^{C}\|\frac{1}{N_{s_j}}\sum_{i=1}^{N_{s_j}}\Phi(g_3(g_2(g_1(f(\mathcal{X}_i^s))))) \\ -\frac{1}{N_t}\sum_{i=1}^{N_{t_j}}\Phi(g_3(g_2(g_1(f(\mathcal{X}_i^t)))))\|_{\mathcal{H}}^2, \quad (3)$$

where $\mathcal{X}_i^{s_j}$ and $\mathcal{X}_i^{t_j}$ correspond to the speech samples belonging to the j^{th} emotion and N_{s_j} and N_{t_j} denote their sample numbers satisfying $N_{s_1} + \cdots + N_{s_C} = N_s$ and $N_{t_1} + \cdots + N_{t_C} = N_t$, respectively. Hence, it is clear that $\mathcal{L}_{fc}$ can be used to measure the fine emotion class aware conditional feature distribution gap between the source and target speech features.

Finally, we consider designing a rough emotion class aware conditional distribution adapted regularization term, i.e., $\mathcal{L}_{rc}$, to guide the feature learning in the second FC layer, whose formulation is as follows:

$$\mathcal{L}_{rc} = \frac{1}{C_r}\sum_{j=1}^{C_r}\|\frac{1}{N_{s_j}}\sum_{i=1}^{N_{s_j}}\Phi(g_2(g_1(f(\mathcal{X}_i^s)))) \\ -\frac{1}{N_t}\sum_{i=1}^{N_{t_j}}\Phi(g_2(g_1(f(\mathcal{X}_i^t))))\|_{\mathcal{H}}^2, \quad (4)$$

where $C_r < C$ can be called a rough emotion class number.

Note that $\mathcal{L}_{rc}$ shown in Equation (4) looks like a new measurement of conditional distribution mismatch between the source and target speech features, which is so similar to $\mathcal{L}_{fc}$ in Equation (3). However, they are actually very different. Specifically, in $\mathcal{L}_{rc}$, a set of emotion classes involved in cross-corpus SER will merge together and then the conditional MMD is calculated. This is motivated by the work of the valance-arousal emotion wheel proposed by Yang et al. (2022) shown in Figure 2. As Figure 2 shows, it is clear to see that

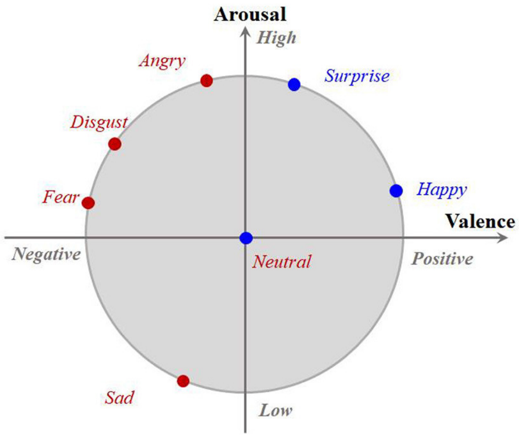


FIGURE 2
The 2D arousal-valence emotion wheel proposed by Yang et al. (2022). It consists of two dimensions, where the horizontal axis denotes the degree of valence while the vertical axis corresponds to the arousal. Each typical discrete emotion can be mapped to one point in the emotion wheel according to its corresponding valence and arousal values.

most of the existing typical emotions are all high-arousal and only a few emotions, e.g., *Sad*, are low-arousal. It is also interesting to see that along the valence dimension, the separability among these emotions would be significantly improved. For example, we can observe from Figure 2 that *Angry*, *Disgust*, and *Fear* are low-valence, while *Surprise* and *Happy* are high-valence although they all belong to the high-arousal ones. Inspired by the above observations, we propose to align the rough emotion-aware conditional distributions with respect to the valence dimension in the second FC layer and hence design $\mathcal{L}_{rc}$ to further improve the corpus invariance of the proposed PDAN together with the resting two ones. It should be noticed that since the features in shallow layers have limited discriminative ability, it may be a tough task to directly align the fine emotion class aware conditional distribution gap between the source and target speech features together with the marginal one in the first FC layer. Therefore, we assign the fine emotion class aware conditional distribution term to the last FC layer instead of the first one because such features in the deepest FC layer would be more emotion-discriminative. According to the granularity of the emotion class information used in calculating these three feature distribution adapted terms, it can be seen that the feature distribution adaption operations of PDAN are present in a progressive way. This is why we call the proposed method PDAN.

Under the above considerations, we are able to arrive at the optimization problem of the proposed PDAN by jointly minimizing the four well-designed losses, which can be expressed as follows:

$$\min_{f,g_1,g_2,g_3} \mathcal{L}_{total} = \mathcal{L}_s + \lambda_1\mathcal{L}_m + \lambda_2\mathcal{L}_{rc} + \lambda_3\mathcal{L}_{fc}, \qquad (5)$$

where λ_1, λ_2, and λ_3 are the trade-off parameters controlling the balance among the four losses.

2.3. Optimization of PDAN

Since the calculation of two conditional distribution adapted loss needs the target label information, we optimize the optimization problem of PDAN by using an alternated direction method. Specifically, we first randomly initialize the parameters of PDAN, i.e., f, g_1, g_2, and g_3, and then predict the pseudo emotion labels of target speech samples denoted by $\mathbf{L}_t^p$. Subsequently, perform the following two major steps until convergence:

1. According to $\mathbf{L}_t^p$, calculate the loss functions $\mathcal{L}_{total}$ and update the parameters of PDAN, i.e., f, g_1, g_2, and g_3, by the typical optimization algorithm, e.g., SGD and Adam.
2. Fix f, g_1, g_2, and g_3, and update the pseudo target emotion labels $\mathbf{L}_t^p$.

Note that in PDAN, the kernel trick can be used to effectively calculate three MMD based losses, which can be formulated as follows:

$$\begin{aligned}\mathbf{MMD}^2(\mathbf{X}^s,\mathbf{X}^t) &= \|\frac{1}{N_s}\sum_{i=1}^{N_s}\Phi(\mathbf{x}_i^s))) - \frac{1}{N_t}\sum_{i=1}^{N_t}\Phi(\mathbf{x}_i^t)))\|_{\mathcal{H}}^2,\\ &= \frac{N_s}{N_s(N_s-1)}\sum_{i\neq j}^{N_s}k(\mathbf{x}_i^s,\mathbf{x}_j^s) + \frac{1}{N_t(N_t-1)}\\ &\quad \sum_{i\neq j}^{N_t}k(\mathbf{x}_i^t,\mathbf{x}_j^t) - \frac{2}{N_sN_t}\sum_{i,j=1}^{N_s,N_t}k(\mathbf{x}_i^s,\mathbf{x}_j^t),\end{aligned} \qquad (6)$$

where $k(\cdot)$ is a kernel function replacing the inner product operation between vectors in RKHS produced by $\Phi(\cdot)$ with calculating a predefined function, and $\mathbf{x}_i^s$ and $\mathbf{x}_i^t$ are the i^{th} column in $\mathbf{X}^s$ and $\mathbf{X}_i^t$.

Finally, we summarize the detailed procedures for updating PDAN in Algorithm 1 such that the readers can better understand how to optimize the proposed PDAN.

3. Experiments

3.1. Speech emotion corpora and protocol

In this section, we design extensive cross-corpus SER tasks to evaluate the proposed PDAN method. Three public available speech emotion corpora including EmoDB (Burkhardt et al.,

```
Input: Source Speech Spectrums: $\mathcal{D}_s = \{\mathcal{X}_1^s, \cdots, \mathcal{X}_{N_s}^s\}$,
    Target Speech Spectrums: $\mathcal{D}_t = \{\mathcal{X}_1^t, \cdots, \mathcal{X}_{N_t}^s\}$,
    Learning Rate: $\alpha$,
    Trade-off Parameters: $\lambda_1, \lambda_2$, and $\lambda_3$,
    Maximal Iterations: $N_{max}$.
Output: Optimal Network Parameters: $f = \hat{f}$, $g_1 = \hat{g}_1$,
    $g_2 = \hat{g}_2$, and $g_3 = \hat{g}_3$.
Initialize the network parameters: $\tilde{f}$, $\tilde{g}_1$, $\tilde{g}_2$, and
    $\tilde{g}_3$, and iteration indicator: $iter = 0$.
while $\mathcal{L}_{total} \neq 0 \quad \| \quad iter < N_{max}$ do
    $iter = iter + 1$;
    Fix $f$, $g_1$, $g_2$, and $g_3$, predict the pseudo label
        $\mathbf{L}_t^p$;
    Fix $\mathbf{L}_t^p$, calculate $\mathcal{L}_{total}$;
    Update $f$, $g_1$, $g_2$, and $g_3$:
        $\nabla_{\boldsymbol{\theta}} \leftarrow \frac{\partial(\mathcal{L}_s + \lambda_1 \mathcal{L}_m + \lambda_2 \mathcal{L}_{rc} + \lambda_3 \mathcal{L}_{fc})}{\partial \boldsymbol{\theta}}$, where $\boldsymbol{\theta} = \{f, g_1, g_2, g_3\}$;
        $\boldsymbol{\theta}^{n+1} \leftarrow \boldsymbol{\theta}^n - \alpha \nabla_{\boldsymbol{\theta}}$;
end while
```

Algorithm 1. The detailed procedures for updating optimization problem of PDAN in Equation (5).

2005), eNTERFACE (Martin et al., 2006), and CASIA (Zhang and Jia, 2008), are chosen. EmoDB is one of the most widely-used German acted speech emotion corpora collected by Burkhardt et al. from TU Berlin, Germany. Ten participants consisting of five women and five men were recruited to simulate seven types of emotions, i.e., *Neutral*, *Angry*, *Fear*, *Happy*, *Sad*, *Disgust*, and *Boredom*, respectively. The total sample number reaches 545 and can be downloaded from the http://www.expressive-speech.net/emodb/. eNTERFACE is an induced audio-video bi-modal emotion database. We only adopted its audio part and the language is English. It consists of 1,257 speech samples from 41 independent speakers comprising six basic emotions, i.e., *Disgust*, *Sad*, *Angry*, *Happy*, *Fear*, and *Surprise*, respectively. CASIA is a Chinese acted speech corpus designed by the Institute of Automation, Chinese Academy of Science. It recruited four speakers including two women and two men to record 1,200 speech samples from six typical emotions, i.e., *Neutral*, *Surprise*, *Angry*, *Happy*, *Fear*, and *Sad*.

By alternatively using either two of these three speech emotion corpora to serve as the source and target domains, six cross-corpus SER tasks are designed denoted by $B \rightarrow E$, $B \rightarrow E$, $B \rightarrow E$, $B \rightarrow E$, $B \rightarrow E$, and $B \rightarrow E$, respectively. Note that B, E, and C are the abbreviations of EmoDB, eNTERFACE, and CASIA. The left and right corpora of the arrow denote the source and target ones in such a cross-corpus SER task. Since these three corpora have inconsistent emotion labeling information, in each task we select the speech samples sharing the same emotion label from the corresponding source and target corpora. To make the readers better know the detail of the sample information in each cross-corpus SER task, we summarize the sample statistics of speech corpora used in all six tasks in Table 1. As for the performance metric, we choose unweighted average recall (UAR) (Schuller et al., 2010b) defined as the accuracy per class averaged by the total emotion class number, which is widely used in evaluating SER methods. For comparison purpose, five typical transfer subspace learning methods, i.e., Transfer Component Analysis (TCA) (Pan et al., 2010), Geodesic Flow Kernel (GFK) (Gong et al., 2012), Subspace Alignment (SA) (Fernando et al., 2013), Domain Adaptive Subspace Learning (DoSL) (Liu et al., 2018), and Joint Distribution Adaptive Regression (JDAR) (Zhang et al., 2021), respectively, and four deep transfer learning methods, i.e., Deep Adaptation Networks (DAN) (Long et al., 2015), Domain-Adversarial Neutral Network (DANN) (Ajakan et al., 2014), Deep-CORAL (Sun and Saenko, 2016), and Deep Subdomain Adaptation Network (DSAN) (Zhu et al., 2020), respectively, are included.

3.2. Implementation details

First, as for the subspace learning comparison methods, we choose two types of speech feature sets, i.e., IS09 (Schuller et al., 2009) and IS10 (Schuller et al., 2010a) to describe speech signals, respectively. The IS09 feature set consists of 384 elements including 16×2 acoustic low-level descriptors (LLDs) such as fundamental frequency (F0), zero-crossing rate (ZCR), and Mel-frequency cepstrum coefficient (MFCC), and their first order difference, and their 12 corresponding functions such as maximal value, mean value, and minimal value. The IS10 feature set has 1,582 elements which are obtained by applying 21 statistical functions to 38 LLDs and their first order derivatives plus 2 single features about F0 (the number of onsets and tern duration) and discarding 16 zero-information features (e.g., minimum F0). The detailed information of these two feature sets are referred to in the works of Schuller et al. (2009) and Schuller et al. (2010a), respectively. In the experiments, the openSIMLE toolkit (Eyben et al., 2010) is used to extract the IS09 and IS10 feature sets. The hyper-parameters of all the subspace learning methods are set as follows:

1. **TCA**, **GFK**, and **SA**: A hyper-parameter, i.e., the reduced dimension denoted by d, needs to be set for TCA, GFK, and SA. In the experiments, we search the d from a parameter interval $[5:5:d_{max}]$, where d_{max} is the maximal dimension reduced by these three methods in each experiment.
2. **DoSL** and **JDAR**: There are two hyper-parameters in DoSL and JDAR methods, i.e., λ and μ. They are used to control the balance between the original regression loss function and two regularization terms including feature selection and feature distribution difference alleviation terms. In the experiments, they are both searched from the parameter interval $[5:5:100]$. In addition, since the JDAR method

TABLE 1 The sample statistics of EmoDB (B), eNTERFACE (E), and CASIA (C) corpora used in the designed six cross-corpus SER tasks.

Tasks	Speech corpus (# Samples belonging to each emotion)	Total
$B \rightarrow E$	B (Angry: 127, Sad: 62, Fear: 69, Happy: 71, Disgust: 46)	375
$E \rightarrow B$	E (Angry: 211, Sad: 211, Fear: 211, Happy: 208, Disgust: 211)	1,052
$B \rightarrow C$	B (Angry: 127, Sad: 62, Fear: 69, Happy: 71, Neutral: 79)	408
$C \rightarrow B$	C (Angry: 200, Sad: 200, Fear: 200, Happy: 200, Neutral: 200)	1,000
$E \rightarrow C$	E (Angry: 211, Sad: 211, Fear: 211, Happy: 208, Surprise: 211)	1,052
$C \rightarrow E$	C (Angry: 200, Sad: 200, Fear: 200, Happy: 200, Surprise: 200)	1,000

needs to iteratively predict the pseudo emotion labels of the target speech signals and calculate the emotion class aware conditional distribution gap between the source and target speech feature sets, we set the iterations as 5 for JDAR in all the cross-corpus SER tasks.

Second, as for the deep learning methods including our PDAN, we first transform the original speech signals into speech spectrums to serve as the inputs of all the methods. Specifically, for each speech sample from the emotion corpora, we set the frame size and overlap as 350 and 175 sampling points, respectively, and then all the speech frames windowed by the Hamming function were transformed to spectrums by using Fourier transformation to compose the speech spectrums. Note that in speech spectrum generation, the sampling frequencies used for EmoDB, eNTERFACE, and CASIA are 16, 44, and 16 kHz, respectively. In the implementations of all the deep learning methods, the Adam optimizer is used to train the model. Its three parameters, i.e., β_1, β_2, and weight decay λ are set as 0.9, 0.999, and 0.005, respectively. During the training stage, the batch size and the initial learning rate are set to 32 and 0.0002, respectively. AlexNet (Krizhevsky et al., 2012) is served as the CNN backbone of all the deep learning methods and only the neuron number of the last fully connected layer is reset as the one involving emotion class number in each cross-corpus SER task. Moreover, since most of the comparison methods adopt MMD losses, following the work of Long et al. (2015) and Zhu et al. (2020), we use the mixed Gaussian function to serve as the kernel function, i.e., $\mathbf{K} = \sum_{i=1}^{5} K_i$, where $K_i(\mathbf{u}, \mathbf{v}; \sigma_i) = e^{\frac{-\|\mathbf{u}-\mathbf{v}\|^2}{2\sigma_i^2}}$, where σ_i denotes the bandwidth and its value range is $[2, 4, 8, 16, 32]$. Finally, the trade-off parameter of each comparison methods is set as follows:

1. **DAN** and **DSAN**: There is only one trade-off parameter in DAN and DSAN. We set its interval as $[0.001, 0.005, 0.01, 0.05, 0.1, 0.5]$.
2. **DANN**: DANN also has only one trade-off parameter. We set its searching range as $[0.001, 0.003, 0.005, 0.01, 0.05, 0.1, 0.5]$.
3. **Deep-CORAL**: Similar to the above deep transfer learning methods, one trade-off parameter in Deep-CORAL needs to be set. In the experiments, its interval is $[1, 10, 20, 30, 50, 100]$.
4. **PDAN**: The proposed PDAN has three trade-off parameters, i.e., λ_1, λ_2, and λ_3. We search them from $[0.001, 0.005, 0.01, 0.05, 0.1, 0.5]$ throughout all the tasks. Moreover, since the proposed PDAN needs to update the target labels in the optimization, in the training stage we will fix the network parameters and update the target labels at the end of each epoch. In addition, we set the rough class number $C_r = 2$ and divide the original emotions into two rough classes including *High-Valence* (*Happy*, *Surprise*, and *Neutral*) and *Low-Valence* (*Angry*, *Sad*, *Fear*, and *Disgust*).

Finally, since the target label information in cross-corpus SER is entirely unknown, it is not possible to use the validation set to determine the optimal model during the training stage for the transfer learning methods. Therefore, to offer a fair comparison, we follow the tradition of transfer learning method evaluation and report the best results corresponding to the best trade-off parameters for all the methods in the experiments.

3.3. Results and discussions

Experimental results are given in Table 2. From Table 2, several interesting observations can be obtained. First, it can be clearly seen that the proposed PDAN method achieved the best average UAR reaching 42.83% among all the transfer learning methods, which has an increase of 1.06% compared with the second best well-performing method (JDAR + IS10 feature set). Moreover, among all the six cross-corpus SER tasks, our PDAN performs better than all the comparison methods in three others, i.e., E→B, B→C, C→B, respectively. Although the proposed PDAN did not achieve the best performance in the resting three tasks, it can be seen from the comparisons that the results obtained from our method are very competitive against the best-performing comparison methods, e.g., 36.19% (PDAN) v.s. 37.95% (JDAR + IS10 feature set) in task B→E. These observations demonstrated the superiority of the PDAN over recent state-of-the-art transfer subspace learning and deep transfer learning methods in dealing with cross-corpus SER tasks.

TABLE 2 The experimental results of all the transfer learning methods for six cross-corpus SER tasks, in which the best results are highlighted in bold.

Method		B→ E	E→B	B→C	C→B	E→C	C→E	Average
Subspace Learning (IS09 Feature Set)	SVM	28.93	23.58	29.60	35.01	26.10	25.14	28.06
	TCA	30.52	44.03	33.40	45.07	31.10	32.32	36.07
	GFK	32.11	42.48	33.10	48.08	32.80	28.13	36.17
	SA	33.50	43.89	35.80	49.03	32.60	28.17	36.33
	DoSL	36.12	38.95	34.40	45.75	30.40	31.59	36.20
	JDAR	36.33	39.97	31.10	46.29	32.40	31.50	36.27
Subspace Learning (IS10 Feature Set)	SVM	34.50	28.13	35.30	35.29	24.30	26.81	30.73
	TCA	32.60	44.53	40.50	51.47	33.20	29.77	38.68
	GFK	36.01	40.11	40.00	45.93	33.00	29.09	37.35
	SA	35.65	43.92	37.50	47.06	32.10	30.61	37.80
	DoSL	36.82	43.33	36.80	48.45	**35.60**	33.91	39.15
	JDAR	**37.95**	47.80	42.70	48.97	**35.60**	**37.58**	41.76
Deep Learning	AlexNet	29.49	31.03	32.90	42.23	27.59	26.30	31.59
	DAN	36.13	40.41	39.00	49.85	29.00	31.47	37.64
	DANN	33.38	43.68	39.20	53.71	29.80	29.25	38.05
	Deep-CORAL	35.03	43.38	38.30	48.28	31.00	30.89	37.81
	DSAN	36.19	46.90	40.30	50.69	29.70	32.61	39.41
	PDAN (Ours)	36.19	**53.78**	**42.90**	**56.88**	33.70	33.54	**42.83**

Second, by comparing the results obtained by the subspace learning methods with IS09 and IS10 feature sets, it can be found that most methods would achieve better performance when using the IS10 feature set to describe speech signals. For example, JDAR achieved the average UAR of 41.76% when using the IS10 feature set, while its average UAR would decrease to 36.27% if the feature set used to describe speech instead adopted IS09. This may attribute to the limited representation ability of the IS09 feature set compared to IS10. According to the works of Schuller et al. (2009, 2010a), it can be known that the IS10 feature set contains more acoustic LLDs (38) and introduces more statistical functions (21) than IS09 (32 and 12), which leads to a greater capacity of IS10 in describing speech signals. Hence, the transfer subspace learning methods may learn more discriminative representations from the IS10 feature set in coping with cross-corpus SER tasks.

Third, it is also interesting to see that several transfer subspace learning methods using the IS10 feature set, e.g., DoSL and JDAR, outperformed most deep transfer learning ones. This may attribute to the more powerful discriminative ability of the IS10 feature set compared with the features directly learned from the speech spectrums by the deep neural networks. Note that besides the corpus invariant ability, the discriminative one is also an important factor affecting the performance of transfer learning methods, which can be supported by the comparison between the results of IS09 and IS10 feature sets. Consequently, with IS10 as the feature set, several subspace learning methods may achieve better performance than the deep learning ones in coping with the cross-corpus SER tasks.

Last but not least, by deeply comparing the results of all the methods for tasks C→B and B→C and others, it is interesting to see that most methods usually performed better in these two tasks. This may be caused by the difference of emotion-induced methods among these three speech corpora. Specifically, it can be found from the works of Burkhardt et al. (2005), Martin et al. (2006), and Zhang and Jia (2008) that EmoDB and CASIA are both acted speech corpora, while eNTERFACE is an induced one. In other words, the emotional speech samples of EmoDB and CASIA are both acted by the speakers, which are quite different from the ones in eNTEFACE. In eNTERFACE, several stimulus materials were first used to induce the speakers' natural emotions, and then their speech signals were synchronously recorded.

3.4. Ablation study

As Figure 1 and Equation (5) show, the proposed PDAN have a set of progressive distribution adapted regularization terms, which enable the network to learn the corpus invariant features for cross-corpus SER and are different from other deep transfer learning methods, e.g., DAN, DANN, and DSAN. Specifically, the proposed progressive distribution adapted regularization term designed for our PDAN has two major

TABLE 3 Experimental results of PDAN with different total loss functions for six cross-corpus SER tasks, in which the best results are highlighted in bold.

Method	B→E	E→B	B→C	C→B	E→C	C→E	Average
$\mathcal{L}_s + \mathcal{L}_m$	34.36	43.39	37.50	48.89	30.00	30.12	37.38
$\mathcal{L}_s + \mathcal{L}_m + \mathcal{L}_{fc}$	35.16	48.96	41.40	54.96	32.70	32.98	41.03
$\mathcal{L}_s + \mathcal{L}_m + \mathcal{L}_{rc} + \mathcal{L}_{fc}$	**36.19**	**53.78**	**42.90**	**56.88**	**33.70**	**33.54**	**42.83**

advantages. First, besides widely-used marginal and fine class aware conditional distribution adaptions, we also introduce a rough emotion class aware conditional one to benefit the alleviation of feature distribution difference between the source and target speech emotion corpora. Second, these distribution adapted terms are added to regularize different FC layers of CNN to guide the corpus invariant feature learning, which takes full advantage of the hierarchical structure of deep neural networks. It is clear to see that the computation of marginal distribution adapted term does not need the emotion label information, while the two conditional ones are opposite. Moreover, the fine class aware conditional one needs more precise emotion label information of the speech samples compared with the rough one. Consequently, following the fact that the features learned in the deeper layers would have more discriminative ability with respect to the depth of neural network, we propose a progressive regularization method to make full use of these three terms, i.e., adding the marginal one to the first FC layer, the rough conditional one to the second FC layer, and the fine conditional one to the last FC layer, respectively.

To see whether the designed progressive adapted regularization terms are indeed effective, we conduct additional experiments by removing one or two of the rough emotion class aware conditional distribution adapted term $\mathcal{L}_{rc}$ and fine emotion class aware one $\mathcal{L}_{fc}$ to obtain the new total loss function to train the PDAN. The reduced versions of PDAN are denoted by $\mathcal{L}_s + \mathcal{L}_m$ and $\mathcal{L}_s + \mathcal{L}_m + \mathcal{L}_{fc}$, respectively. The experimental results are shown in Table 3. From Table 3, it can be found that the PDAN trained under the guidance of $\mathcal{L}_s + \mathcal{L}_m + \mathcal{L}_{rc} + \mathcal{L}_{fc}$ and $\mathcal{L}_s + \mathcal{L}_m + \mathcal{L}_{fc}$ performed promisingly better than the one associated with $\mathcal{L}_s + \mathcal{L}_m$ in all six cross-corpus SER tasks. This observation indicates that the performance of PDAN introducing the conditional distribution adaptions would be remarkably increased compared with merely using the marginal distribution adaption. Moreover, it can also be seen that the results achieved by PDAN under the guidance of $\mathcal{L}_s + \mathcal{L}_m + \mathcal{L}_{rc} + \mathcal{L}_{fc}$ are better than $\mathcal{L}_s + \mathcal{L}_m + \mathcal{L}_{fc}$, which demonstrates the effectiveness of further introducing the rough conditional distribution adaption and the superiority of the proposed progressive distribution adaptions used in PDAN for dealing with cross-corpus SER tasks.

4. Conclusion

In this paper, we have proposed a novel deep transfer learning method called progressive distribution adapted neural networks (PDAN) to deal with the problem of cross-corpus SER. Unlike existing deep transfer learning methods, PDAN absorbs the knowledge of the emotion wheel and makes full use of the hierarchical structure of deep neural networks. Specifically, we design a progressive distribution adapted regularization term consisting of a marginal distribution adaption and two different types of conditional distribution adaptions to layer-by-layer guide the feature learning of PDAN. Hence, PDAN can learn the emotion discriminative and corpus invariant features for speech signals and be effective to deal with cross-corpus SER tasks. Extensive experiments on three widely-used speech emotion corpora were conducted to evaluate the performance of the proposed PDAN. Experimental results showed that the proposed PDAN can achieve a more satisfactory overall performance than recent state-of-the-art transfer subspace learning and deep transfer learning methods in coping with cross-corpus SER tasks.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: EmoDB, http://emodb.bilderbar.info/start.html, eNTERFACE, http://www.enterface.net/enterface05, and CASIA, http://www.chineseldc.org.

Author contributions

YZ: conceptualization, methodology, and funding acquisition. YZ and HL: writing and original draft preparation. HL and JZ: formal analysis. EF: investigation. CL: resources and data curation. HC and CT: review and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported in part by the Natural National Science Foundation of China (NSFC) under the Grant Nos.

U2003207, 61902064, and 62076195, in part by the Jiangsu Frontier Technology Basic Research Project under the Grant No. BK20192004, in part by the Zhishan Young Scholarship of Southeast University, and in part by the Yangtze River Delta Regional Leading Talents Research Project on Immunization under the Grant No. CSJP005.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abdelwahab, M., and Busso, C. (2018). Domain adversarial for acoustic emotion recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* 26, 2423–2435. doi: 10.1109/TASLP.2018.2867099

Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., and Marchand, M. (2014). Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*. doi: 10.48550/arXiv.1505.07818

Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., and Smola, A. J. (2006). Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* 22, e49-e57. doi: 10.1093/bioinformatics/btl242

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., Weiss, B., et al. (2005). "A database of german emotional speech," in *Proceedings of the 2005 Annual Conference of the International Speech Communication Association (INTERSPEECH)* (Lisbon: ISCA), 1517–1520.

Deng, J., Xu, X., Zhang, Z., Frühholz, S., and Schuller, B. (2017). Universum autoencoder-based domain adaptation for speech emotion recognition. *IEEE Signal Process. Lett.* 24, 500–504. doi: 10.1109/LSP.2017.2672753

Deng, J., Zhang, Z., Eyben, F., and Schuller, B. (2014). Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Process. Lett.* 21, 1068–1072. doi: 10.1109/LSP.2014.2324759

El Ayadi, M., Kamel, M. S., and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognit.* 44, 572–587. doi: 10.1016/j.patcog.2010.09.020

Eyben, F., Wöllmer, M., and Schuller, B. (2010). "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia (MM)* (Florence: ACM), 1459–1462.

Fernando, B., Habrard, A., Sebban, M., and Tuytelaars, T. (2013). "Unsupervised visual domain adaptation using subspace alignment," in *Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV)* (Sydney, NSW: IEEE), 2960–2967.

Gideon, J., McInnis, M. G., and Provost, E. M. (2019). Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG). *IEEE Trans. Affect. Comput.* 12, 1055–1068. doi: 10.1109/TAFFC.2019.2916092

Gong, B., Shi, Y., Sha, F., and Grauman, K. (2012). "Geodesic flow kernel for unsupervised domain adaptation," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Providence, RI: IEEE), 2066–2073.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. (2009). Covariate shift by kernel mean matching. *Dataset Shift Mach. Learn.* 3, 5. doi: 10.7551/mitpress/9780262170055.003.0008

Hassan, A., Damper, R., and Niranjan, M. (2013). On acoustic emotion recognition: compensating for covariate shift. *IEEE Trans. Audio Speech Lang. Process.* 21, 1458–1468. doi: 10.1109/TASL.2013.2255278

Kanamori, T., Hido, S., and Sugiyama, M. (2009). A least-squares approach to direct importance estimation. *J. Mach. Learn. Res.* 10, 1391–1445. doi: 10.5555/1577069.1755831

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS), Vol. 25* (Lake Tahoe, NV).

Kwon, S. (2021). Mlt-dnet: speech emotion recognition using 1d dilated cnn based on multi-learning trick approach. *Expert. Syst. Appl.* 167, 114177. doi: 10.1016/j.eswa.2020.114177

Liu, N., Zong, Y., Zhang, B., Liu, L., Chen, J., Zhao, G., et al. (2018). "Unsupervised cross-corpus speech emotion recognition using domain-adaptive subspace learning," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Calgary, AB: IEEE), 5144–5148.

Long, M., Cao, Y., Wang, J., and Jordan, M. (2015). "Learning transferable features with deep adaptation networks," in *Proceedings of the 2015 International Conference on Machine Learning (ICML)* (Lille), 97–105.

Lu, C., Zong, Y., Zheng, W., Li, Y., Tang, C., and Schuller, B. (2022). Domain invariant feature learning for speaker-independent speech emotion recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* 30, 2217–2230. doi: 10.1109/TASLP.2022.3178232

Martin, O., Kotsia, I., Macq, B., and Pitas, I. (2006). "The enterface'05 audio-visual emotion database," in *Proceedings of the 22nd International Conference on Data Engineering Workshops* (Atlanta, GA: IEEE), 8–8.

Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q. (2010). Domain adaptation via transfer component analysis. *IEEE Trans. Neural Netw.* 22, 199–210. doi: 10.1109/TNN.2010.2091281

Schuller, B., Steidl, S., and Batliner, A. (2009). "The interspeech 2009 emotion challenge," in *Proceedings of the 2009 Annual Conference of the International Speech Communication Association (INTERSPEECH)* (Brighton: ISCA).

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., et al. (2010a). "The interspeech 2010 paralinguistic challenge," in *Proceedings of the 2010 Annual Conference of the International Speech Communication Association (INTERSPEECH)* (Makuhari: ISCA), 2794–2797.

Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., et al. (2010b). Cross-corpus acoustic emotion recognition: variances and strategies. *IEEE Trans. Affect. Comput.* 1, 119–131. doi: 10.1109/T-AFFC.2010.8

Schuller, B. W. (2018). Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM.* 61, 90–99. doi: 10.1145/3129340

Song, P., Zheng, W., Ou, S., Zhang, X., Jin, Y., Liu, J., et al. (2016). Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization. *Speech Commun.* 83, 34–41. doi: 10.1016/j.specom.2016.07.010

Sun, B., and Saenko, K. (2016). "Deep coral: correlation alignment for deep domain adaptation," in *Proceedings of the 2016 European Conference on Computer Vision (ECCV)* (Amsterdam: Springer), 443–450.

Tsuboi, Y., Kashima, H., Hido, S., Bickel, S., and Sugiyama, M. (2009). Direct density ratio estimation for large-scale covariate shift adaptation. *J. Inf. Process.* 17, 138–155. doi: 10.2197/ipsjjip.17.138

Yang, L., Shen, Y., Mao, Y., and Cai, L. (2022). "Hybrid curriculum learning for emotion recognition in conversation," in *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI)* (AAAI).

Zhang, J., Jiang, L., Zong, Y., Zheng, W., and Zhao, L. (2021). "Cross-corpus speech emotion recognition using joint distribution adaptive regression," in *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Toronto, ON: IEEE), 3790–3794.

Zhang, J. T. F. L. M., and Jia, H. (2008). "Design of speech corpus for mandarin text to speech," in *Proceedings of the Blizzard Challenge 2008 Workshop at INTERSPEECH* (Brisbane: ISCA).

Zhang, S., Zhang, S., Huang, T., and Gao, W. (2017). Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Trans. Multimedia* 20, 1576–1590. doi: 10.1109/TMM.2017.2766843

Zhang, S., Zhao, X., and Tian, Q. (2022). Spontaneous speech emotion recognition using multiscale deep convolutional lstm. *IEEE Trans. Affect. Comput*. 13, 680–688. doi: 10.1109/TAFFC.2019.2947464

Zhao, Y., Wang, J., Ye, R., Zong, Y., Zheng, W., and Zhao, L. (2022). "Deep transductive transfer regression network for cross-corpus speech emotion recognition," in *Proceedings of the 2022 Annual Conference of the International Speech Communication Association (INTERSPEECH)* (Incheon: ISCA).

Zhu, Y., Zhuang, F., Wang, J., Ke, G., Chen, J., Bian, J., et al. (2020). Deep subdomain adaptation network for image classification. *IEEE Trans. Neural Netw. Learn. Syst*. 32, 1713–1722. doi: 10.1109/TNNLS.2020.2988928

Zong, Y., Zheng, W., Cui, Z., and Li, Q. (2016). Double sparse learning model for speech emotion recognition. *Electron. Lett*. 52, 1410–1412. doi: 10.1049/el.2016.1211

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 16 September 2022
DOI 10.3389/fnins.2022.985709

Check for updates

OPEN ACCESS

EDITED BY
Xiaopeng Hong,
Harbin Institute of Technology, China

REVIEWED BY
Guojun Dai,
Hangzhou Dianzi University, China
Ling Zhao,
Hebei University, China

*CORRESPONDENCE
Suh-Yeon Dong
sydong@sookmyung.ac.kr

SPECIALTY SECTION
This article was submitted to
Perception Science,
a section of the journal
Frontiers in Neuroscience

RECEIVED 04 July 2022
ACCEPTED 29 August 2022
PUBLISHED 16 September 2022

CITATION
Ji Y and Dong S-Y (2022) Deep learning-based self-induced emotion recognition using EEG.
Front. Neurosci. 16:985709.
doi: 10.3389/fnins.2022.985709

Deep learning-based self-induced emotion recognition using EEG

Yerim Ji and Suh-Yeon Dong*

Department of Information Technology Engineering, Sookmyung Women's University, Seoul, South Korea

Emotion recognition from electroencephalogram (EEG) signals requires accurate and efficient signal processing and feature extraction. Deep learning technology has enabled the automatic extraction of raw EEG signal features that contribute to classifying emotions more accurately. Despite such advances, classification of emotions from EEG signals, especially recorded during recalling specific memories or imagining emotional situations has not yet been investigated. In addition, high-density EEG signal classification using deep neural networks faces challenges, such as high computational complexity, redundant channels, and low accuracy. To address these problems, we evaluate the effects of using a simple channel selection method for classifying self-induced emotions based on deep learning. The experiments demonstrate that selecting key channels based on signal statistics can reduce the computational complexity by 89% without decreasing the classification accuracy. The channel selection method with the highest accuracy was the kurtosis-based method, which achieved accuracies of 79.03% and 79.36% for the valence and arousal scales, respectively. The experimental results show that the proposed framework outperforms conventional methods, even though it uses fewer channels. Our proposed method can be beneficial for the effective use of EEG signals in practical applications.

KEYWORDS

self-induced emotion recognition, high-density EEG, channel selection, deep learning, convolutional neural network

1. Introduction

Emotion plays a crucial role in human decision-making. Hence, recognition of different emotions can effectively improve communication between humans and machines in human-computer interaction (HCI) systems. Human emotions have been recognized using non-physiological signals, such as facial expressions (Ko, 2018), speech (Khalil et al., 2019), and gestures (Noroozi et al., 2018). However, non-physiological signals can be intentionally hidden. In contrast, physiological signals cannot be directly altered because the human body produces them spontaneously. For this reason, many researchers have attempted to identify emotions in physiological signals, such as those detected by electroencephalograms (EEGs), electrocardiograms (ECGs), galvanic skin responses (GSRs), and electromyograms (EMGs) (Wei, 2013; Goshvarpour et al., 2017; Katsigiannis and Ramzan, 2017). In this study, we focus on recognizing emotions using EEG signals.

Previous EEG-based emotion recognition techniques have performed well, but most of them focused on externally induced emotion, using audiovisual materials as emotional stimuli (Koelstra et al., 2011; Soleymani et al., 2011; Zheng and Lu, 2017). This type of method requires subjects to continually pay attention to visual or auditory stimuli. External stimuli may be useful to elicit strong emotions, but because there are individual differences in emotional sensitivity, the selected stimuli may not be suitable for all subjects. Accordingly, some researchers have asked subjects to recall episodic memories or imagine situations associated with certain emotions (Damasio et al., 2000; Onton and Makeig, 2009). This enables the subjects to self-induce emotions based on past experience instead of audiovisual materials determined by researchers in advance. The EEG signals produced by this method are more ecologically valid because they capitalize on individual events that have personal meaning (Salas et al., 2012). However, subjects may lose their concentration when they close their eyes and perform the emotional imagery (EI) task. Therefore, the raw EI signals obtained through this method have a lower amplitude than the signals generated by external stimuli (Iacoviello et al., 2015). This increases the difficulty with which emotions are classified using EEG signals. For this reason, classifying self-induced emotions without using external stimuli remains challenging.

In recent years, deep learning methods have been applied to automatically classify emotions using raw EEG signals without handcrafted features (Craik et al., 2019; Huang et al., 2021). In particular, convolutional neural networks (CNNs) have produced promising results for EEG-based emotion recognition because of their ability to automatically extract robust features (Yang et al., 2018; Hu et al., 2021). However, most CNN-based studies still rely on complex preprocessing techniques, such as the conversion of raw EEG signals into other representations (Kwon et al., 2018; Wang et al., 2020). In this study, we employ a CNN for end-to-end classification, which utilizes raw EEG signals as the input and eliminates the need to perform a complex transformation. Feeding raw EEG signals as input into deep learning models is suitable for analyzing time-series EEG signals (Liang et al., 2021). However, this results in a high computational complexity because of the long training time required when using a large number of EEG channels (Tong et al., 2018). In addition, using all channels, including irrelevant channels, causes the CNN to generate complex features, which decreases the classification accuracy (Wang et al., 2019; Li et al., 2020; Zheng et al., 2021). Consequently, EEG channel selection is advantageous not only for reducing the time required for computation, but also for improving the accuracy.

The most commonly used EEG channel selection methods are the wrapper and filtering methods (Shi et al., 2021). The wrapper method uses recursive techniques to select the optimal subset of all EEG channel combinations (Lal et al., 2004). Wrapper-based methods exhibit superior performance in selecting the optimal channel subset, but they are time-consuming (González et al., 2019) and are prone to overfitting (Alotaiby et al., 2015). Two filtering methods are used to solve this problem. The first involves manually selecting channels related to emotions, and the second automatically selects a subset of channels based on certain standards. For example, many studies have selected EEG channels representing the frontal lobe to capture emotions (Atkinson and Campos, 2016; Thammasan et al., 2016; Xu et al., 2019) because previous results have suggested that the neural activity in the frontal lobe is related to emotional processing. However, manually selecting channels based on previous observations does not necessarily yield better results compared to using all EEG channels. Therefore, this study proposes a statistical method for selecting a smaller number of EEG channels in order to robustly reduce the computational load while simultaneously increasing performance. In this method, the most suitable channels are automatically selected by calculating the EEG signal statistics for each subject before the high-density EEG data are used as input for the CNN.

In summary, we propose a novel framework for deep learning-based systems using high-density EEG data. In this framework, the optimal frequency band is first selected. Then, after applying a channel selection method using the statistical characteristics of the raw EEG signal data, a CNN is utilized for feature extraction and classification. The flow diagram of the proposed system is shown in Figure 1, and the main contributions of this study are as follows: (1) To the best of our knowledge, this is the first work to classify self-induced emotion in EEG signals using a deep learning model and demonstrate the efficiency of statistical channel selection methods using signal amplitudes; (2) Frequency band and channel selection strategies were applied to pre-select the prominent features of low-amplitude EEG signals to improve the classification accuracy. In particular, a signal statistics-based channel selection strategy that used fewer channels reduced the computational complexity of the system and improved the efficiency of the brain-computer interface (BCI) system, and (3) Experiments were conducted on the publicly available "Imagined Emotion Study" dataset (IESD) to evaluate the performance of our deep learning-based method for classifying self-induced emotion.

2. Related work

Many studies have investigated EEG-based emotion recognition, but only a few have classified self-induced emotions using internal EEG signals. For example, Kothe et al. (2013) collected EEG signals of self-induced emotions produced through the recall of experiences associated with 15 different emotions. They used the filter bank common spatial pattern (FBCSP) algorithm to extract temporal-spatial features from 124 channels in the EEG signals and

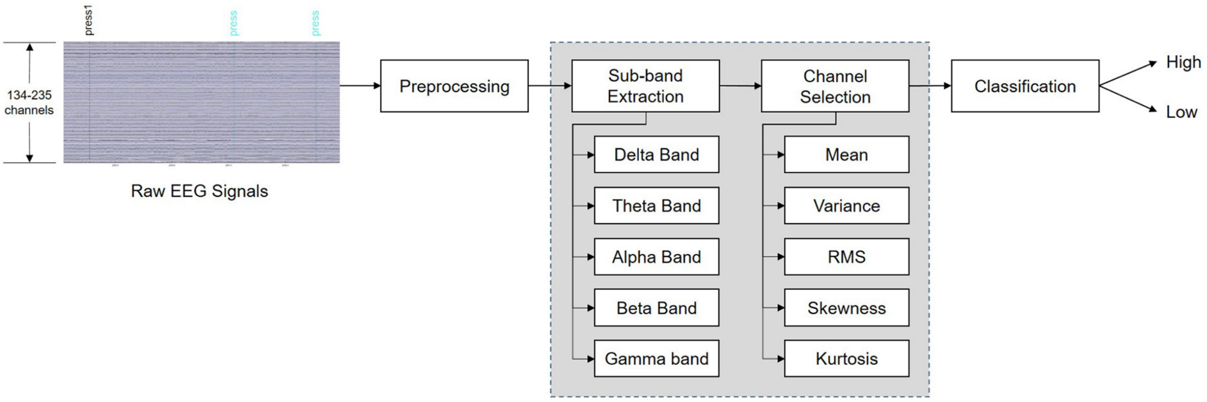


FIGURE 1
Flow diagram of the proposed system for recognition of self-induced emotions.

a linear discriminant analysis (LDA) classifier for valence level recognition. They reported an average accuracy of 71.3%, but excluded three ambiguous emotions (compassion, disgust, and excitement). Similarly, Bigirimana et al. (2020) used common spatial pattern (CSP) features to extract the temporal-spatial-frequency representations. They obtained an accuracy of 80% using LDA for imagery induced by recalling sad and happy events. Iacoviello et al. (2015) proposed an automatic real-time classification method based on a discrete wavelet transform (DWT) that used a support vector machine (SVM). They achieved an accuracy of 90.2% for the emotion of disgust self-induced by remembering an unpleasant odor.

Previous studies on self-induced emotions found that emotion-inducing imagery tasks designed to elicit specific discrete emotions (e.g., disgust) achieved higher performance than other methods; however, emotions do not usually occur in isolation (Mills and D'Mello, 2014). To consider emotions similar to those that occur in real life, more studies are needed to classify complex emotions that are mixed with previously experienced emotions. This can be accomplished by including a variety of emotions in the imagery task. Therefore, in this study, we aimed to recognize various self-induced emotions at the valence and arousal levels. In addition, all existing studies on self-induced emotion are based on machine learning (ML) methods. In contrast to these studies, we propose a deep learning-based system to improve the recognition performance and system efficiency. Deep learning methods outperform traditional ML methods in several fields of research (Craik et al., 2019; Roy et al., 2019), but deep learning techniques have not been fully utilized in the classification of self-induced emotion. To the best of our knowledge, this is the first attempt to detect self-induced emotion in EEG signals using a CNN.

3. Data description

The EEG dataset we used for training and testing was the "Imagined Emotion Study" dataset (IESD) (Onton and Makeig, 2021), which is publicly available on the OpenNeuro.org platform. To the best of our knowledge, this is the only publicly available dataset that contains EEG signals collected for emotion-inducing imagery tasks. In this dataset, all 34 subjects (with ages ranging from 18 to 35 years) listened to 15- to 30-s audio clips that induced an emotional experience, which helped them imagine what they had felt in the past. Next, the subjects performed EI for an average of 3–5 min for each trial. The EI trials consisted of 15 self-paced emotional images that reflected the emotions of anger, awe, compassion, contentedness, disgust, excitement, fear, frustration, grief, happiness, jealousy, joy, love, relief, and sadness. While the subjects imagined the emotional experience, they pressed the "feeling it" button when they felt the suggested emotion strongly enough. Among the 34 subjects, five were excluded from future analysis because they pressed the "feeling it" button only once per emotion or did not press the button at all. The EEG signals for each subject were collected using a 250-channel BioSemi ActiveTwo system (Amsterdam, Netherlands) with a sampling rate of 256 Hz.

4. Preprocessing

4.1. Data processing

The raw EEG signals were preprocessed using MATLAB (R2021a, MathWorks Inc., Natick, MA, USA) and its EEGLAB toolbox (EEGLAB, Boston, MA, USA) (Delorme and Makeig, 2004). Four channels (E3, G23, H25, and H26) were not used in this study at all because they were bad channels for all subjects (the E3 and G23 channels were located in the right

TABLE 1 Number of samples for each class of emotion.

Classification scheme	Class	Number of samples
Valence	Low (Negative)	498
	High (Positive)	636
Arousal	Low (Calm)	489
	High (Active)	645

and left temporal regions, respectively, and the H25 and H26 channels were located in the prefrontal region). Thus, the number of all available channels (C) was 246. Furthermore, the data produced by electrodes with poor skin contact were removed from the recorded signals, leaving 134–235 channels per participant (the number of channels differed for each subject because different selections of bad channels were removed for different subjects). Subsequently, artifacts were eliminated by performing independent component analysis (ICA). After the channel subset for each subject was determined, we interpolated across the channels by applying a spherical spline interpolation (Perrin et al., 1989).

In this study, we only used the periods during which the subjects felt the 15 emotions listed in Section 3. We did this because most of the EI trial period covered neutral states that were not related to emotions (Damasio et al., 2000), and thus including the entire period for training could have led to incorrect classification results. Taking this into account, the continuous EEG signals were preprocessed by excluding periods that did not contain data produced by EI. This generated 2-s segments centered on the moment when the subjects pressed the "feeling it" button. Therefore, the number of segments linked to each subject was the same as the number of times the subject pressed the "feeling it" button; this number ranged from 16 to 149 for each subject. The total number of segments used in our study was 1,134.

4.2. Label processing

Each segment was associated with a label grouped according to the valence and arousal scales, which are the emotional states quantified using Russell's circumplex model (Russell, 1980). Low valence (LV) indicates "negative" emotions (anger, jealousy, disgust, etc.), and high valence (HV) indicates "positive" emotions (love, joy, happiness, etc.). Low arousal (LA) indicates "calm" emotions (sadness, contentedness, grief, etc.), and high arousal (HA) indicates "active" emotions (excitement, fear, anger, etc.). Low and high values were assigned as 0 and 1, respectively. The labeling results are summarized in Table 1. On both the valence and arousal scales, the subjects felt the emotions belonging to the "high" class more easily, which resulted in more samples being generated for that class.

5. Methods

In this section, we present a novel method that quickly recognizes self-induced emotions in EEG signals. It employs a simple technique that selects frequency bands and channels suitable for classification, and therefore it is computationally efficient and suitable for real-time recognition of emotions

5.1. Problem formulation

Let $D_i = (X^1, y^1), \ldots, (X^{N_i}, y^{N_i})$ denote the dataset and N_i denote the number of segments for subject i. Given an EEG input for the k-th segment, X^k, the task is to predict the emotion label y^k corresponding to the k-th segment. The input segment X^k of the network is the tensor $(P \times C \times N_i)$, where P denotes the total number of data points in each segment and C denotes the number of EEG channels. Furthermore, $P = F_s \times T_s$ where F_s denotes the sampling frequency and T_s denotes the duration of the segment. In this context, this study proposes a channel selection method that reduces the number of necessary channels from C (all available channels) to K without compromising performance.

5.2. Frequency band selection

EEG signals are typically categorized according to rhythmic characteristics, resulting in five different sub-bands: delta (δ), theta (θ), alpha (α), beta (β), and gamma (γ). In this study, the EEG signals were band-pass filtered by applying a Butterworth filter to each frequency band. The extracted frequency bands included 1–4 Hz (δ), 4–8 Hz (θ), 8–14 Hz (α), 14–30 Hz (β), 30–50 Hz (γ), and a combination of all these bands. In general, previous EEG-based studies that externally induced emotions using dynamic stimuli, such as video clips, have reported that high-frequency bands are suitable for classifying emotions (Zheng and Lu, 2015; Song et al., 2018; Islam et al., 2021; Rahman et al., 2021). Similarly, the γ band is known to have more of a connection to emotional states than other frequency bands, especially for static stimuli such as images (Li and Lu, 2009; Yang et al., 2020). Accordingly, we hypothesized that the γ rhythm will exhibit a larger difference with different emotions compared to other bands.

However, self-induced emotions evoked by imagining emotional situations in a static environment differ from externally induced emotions. Because the optimal frequency bands for classifying self-induced emotions have not been sufficiently investigated, we investigated all sub-bands in an effort to find a suitable frequency band that maximizes the classification performance.

TABLE 2 EEG signal statistics used for channel selection.

Statistic	Equation
Mean	$\mu(c) = \frac{1}{N}\sum_{i=1}^{N} x_c(i)$
Variance	$V(c) = \frac{1}{N}\sum_{i=1}^{N}(x_c(i) - \mu_c)^2$
Root mean square	$RMS(c) = \sqrt{\frac{\sum_{i=1}^{N} \lvert x_c(i)\rvert^2}{N}}$
Skewness	$SV(c) = \frac{1}{N}\sum_{i=1}^{N}(\frac{x_c(i) - \bar{x_c}}{\sigma})^3$
Kurtosis	$KV(c) = \frac{1}{N}\sum_{i=1}^{N}(\frac{x_c(i) - \bar{x_c}}{\sigma})^4$

In the equations, $x_c(i)$ is the i-th data point of the EEG signal for channel c and N denotes the total number of data points.

5.3. Channel selection

Channel selection removes irrelevant channels; this task simultaneously reduces the calculation complexity and improves the classification accuracy. An automatic channel selection method has not been developed in the field of emotion recognition, and studies in this field are mainly focused on manually selecting channels based on experience (Xu et al., 2019). A simple method for automatically selecting channels is to use the amplitude statistics of EEG signals as a threshold (Alotaiby et al., 2015). This selection criterion is based on the fact that brain activity is most intense when emotional states are being experienced.

To select channels suitable for classifying self-induced emotions, we considered the typical statistics used in the literature, such as the time-domain statistical values (mean, variance, skewness, and kurtosis) and root mean square (RMS), which can be derived from EEG time series. The variance (standard deviation) has been used for channel selection in epileptic seizure (Duun-Henriksen et al., 2012) and motor imagery classification (Azalan et al., 2019). However, appropriate statistics for channel selection in EI classification have not been reported. Therefore, we propose optimal statistics for classifying self-induced emotions based on the experiments we conducted.

Table 2 presents the mathematical formulation of the statistics used in this study. In these equations, $x_c(i)$ is the i-th data point of the EEG signal for channel c and N denotes the total number of data points. The signal statistics were calculated for all channels, and the channels with the highest statistical values were chosen in the channel selection algorithm. Finally, the top K channels with the highest classification accuracies were selected.

5.4. Convolutional neural network

After the frequency bands and channels were selected, a CNN automatically extracted features from both the temporal and spatial dimensions of the raw EEG segments. The CNN architecture used in this study 8was based on the shallow CNN (ShallowConvNet) proposed in Schirrmeister et al. (2017). Due to the shallow architecture of ShallowConvNet, a high accuracy can be achieved without significantly increasing the computational cost (Schirrmeister et al., 2017). The architecture of ShallowConvNet is presented in Table 3.

TABLE 3 Architecture of ShallowConvNet.

Layer	Operation and parameters
L1	40 × Conv(3 × 1), stride(1 × 1)
	40 × Conv(1 × C), stride(1 × 1)
	BatchNorm
	Activation(Square)
	AvgPool(30 × 1), stride(4 × 1)
	Activation(Log)
	Dropout(0.5)
Output	Dense
	Softmax classification

The first convolutional layer was split into two layers, performing temporal and spatial convolutions. This was performed because splitting the first convolutional block is known to yield better results when the number of channels is large (Schirrmeister et al., 2017). Hence, this setup is suitable for extracting features from high-density raw EEG signals. Temporal convolution learns how the amplitude changes over time for all channels of the input segment. Because temporal convolution performs computations for all channels, the volume of computations inevitably depends on the number of channels C. Therefore, C was reduced to K through the channel selection method proposed in Section 5.3. Spatial convolution was used to extract the spatial features of each temporal filter. These steps are similar to the band-pass and common spatial patterns (CSP) spatial filter functions in FBCSP (Ang et al., 2008).

The initial convolutional layer was followed by squaring nonlinearity, an average pooling layer, and a logarithmic activation function. These steps are similar to the trial log-variance computations in FBCSP. In the last output layer, the dense and softmax layers were used for classification.

6. Experimental results

6.1. Implementation details

In this section, we evaluate our proposed method for the task of classifying self-induced emotions in the IESD dataset, using a CNN as the feature extractor and classifier. As mentioned in Section 3, the EEG data for 29 subjects (subject numbers 1–8, 10–21, 23–27, and 29–32) out of a total of 34 were utilized in our experiment. Continuous EEG data were processed into 2-s EEG

TABLE 4 Hyperparameter values of ShallowConvNet.

Hyper-parameter	Value
Optimizer	Adam
Learning rate	0.000625
Batch size	8
Epochs	150 [valence]
	50 [arousal]
Loss function	Negative log likelihood

TABLE 5 Average classification performance for different frequency bands using all channels.

Frequency band	Valence		Arousal	
	Accuracy (%)	F1 (%)	Accuracy (%)	F1 (%)
δ band	62.07	56.33	60.81	56.45
θ band	62.80	57.70	60.32	56.43
α band	64.67	59.62	65.30	61.61
β band	73.39	70.60	71.76	69.16
γ band	**75.97**	**73.28**	**77.68**	**75.54**
All $(\delta, \theta, \alpha, \beta, \gamma)$	72.37	68.93	71.24	68.87

The best results are in bold.

segments, as described in Section 4.1, and fed as input to the CNN for training and testing. For each subject, 80% of the 2-s EEG segments were used for the training set and 20% were used for the test set. The average values from all fold results using five-fold cross-validation were calculated. Next, we experimentally set the appropriate hyperparameters for ShallowConvNet. The optimized hyperparameters used in this study are listed in Table 4. The experiment was performed on a computer with an Intel(R) Core(TM) i7-10700K CPU @ 3.80 GHz 3.79 GHz and NVIDIA GeForce RTX 3080 graphics processing unit (GPU).

6.2. Effect of frequency band on classification performance

In the first set of experiments, the influence of the frequency band on the classification accuracy of the CNN was investigated. Prior to channel selection and feature extraction, all 246 channels were used to find sub-bands suitable for classifying the self-induced emotions. ShallowConvNet was trained separately for the EEG rhythms of the δ, θ, α, β, and γ bands, as well as the entire frequency range of all these sub-bands (1-50 Hz). The average classification results for the 29 subjects on the valence and arousal scales for each sub-band and for all bands using all the channels are shown in Table 5.

Among the five EEG frequency bands, the γ and β bands achieved higher valence and arousal classification results than did the other frequency bands. This result indicates that the higher frequency bands are more closely associated with valence and arousal than the lower frequency bands. The γ band achieved recognition accuracies of 75.97 and 77.68% on the valence and arousal scales, respectively; these were the highest recognition accuracies for each scale. We also considered the F1 score, which is a class-balanced measure of accuracy. Compared to the F1 score of the lowest frequency band (δ), the F1 score of the γ band increased by 16.95% on the valence scale and by 19.09% on the arousal scale. This indicates that the input signals filtered in the γ band (30–50 Hz) improve the precision and recall of the system. In addition, a high average recognition accuracy was achieved for all bands (1–50 Hz). In summary, the CNN performed the best when learning the features in the 30–50 Hz frequency range (the γ band).

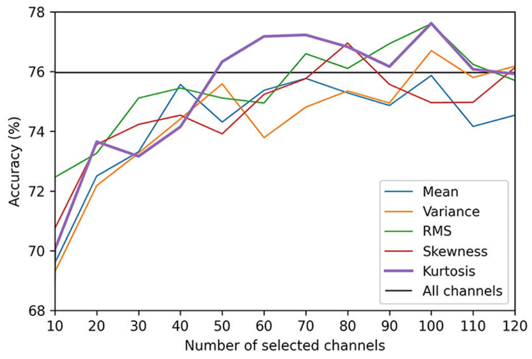


FIGURE 2
Comparison of valence classification accuracies for different EEG channel selection methods.

6.3. Performance comparison of different channel selection methods

Before comparing the results of the channel selection methods, we first evaluated the influence of the number of selected channels (K) on the performance of the CNN. The results produced by varying K from 1 to 123 (half the total number of channels) for the valence and arousal scales are presented in Figures 2, 3, respectively. We did not evaluate the channel selection method using more than 124 channels because, in that case, the channel selection had no significant effect on the results. When K was too small (e.g., $K = 10$), the representation could not be maintained. This led to a decrease in decoding performance, which degraded the accuracy of self-induced emotion recognition. However, when K was too large, similar channels that did not contribute

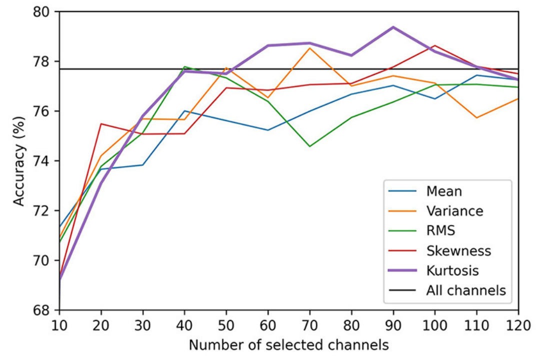


FIGURE 3
Comparison of arousal classification accuracies for different EEG channel selection methods.

to the classification were also included, which limited the representation capacity of the CNN. Moreover, although there was a minimal improvement in performance, the computational cost of the model significantly increased. In Figures 2, 3, the black horizontal line indicates the accuracy that was achieved when all the channels were considered. On both scales, the accuracy of the kurtosis-based channel selection method began to stabilize after 50 channels. Therefore, in order to determine the optimal number of channels, it is necessary to include more than 50 channels.

Table 6 shows the performance of all the channel selection methods for the γ band. For the kurtosis-based channel selection method, the self-induced emotion recognition accuracy reached 79.03% for the valence scale using the top 68 channels and 79.36% for the arousal scale using the top 90 channels. For arousal classification, the skewness-based channel selection method achieved the highest accuracy (but only marginally) using the top 119 channels. Overall, therefore, the kurtosis-based channel selection method performed the best considering the low number of channels it used.

We also compared the performance of each method using the same number of channels (K=64). This number of channels is commonly used in EEG-based emotion recognition studies. On the valence scale, the kurtosis-based method demonstrated a higher performance than the other statistics. On the arousal scale, the skewness-based method demonstrated the highest accuracy, but it was only 0.08% higher than that of the kurtosis-based method. This illustrates how selecting the minimum number of EEG channels that yields the best or required accuracy can balance the performance and computational complexity (Arvaneh et al., 2011). Therefore, although there was a slight difference in accuracy, the kurtosis-based channel selection method exhibited higher accuracy with fewer channels, and thus it is the most suitable channel selection method for self-induced emotion recognition.

TABLE 6 Comparison of the accuracy (%) of different channel selection methods for the γ band.

Classification scheme	Statistic	Maximum accuracy (K)	$K = 64$
Valence	Mean	77.13 (79)	74.23
	Variance	77.28 (108)	75.06
	RMS	78.15 (87)	75.22
	Skewness	76.97 (78)	75.33
	Kurtosis	**79.03** **(68)**	**76.50**
Arousal	Mean	79.01 (122)	75.86
	Variance	78.52 (70)	76.38
	RMS	77.78 (114)	75.33
	Skewness	**79.50** (119)	**76.88**
	Kurtosis	**79.36** **(90)**	**76.80**

K is the number of selected channels. The best results are in bold.

6.4. Effect of frequency band on kurtosis-based channel selection

Figure 4 shows the performance of each frequency band for the kurtosis-based channel selection method. The classification accuracies of the γ band were significantly higher than those of the other frequency bands, regardless of the number of selected channels. In contrast, the classification accuracies of the δ and θ bands were the lowest for the valence and arousal scales, respectively. These results are similar to those obtained using all the channels, as shown in Table 5. This demonstrates that using both the optimal frequency band and optimal channel selection method in our proposed framework improves the EI classification accuracy.

6.5. Effect of computational cost reduction

Table 7 presents a comparison of the overall results of the experiments performed in this study. The table displays the average accuracy and standard deviation of the 29 subjects for the valence and arousal classification tasks in terms of the classification accuracy and execution time. The execution

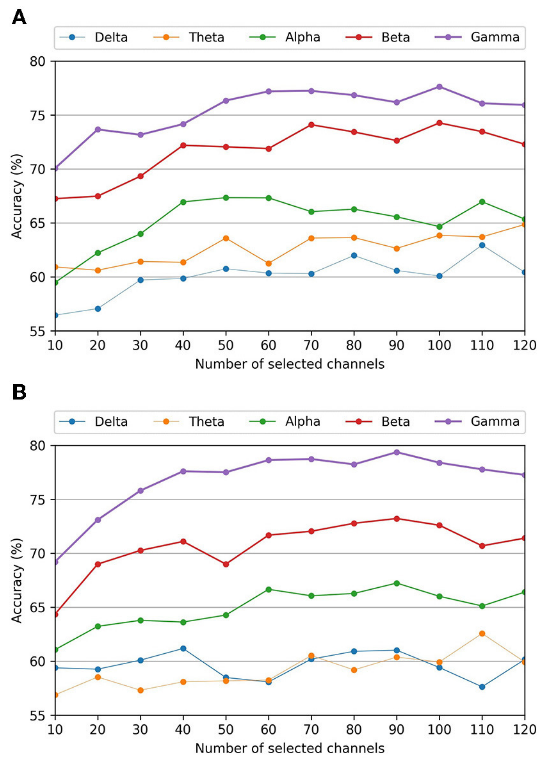


FIGURE 4
Average classification accuracies of different frequency bands for the kurtosis-based channel selection as a function of the number of selected channels. **(A)** Valence scale classification accuracy. **(B)** Arousal scale classification accuracy.

time includes the time required for preprocessing, training, and inference, and it represents the overall computational complexity of the system. According to the results, the feature selection (sub-band and channel selection) process significantly reduced the execution time and improved the accuracy. The BPF and channel selection methods were both effective in improving the performance. In particular, the proposed channel selection method exhibited superior performance in terms of reducing the execution time. Here, the channel selection time is less than 0.1-s and accounts for less than 0.01% of the execution time. This confirms that effective channel selection reduces the training time without compromising the accuracy.

6.6. Channel selection results of different model

The advantage of the proposed method is that it does not overfit a specific model. To validate this fact, we applied kurtosis-based channel selection to DeepConvNet (Schirrmeister et al., 2017), which is widely used as a comparison model for ShallowConvNet. Although the optimal set of channels for ShallowConvNet was used as the input for DeepConvNet, the results produced a 79% reduction in execution time without decreasing the accuracy. Thus, the proposed channel selection method can be expected to further improve the accuracy by determining the optimal number of channels for a given CNN.

6.7. Subject-independent evaluation

We also conducted experiments on subject-independent evaluation to verify the effectiveness of the proposed method. In this experiment, leave-one-out cross-validation was used for evaluation. In each fold, the EEG data of 28 subjects are used for the training, and the remaining 1 subject's EEG data is used for the testing. Since the data of all subjects except the target subject are used for the training, subject-independent channel selection was performed. Table 8 shows the performance of ShallowConvNet on IESD in 10 epochs. The overall accuracy is lower than that of the subject-dependent experiment. However, after applying BPF and channel selection, performance was improved by the proposed framework. This demonstrates that the proposed method can improve performance in both subject-dependent and subject-independent scenarios.

7. Discussion

In this study, we automatically classified self-induced emotions *via* a CNN without using complex preprocessing techniques. We demonstrated that the proposed kurtosis-based channel selection method improved the classification accuracy and significantly reduced the computational complexity. In particular, selecting channels from the γ band maximized the overall classification performance.

High-frequency bands have been widely used to study advanced cognitive functions such as emotions (Yang et al., 2020). As a result of evaluating different frequency bands in this study, we also found that the high-frequency bands contributed more significantly to self-induced emotion classification than did the low-frequency bands. In particular, our results demonstrate that the γ band can identify self-induced emotions more clearly than other bands. However, because CNN-based studies have not been conducted for EI classification before, the classification accuracy achieved in this study by ShallowConvNet for each frequency band can be used as a suggestion for future studies.

Statistical channel selection is a classifier-independent (filtering) method. As mentioned in Section 1, filtering methods do not always find the optimal channel subset or improve performance. Despite this fact, the proposed kurtosis-based channel selection method achieved higher performance using fewer channels. We also applied the proposed method to

TABLE 7 Performance of the proposed framework in terms of average accuracy (%) and execution time.

Deep learning model	Method	Valence		Arousal	
		Accuracy (K)	Execution time (for 150 epochs)	Accuracy (K)	Execution time (for 50 epochs)
ShallowConvNet (Schirrmeister et al., 2017)	Baseline (1-50 Hz)	72.37 ± 15.40 (246)	6 m 60 s	71.24 ± 16.11 (246)	2 m 20 s
	BPF (30–50 Hz)	75.97 ± 16.24 (246)	6 m 50 s	77.68 ± 13.38 (246)	2 m 18 s
	BPF + channel selection (Ours)	**79.03 ± 15.22 (68)**	**28 s**	**79.36 ± 12.33 (90)**	**22 s**
DeepConvNet (Schirrmeister et al., 2017)	Baseline (1-50 Hz)	69.67 ± 16.66 (246)	30 m 31 s	65.93 ± 15.15 (246)	10 m 17 s
	BPF (30–50 Hz)	73.27 ± 17.63 (246)	30 m 02 s	72.89 ± 14.59 (246)	10 m 11 s
	BPF + channel selection (Ours)	**75.29 ± 15.76 (68)**	**5 m 45 s**	**76.10 ± 13.98 (90)**	**2 m 42 s**

BPF stands for "band-pass filter", which indicates the frequency band selection process. The best results are in bold.

another model (DeepConvNet) to verify the advantages of the filtering method. Although we did not use the optimal channel subset as the input for that model, the computational complexity was significantly reduced without compromising the performance. This is the first study to demonstrate the efficiency of statistical channel selection methods using signal amplitudes, which is based on the observation that self-induced emotions have a lower signal amplitude than those induced by external stimuli.

To the best of our knowledge, no previous study has attempted to classify emotions using the same IESD dataset. In a similar study, Hsu et al. (2022) proposed using unsupervised learning approaches to characterize emotional state changes by clustering emotional states in terms of EEG activity differences rather than using subjective labels within the same dataset. Kothe et al. (2013) used the same experimental paradigm that we used, and their binary classification results for the valence scale produced an accuracy of 71.3%. Therefore, our study outperformed this study in that it yielded a valence classification accuracy of 79.03% using all 15 emotions (as opposed to the 12 emotions Kothe et al., 2013 used) and only 68 channels (as opposed to the 124 channels Kothe et al., 2013 used). Moreover, we achieved an accuracy of 79.36% using 90 channels for the arousal scale, which has not been achieved before in previous studies. Furthermore, the FBCSP algorithm used in the previous study is not suitable for deep learning-based systems because it utilizes multiple sub-bands and incurs high computational costs (Kumar et al., 2017). For this reason, the proposed method is effective in that it selects channels based on amplitude statistics without significant computational demands and reduces the overall computational complexity of the system.

TABLE 8 Performance of subject-independent classification using ShallowConvNet.

Method	Valence		Arousal	
	Accuracy	Execution time	Accuracy	Execution time
Baseline (1-50 Hz)	59.95 ± 8.96	4 m 10 s	57.71 ± 8.40	4 m 11 s
BPF (30-50 Hz)	62.67 ± 8.57	4 m 08 s	60.29 ± 8.33	4 m 10 s
BPF + channel selection (Ours)	**63.46 ± 8.34**	**53 s**	**63.75 ± 7.11**	**1 m 28 s**

The best results are in bold.

Like other studies, this study has limitations. Based on the fact that the optimal channel subset varies from individual to individual (Almarri et al., 2021), we performed a subject-specific channel selection, but we did not analyze the selected channels themselves. Therefore, our results did not show the relationship between self-induced emotion and selected channels. Further studies need to be done to investigate the relationship between the channels selected by the kurtosis-based channel selection method and channels that are active in the EI tasks. In addition, this study used only EEG signals collected from 29 subjects in the IESD dataset. Therefore, further work will verify our findings and improve classification accuracy by using larger datasets and data augmentation techniques. Furthermore, fusion with other modalities, such as facial expressions, speech, and ECGs, will be considered to improve the classification accuracy.

8. Conclusion

This paper presented a new deep learning-based framework for self-induced emotion recognition using high-density EEG signals. We proposed a channel selection method based on signal amplitude statistics to improve the performance by removing irrelevant channels, which avoided the large computational load required by high-density EEG signals. The kurtosis-based channel selection method was the most effective method for maximizing the accuracy of self-induced emotion classification. It achieved average classification accuracies of 79.03 and 79.36% for the valence and arousal scales, respectively, using the IESD dataset. We used only 68 channels for valence scale and 90 channels for arousal scale instead of using all 246 channels in the gamma band. This channel selection method reduced the computational complexity of the system by approximately 89% without causing a decrease in accuracy. In addition, we found that selecting channels from only the γ band generated the highest overall classification accuracy. The experimental results demonstrate that appropriate sub-band and channel selection improve the CNN's ability to learn and extract meaningful features. The selected channel combinations were also applied to other models to evaluate the generalization capability of the channel selection method. This analysis shows that our proposed framework can be applied in future CNN-based emotion recognition studies that use high-density EEG signals. The results of this study may contribute to the efficiency and real-time performance of BCI systems.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://doi.org/10.18112/openneuro.ds003004.v1.1.0.

Author contributions

YJ designed the methods, performed the experiments, analyzed the results, and wrote the manuscript. S-YD designed the methods, discussed the results, and extensive revisions to the paper. Both authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (Ministry of Science and ICT, MSIT) (No. NRF-2021R1F1A1052389), by the Commercialization Promotion Agency for R&D Outcomes (COMPA) funded by the Ministry of Science and ICT (MSIT) [Commercialization of health functional foods by verifying the efficacy of functional ingredients and developing the selection method of appropriate content based on AI], and by the MSIT, Korea, under the ICAN (ICT Challenge and Advanced Network of HRD) Program (IITP-2022-RS-2022-00156299) supervised by the IITP (Institute of Information and Communications Technology Planning and Evaluation).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Almarri, B., Rajasekaran, S., and Huang, C.-H. (2021). Automatic subject-specific spatiotemporal feature selection for subject-independent affective BCI. *PLoS ONE* 16, e0253383. doi: 10.1371/journal.pone.0253383

Alotaiby, T., Abd El-Samie, F. E., Alshebeili, S. A., and Ahmad, I. (2015). A review of channel selection algorithms for EEG signal processing. *EURASIP J. Adv. Signal Process.* 2015, 1–21. doi: 10.1186/s13634-015-0251-9

Ang, K. K., Chin, Z. Y., Zhang, H., and Guan, C. (2008). "Filter bank common spatial pattern (FBCSP) in brain-computer interface," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* (Hong Kong: IEEE), 2390–2397.

Arvaneh, M., Guan, C., Ang, K. K., and Quek, C. (2011). Optimizing the channel selection and classification accuracy in EEG-based bci. *IEEE Trans. Biomed Eng.* 58, 1865–1873. doi: 10.1109/TBME.2011.2131142

Atkinson, J., and Campos, D. (2016). Improving bci-based emotion recognition by combining EEG feature selection and kernel classifiers. *Expert Syst. Appl.* 47, 35–41. doi: 10.1016/j.eswa.2015.10.049

Azalan, M. S. Z., Paulraj, M., and Adom, A. H. (2019). "Enhancement of motor imagery brain computer interface performance using channel reduction method based on statistical parameters," in *IOP Conference Series: Materials Science and Engineering, Vol. 557* (Bogor: IOP Publishing), 012016.

Bigirimana, A. D., Siddique, N., and Coyle, D. (2020). Emotion-inducing imagery versus motor imagery for a brain-computer interface. *IEEE Trans. Neural Syst. Rehabil. Eng.* 28, 850–859. doi: 10.1109/TNSRE.2020.2978951

Craik, A., He, Y., and Contreras-Vidal, J. L. (2019). Deep learning for electroencephalogram (EEG) classification tasks: a review. *J. Neural Eng.* 16, 031001. doi: 10.1088/1741-2552/ab0ab5

Damasio, A. R., Grabowski, T. J., Bechara, A., Damasio, H., Ponto, L. L., Parvizi, J., et al. (2000). Subcortical and cortical brain activity during the feeling of self-generated emotions. *Nat. Neurosci* 3, 1049–1056. doi: 10.1038/79871

Delorme, A., and Makeig, S. (2004). Eeglab: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* 134, 9–21. doi: 10.1016/j.jneumeth.2003.10.009

Duun-Henriksen, J., Kjaer, T. W., Madsen, R. E., Remvig, L. S., Thomsen, C. E., and Sorensen, H. B. D. (2012). Channel selection for automatic seizure detection. *Clin. Neurophysiol.* 123, 84–92. doi: 10.1016/j.clinph.2011.06.001

González, J., Ortega, J., Damas, M., Martín-Smith, P., and Gan, J. Q. (2019). A new multi-objective wrapper method for feature selection-accuracy and stability analysis for bci. *Neurocomputing* 333, 407–418. doi: 10.1016/j.neucom.2019.01.017

Goshvarpour, A., Abbasi, A., and Goshvarpour, A. (2017). An accurate emotion recognition system using ecg and gsr signals and matching pursuit method. *Biomed. J.* 40, 355–368. doi: 10.1016/j.bj.2017.11.001

Hsu, S.-H., Lin, Y., Onton, J., Jung, T.-P., and Makeig, S. (2022). Unsupervised learning of brain state dynamics during emotion imagination using high-density EEG. *Neuroimage* 249, 118873. doi: 10.1016/j.neuroimage.2022.118873

Hu, J., Wang, C., Jia, Q., Bu, Q., Sutcliffe, R., and Feng, J. (2021). Scalingnet: extracting features from raw EEG data for emotion recognition. *Neurocomputing* 463, 177–184. doi: 10.1016/j.neucom.2021.08.018

Huang, D., Chen, S., Liu, C., Zheng, L., Tian, Z., and Jiang, D. (2021). Differences first in asymmetric brain: a bi-hemisphere discrepancy convolutional neural network for EEG emotion recognition. *Neurocomputing* 448, 140–151. doi: 10.1016/j.neucom.2021.03.105

Iacoviello, D., Petracca, A., Spezialetti, M., and Placidi, G. (2015). A real-time classification algorithm for EEG-based bci driven by self-induced emotions. *Comput. Methods Programs Biomed.* 122, 293–303. doi: 10.1016/j.cmpb.2015.08.011

Islam, M. R., Islam, M. M., Rahman, M. M., Mondal, C., Singha, S. K., Ahmad, M., et al. (2021). EEG channel correlation based model for emotion recognition. *Comput. Methods Programs Biomed.* 136, 104757. doi: 10.1016/j.compbiomed.2021.104757

Katsigiannis, S., and Ramzan, N. (2017). Dreamer: A database for emotion recognition through EEG and ecg signals from wireless low-cost off-the-shelf devices. *IEEE J. Biomed. Health Inform.* 22, 98–107. doi: 10.1109/JBHI.2017.2688239

Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., and Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: a review. *IEEE Access* 7, 117327–117345. doi: 10.1109/ACCESS.2019.2936124

Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. *Sensors* 18, 401. doi: 10.3390/s18020401

Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., et al. (2011). Deap: A database for emotion analysis; using physiological signals. *IEEE Trans. Affective Comput.* 3, 18–31. doi: 10.1109/T-AFFC.2011.15

Kothe, C. A., Makeig, S., and Onton, J. A. (2013). "Emotion recognition from EEG during self-paced emotional imagery," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* (Geneva: IEEE), 855–858.

Kumar, S., Sharma, A., and Tsunoda, T. (2017). An improved discriminative filter bank selection approach for motor imagery EEG signal classification using mutual information. *BMC Bioinformatics* 18, 125–137. doi: 10.1186/s12859-017-1964-6

Kwon, Y.-H., Shin, S.-B., and Kim, S.-D. (2018). Electroencephalography based fusion two-dimensional (2d)-convolution neural networks (cnn) model for emotion recognition system. *Sensors* 18, 1383. doi: 10.3390/s18051383

Lal, T. N., Schroder, M., Hinterberger, T., Weston, J., Bogdan, M., Birbaumer, N., et al. (2004). Support vector channel selection in bci. *IEEE Trans. Biomed Eng.* 51, 1003–1010. doi: 10.1109/TBME.2004.827827

Li, M., and Lu, B.-L. (2009). "Emotion classification based on gamma-band EEG," in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Minneapolis, MN: IEEE), 1223–1226.

Li, Y., Yang, H., Li, J., Chen, D., and Du, M. (2020). EEG-based intention recognition with deep recurrent-convolution neural network: performance and channel selection by grad-cam. *Neurocomputing* 415, 225–233. doi: 10.1016/j.neucom.2020.07.072

Liang, Z., Zhou, R., Zhang, L., Li, L., Huang, G., Zhang, Z., et al. (2021). Eegfusenet: Hybrid unsupervised deep feature characterization and fusion for high-dimensional EEG with an application to emotion recognition. *IEEE Trans. Neural Syst. Rehabil. Eng.* 29, 1913–1925. doi: 10.1109/TNSRE.2021.3111689

Mills, C., and D'Mello, S. (2014). On the validity of the autobiographical emotional memory task for emotion induction. *PLoS ONE* 9, e95837. doi: 10.1371/journal.pone.0095837

Noroozi, F., Corneanu, C. A., Kamińska, D., Sapiński, T., Escalera, S., and Anbarjafari, G. (2018). Survey on emotional body gesture recognition. *IEEE Trans. Affective Comput.* 12, 505–523. doi: 10.1109/TAFFC.2018.2874986

Onton, J., and Makeig, S. (2021). Imagined emotion study. doi: 10.18112/openneuro.ds003004.v1.1.0

Onton, J. A., and Makeig, S. (2009). High-frequency broadband modulation of electroencephalographic spectra. *Front. Hum. Neurosci.* 3, 61. doi: 10.3389/neuro.09.061.2009

Perrin, F., Pernier, J., Bertrand, O., and Echallier, J. F. (1989). Spherical splines for scalp potential and current density mapping. *Electroencephalogr. Clin. Neurophysiol.* 72, 184–187. doi: 10.1016/0013-4694(89)90180-6

Rahman, M. M., Sarkar, A. K., Hossain, M. A., Hossain, M. S., Islam, M. R., Hossain, M. B., et al. (2021). Recognition of human emotions using EEG signals: a review. *Comput. Methods Programs Biomed.* 136, 104696. doi: 10.1016/j.compbiomed.2021.104696

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., and Faubert, J. (2019). Deep learning-based electroencephalography analysis: a systematic review. *J. Neural Eng.* 16, 051001. doi: 10.1088/1741-2552/ab260c

Russell, J. A. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol.* 39, 1161. doi: 10.1037/h0077714

Salas, C. E., Radovic, D., and Turnbull, O. H. (2012). Inside-out: comparing internally generated and externally generated basic emotions. *Emotion* 12, 568. doi: 10.1037/a0025811

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., et al. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp.* 38, 5391–5420. doi: 10.1002/hbm.23730

Shi, B., Wang, Q., Yin, S., Yue, Z., Huai, Y., and Wang, J. (2021). A binary harmony search algorithm as channel selection method for motor imagery-based bci. *Neurocomputing* 443, 12–25. doi: 10.1016/j.neucom.2021.02.051

Soleymani, M., Lichtenauer, J., Pun, T., and Pantic, M. (2011). A multimodal database for affect recognition and implicit tagging. *IEEE Trans. Affective Comput.* 3, 42–55. doi: 10.1109/T-AFFC.2011.25

Song, T., Zheng, W., Song, P., and Cui, Z. (2018). EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Trans. Affective Comput.* 11, 532–541. doi: 10.1109/TAFFC.2018.2817622

Thammasan, N., Fukui, K.-I., and Numao, M. (2016). "Application of deep belief networks in EEG-based dynamic music-emotion recognition," in *2016 International Joint Conference on Neural Networks (IJCNN)* (Vancouver, BC: IEEE), 881–888.

Tong, L., Zhao, J., and Fu, W. (2018).“ Emotion recognition and channel selection based on EEG signal," in *2018 11th International Conference on Intelligent Computation Technology and Automation (ICICTA)* (Changsha: IEEE), 101–105.

Wang, F., Wu, S., Zhang, W., Xu, Z., Zhang, Y., Wu, C., et al. (2020). Emotion recognition with convolutional neural network and EEG-based efdms. *Neuropsychologia* 146, 107506. doi: 10.1016/j.neuropsychologia.2020.107506

Wang, Z.-M., Hu, S.-Y., and Song, H. (2019). Channel selection method for EEG emotion recognition using normalized mutual information. *IEEE Access* 7, 143303–143311. doi: 10.1109/ACCESS.2019.2944273

Wei, C. Z. (2013). Stress emotion recognition based on rsp and emg signals. *Adv. Mater. Res*. 709, 827–831. doi: 10.4028/www.scientific.net/AMR.709.827

Xu, H., Wang, X., Li, W., Wang, H., and Bi, Q. (2019). "Research on EEG channel selection method for emotion recognition," in *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)* (Dali: IEEE), 2528–2535.

Yang, K., Tong, L., Shu, J., Zhuang, N., Yan, B., and Zeng, Y. (2020). High gamma band EEG closely related to emotion: evidence from functional network. *Front. Hum. Neurosci*. 14, 89. doi: 10.3389/fnhum.2020.00089

Yang, Y., Wu, Q., Qiu, M., Wang, Y., and Chen, X. (2018). "Emotion recognition from multi-channel EEG through parallel convolutional recurrent neural network," in *2018 International Joint Conference on Neural Networks (IJCNN)* (Rio de Janeiro: IEEE), 1–7.

Zheng, W.-L., and Lu, B.-L. (2015). Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Trans. Auton. Mental Dev*. 7, 162–175. doi: 10.1109/TAMD.2015.2431497

Zheng, W.-L., and Lu, B.-L. (2017). A multimodal approach to estimating vigilance using EEG and forehead eog. *J. Neural Eng*. 14, 026017. doi: 10.1088/1741-2552/aa5a98

Zheng, X., Liu, X., Zhang, Y., Cui, L., and Yu, X. (2021). A portable hci system-oriented EEG feature extraction and channel selection for emotion recognition. *Int. J. Intell. Syst*. 36, 152–176. doi: 10.1002/int.22295

TYPE Original Research
PUBLISHED 16 September 2022
DOI 10.3389/fpsyg.2022.959124

OPEN ACCESS

EDITED BY
Xiaopeng Hong,
Harbin Institute of Technology, China

REVIEWED BY
Qi Wu,
Hunan Normal University, China
Wenfeng Chen,
Renmin University of China, China

*CORRESPONDENCE
Su-Jing Wang
wangsujing@psych.ac.cn

SPECIALTY SECTION
This article was submitted to Perception Science, a section of the journal Frontiers in Psychology

RECEIVED 01 July 2022
ACCEPTED 11 August 2022
PUBLISHED 16 September 2022

CITATION
Lin Q, Dong Z, Zheng Q and Wang S-J (2022) The effect of facial attractiveness on micro-expression recognition.
Front. Psychol. 13:959124.
doi: 10.3389/fpsyg.2022.959124

The effect of facial attractiveness on micro-expression recognition

Qiongsi Lin[1,2], Zizhao Dong[1,2], Qiuqiang Zheng[3] and Su-Jing Wang[1,2]*

[1]Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing, China, [2]Department of Psychology, University of the Chinese Academy of Sciences, Beijing, China, [3]Teacher Education Curriculum Center, School of Educational Science, Huizhou University, Huizhou, China

Micro-expression (ME) is an extremely quick and uncontrollable facial movement that lasts for 40–200 ms and reveals thoughts and feelings that an individual attempts to cover up. Though much more difficult to detect and recognize, ME recognition is similar to macro-expression recognition in that it is influenced by facial features. Previous studies suggested that facial attractiveness could influence facial expression recognition processing. However, it remains unclear whether facial attractiveness could also influence ME recognition. Addressing this issue, this study tested 38 participants with two ME recognition tasks in a static condition or dynamically. Three different MEs (positive, neutral, and negative) at two attractiveness levels (attractive, unattractive). The results showed that participants recognized MEs on attractive faces much quicker than on unattractive ones, and there was a significant interaction between ME and facial attractiveness. Furthermore, attractive happy faces were recognized faster in both the static and the dynamic conditions, highlighting the happiness superiority effect. Therefore, our results provided the first evidence that facial attractiveness could influence ME recognition in a static condition or dynamically.

KEYWORDS

facial attractiveness, micro-expression, micro-expression recognition, emotion recognition, happy-face-advantage

1. Introduction

Micro-expression (ME) is an instinctive facial movement that expresses emotion and cognition. It is difficult for individuals to identify MEs since they are rapid (usually lasting for 40–200 ms), local, low-intensity facial responses (Liang et al., 2013). On the contrary, macro-expression is easily identifiable and lasts between 500 ms and 4 s (Takalkar et al., 2021). Ekman and Friesen (1969) indicated that the only difference between ME and macro-expression is their duration. According to Shen et al. (2012), the duration of the expressions influences the accuracy of ME recognition, the proper upper limit of duration of ME may be 200 ms or less. Shen et al. (2016) utilized electroencephalogram (EEG) and event-related potentials (ERPs) and found that the EEG/ERPs neural mechanisms for recognizing MEs differ from those for recognizing macro-expressions. From their findings, the vertex positive potential (VPP) at the electrodes Cz and CPz were significantly different between MEs (duration of less than 200 ms) and macro-expressions (duration of greater than 200 ms), and the VPP amplitude of negative expression was larger than that of positive and neutral expression with the duration of less than 200 ms, while when the duration was greater than 200 ms, there was no difference in VPP amplitude induced by different emotional expressions.Previous studies discovered that emotional contexts influence ME processing at an early stage. Zhang et al. (2018) found that early ERP differences in emotional contexts on ME processing, more positive P1 (an early component related to the visual processing of faces, peaking at approximately 100 ms) and N170 (peaking at around 160 ms) elicited by targeting ME followed negative and positive contexts rather than neutral contexts. Previous functional magnetic resonance imaging (fMRI) research found that emotional contexts reduce the accuracy of ME recognition while increasing context-related activation in some emotional and attentional regions (Zhang et al., 2020). Due to the additional monitoring and attention required for emotional context inhibition, the increased perceptual load of negative and positive contexts results in increased brain activation as well as decreased behavioral performance (Siciliano et al., 2017). Studies of emotion perception have demonstrated that ME recognition is similar to macro-expression recognition and that it is affected by variety of factors, such as gender (Abbruzzese et al., 2019), age (Abbruzzese et al., 2019), occupation (Hurley, 2012), culture (Iria et al., 2019), and individual psychological characteristics (Zhang et al., 2017). ME recognition is widely used in the fields of national security, judicial interrogation, and clinical fields as an effective clue for detecting deceptions (Ekman, 2009), as MEs occurred too quickly and are very difficult to detect, scholars have long endeavored to explore and improve individuals' ability to recognize MEs. Previous studies have typically focused on how facial attractiveness moderates macro-expression recognition. To the best of our knowledge, no previous study on macro-expressions has employed facial expressions of 200 ms or less as their stimuli, it remains unclear whether the durations of facial expressions are able to modulate the effects of facial attractiveness on facial emotion recognition (FER).

Facial attractiveness is the extent to which a face makes an individual feel good and happy, and how much it makes them want to get closer to it (Rhodes, 2006). Attractiveness is a strong signal of social interaction, reflecting all facial features (Rhodes, 2006; Li et al., 2019). Attractive faces are commonly connected with good features such as personal attributes (Eagly et al., 1991; Lindeberg et al., 2019) and higher intelligence levels (Jackson et al., 1995; Mertens et al., 2021). Abundant evidence showed that facial attractiveness affects the ability to recognize facial expressions (e.g., Dion et al., 1972; Cunningham, 1986; Otta et al., 1996; Hugenberg and Sczesny, 2006; Krumhuber et al., 2007; Zhang et al., 2016). For example, Lindeberg et al. (2019) asked participants to recognize happy or angry expressions and rate the level of attractiveness of their faces, the results show that attractiveness has a strong influence on emotion perception. According to Lindeberg et al. (2019) facial attractiveness moderates expressions recognition, participants showed the happiness superiority effect for the faces with higher attractiveness levels but not for the unattractive ones, i.e., people tend to recognize happiness faster in attractive faces than in unattractive faces, while there is no such effect in other emotions recognition (i.e., anger, sadness, surprise, Leppänen and Hietanen, 2004). Li et al. (2019) also observed that facial attractiveness moderates the happiness superiority effect, participants could identify the happy expression faster in higher attractive faces, which is consistent with the findings of Lindeberg et al. (2019). Furthermore, in the study by Golle et al. (2014), the authors utilized two-alternative-forced choice paradigms, which required participants to choose one stimulus above the other. The result revealed that facial attractiveness affects happy expression recognition. When happy faces were likewise more attractive, identifying them was easier. Mertens et al. (2021) employ the mood-of-the-crowd task to compare attractive and unattractive crowds. According to the research, participants were more quick and accurate when rating happy crowds. Attractive crowds were perceived as happier than unattractive crowds, that is, people in crowds with unattractive faces were regarded to be in a negative mood, which supports the assumption that attractiveness could moderate emotion perception.

However, a few studies failed to demonstrate that facial attractiveness influences facial emotion recognition (e.g., Jaensch et al., 2014). For example, Taylor and Bryant (2016) asked participants to classify happiness, neutral, or anger emotions at two attractiveness levels (attractive, unattractive), according to the findings of their study, the detection of happiness or anger is not significantly influenced by facial attractiveness. It should be noted that Taylor and Bryant (2016) used anger as the negative expression, however, anger is often mistaken for those other emotions (Taylor and Jose, 2014), which may have contributed to the masculinization of attractive female faces that made them seem less attractive (Jaensch et al., 2014) and lead to unreliable results. Thus, this study used disgust expression as experimental material which extends the existing research. Furthermore, previous research on recognizing facial expressions has employed static stimuli, while human faces in real life are not static. As humans utilize dynamic facial expressions in everyday conversation, the ability to accurately recognize dynamic expressions makes more sense (Li et al., 2019). In contrast to static facial expressions, previous studies show that dynamic facial expressions are more ecologically valid and could induce more obvious behavioral responses, such as emotion perception (Recio et al., 2011), emotion elicitation (Scherer et al., 2019), and imitation of facial expressions (Sato and Yoshikawa, 2007). This evidence suggests that dynamic stimuli are better identified than static ones, according to face processing literature (Zhang et al., 2015). In this study, we showed participants static and dynamic stimuli to recognize MEs.

To this end, we aimed to explore whether facial attractiveness moderates ME recognition processing. In Experiment 1, static expressions of disgust, neutral, and happiness were presented. Furthermore, Experiment 2 replicated and extended Experiment 1's results by using dynamic stimuli (happy, disgust). We hypothesized that attractive faces could be judged faster overall in a static condition or dynamically; participants could recognize happiness more accurately in attractive faces than in unattractive faces.

2. Experiment 1

We adopted a recognition task modified from the Brief Affect Recognition Test (BART) to simulate a ME (Shen et al., 2012). In the BART paradigm (Ekman and Friesen, 1974), one of the six emotions (happiness, disgust, anger, fear, surprise, and sadness) was presented for 10 ms to 250 ms. In Experiment 1 we presented static stimuli with a duration of 200 ms (happiness as positive ME, disgust as negative ME, and neutral as a control condition) to investigate the effects of facial attractiveness on the processing of MEs. We hypothesized that participants could judge attractive faces faster in static faces, and facial attractiveness moderates the happiness superiority effect, participants could identify the happy expression faster in higher attractive faces but not for the unattractive ones.

2.1. Methods

2.1.1. Participants

The number of participants was similar to or larger than previous research examining the effect of facial attractiveness on expression recognition (e.g., Taylor and Bryant, 2016; Li et al., 2019). Based on a *post hoc* power analysis by using G∗Power 3.1 (Faul et al., 2007) and calculating power analysis for the main effect of ME (a partial η^2 equal to 0.349, an alpha of 0.05, and a total sample size of 38) and attractiveness (a partial η^2 equal to 0.535, an alpha of 0.05, and a total sample size of 38), we observed that this sample size generated a high power of 1-β equal to 0.978 and 0.999 separately. Thus, thirty-eight right-handed participants from Beijing Normal University, Zhuhai (M = 20.24 years, SD = 0.675 years, 20 women) were recruited and received remuneration for completing the experiment. All participants had a normal or corrected-to-normal vision and no psychiatric history. This study adhered to the Declaration of Helsinki and was approved by the Institutional Review Board of the Institute of Psychology, Chinese Academy of Sciences.

2.1.2. Design

Experiment 1 adopted a 3 (ME: happy, neutral, disgust) ×2 (Attractiveness: attractive, unattractive) within-subject factors design. The dependent variables were the participants' mean accuracy score (%) and the mean reaction times (ms) for participants to accurately detect MEs.

2.1.3. Materials

The Extended Cohn-Kanade Dataset (CK+) face database was used to choose images of faces (Lucey et al., 2010). CK+ is the most frequently used laboratory-controlled facial expression classification database that conforms to the Facial Action Coding System (Ekman and Friesen, 1978). At the individual (within-culture) level, Matsumoto et al. (2007) observed consistent and dependable positive connections among the response systems across all seven emotions (happiness, disgust, sadness, contempt, fear, anger, and surprise). These associations indicated that the response systems were coherent with one another. According to Ekman (1992), the response systems for anger, fear, happiness, sadness, and disgust are coherent across cultures which are based not only on a high level of agreement in the labeling of what these expressions signal across literate and preliterate cultures, but also on studies of the actual expression of emotions, both deliberately and spontaneously, as well as the association of expressions with social interactive contexts. Therefore, Caucasian faces can be used to measure Chinese college students (Zhang et al., 2017). From the CK+ face database, we picked 120 pictures of 40 different models whose facial expressions included disgust, happiness, and neutral. Twenty-two additional Chinese participants rated each neutral expression's level of attractiveness on a 7-point Likert scale (1 = very unattractive, 7 = very attractive). A paired sample t-test confirmed that the attractive faces (M = 4.18, SD = 0.152) were significantly higher than unattractive faces (M = 2.23, SD = 0.148), $t_{(4)}$ = 15.764, p < 0.001. The five faces with the highest and lowest average attractiveness ratings were chosen for the research, resulting in a total of 60 trials. In these trials, ten different model faces were used for each emotion: five attractive models representing the three emotions (happiness, neutral, and disgust) and five unattractive models expressing the same emotions. All photos were 350×418 pixels in size and shown on a white background. A Lenovo computer (23.8-inch CRT monitor, resolution 1,920 × 1,080 pixels) and E-Prime (version 2.0) were used to present the stimuli and collect the data.

2.1.4. Procedure

In a quiet environment, participants were tested individually. First, they were given a practice block consisting of nine trials, to begin with, so that they could get familiar with the task. It was requested of the participants that they maintain their gaze on a center fixation cross that was shown on the screen for a duration of 500 ms, then one of the three basic expressions was shown for the duration of 200 ms in the middle of the screen. Participants were told to press the appropriate key according to the micro-expression they considered the face revealed (the "J" key for happy, "K" key for neutral, or the "L" key for disgust) and rate each face on attractiveness using a 7-point Likert scale (1 = very unattractive, 7 = very attractive), each trial only displayed a single image. After 2,000 ms, the reaction screen vanished automatically. The participants were instructed to complete the task in as little time as possible while maintaining the highest level of accuracy. The experimental blocks didn't utilize the practice block's images. Each experimental block included all 30 photographs, one of each face shown twice in random order. Testing took about 15 min (refer to Figure 1).

2.2. Data processing

The average accuracy and mean reaction times for each combination were calculated in both experiments. To deal with the reaction time outliers, we adopted an approach suggested in Ratcliff (1993) and set up a cut-off point of 1.5 SDs above the mean. After that, the reaction time was processed in the same way as the accuracy. We utilized Greenhouse- Geisser correction for heterogeneity of covariances (if sphericity could not be assumed) and Bonferroni correction for *post-hoc* pairwise comparisons. SPSS 26.0 program was used for the data analysis.

2.3. Results and discussion

We launched a 3 × 2 repeated measures ANOVA with ME (happy, neutral, disgust) and Attractiveness (attractive, unattractive) as within-subject factors, and with mean accuracy as dependent variables. The mean accuracy of the three MEs is shown in Figure 2. The results revealed a significant main effect of ME, [$F_{(2,\,74)}$ = 19.823, p < 0.001,η_p^2 = 0.349], a significant main effect of attractiveness, [$F_{(1,\,37)}$ = 42.519, p < 0.001, η_p^2 = 0.535]. The interactions between ME and attractiveness were significant, [$F_{(1.580,\,2.019)}$ = 41.447, p < 0.001, η_p^2 = 0.528]. Pairwise comparisons with Bonferroni correction show that for ME, mean accuracy were significantly higher when responding to happiness compared to disgust (p = 0.011, 95% *CI* [0.024, 0.228]) neutral identified higher recognition accuracy than happiness (p = 0.002, 95% *CI* [0.041, 0.209]), and disgust (p < 0.001, 95% *CI* [0.139, 0.364]). A simple main effect of ME was analyzed to examine the interaction between attractiveness and ME. The results revealed a significant simple main effect

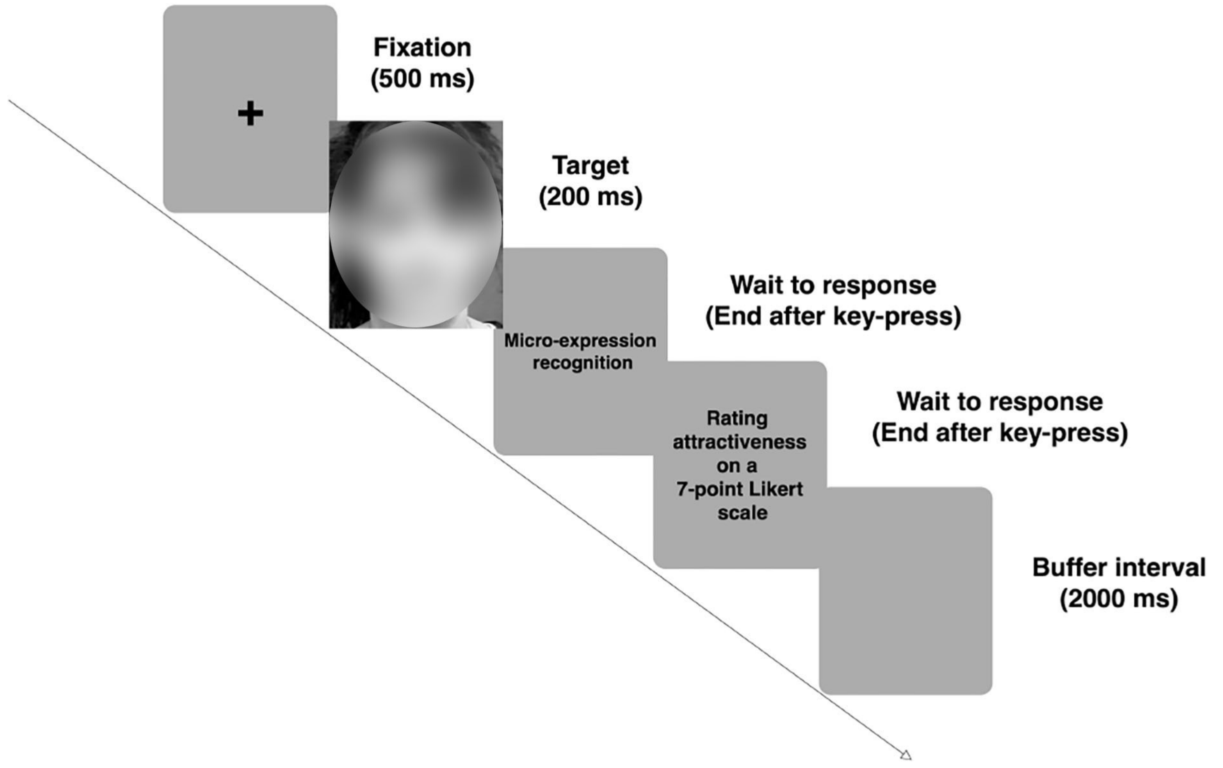


FIGURE 1
The procedure of the micro-expression recognition task and 7-point Likert rating task.

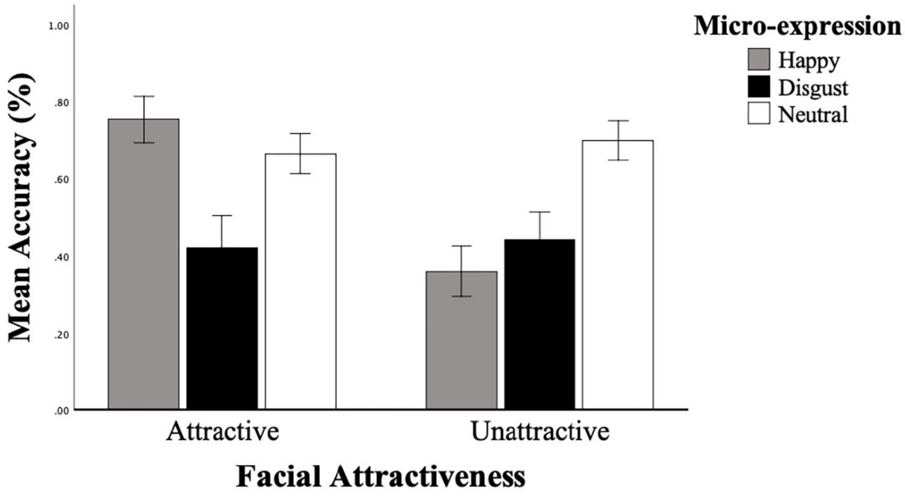


FIGURE 2
Participants' mean accuracy of the static micro-expression recognition task in two facial attractiveness levels (attractive, unattractive). Error bars reflect the 95% CIs for the mean accuracy.

TABLE 1 Mean accuracy of recognition of each Micro-expression in Experience 1.

	Accuracy of recognition (%)	
	Attractive	Unattractive
Micro-expression	M ± SD	M ± SD
Happy	0.775 ± 0.184	0.361 ± 0.199
Disgust	0.421 ± 0.259	0.442 ± 0.223
Neutral	0.665 ± 0.159	0.700 ± 0.156

of ME under the attractive faces condition, [$F_{(2,\,36)} = 27.777$, $p < 0.001$, $\eta_p^2 = 0.607$], and a significant simple main effect of ME under the unattractive faces condition, [$F_{(2,\,36)} = 38.731$, $p < 0.001$, $\eta_p^2 = 0.683$]. Under the attractive faces condition, happiness ($M = 0.755$, $SD = 0.030$) identified higher recognition accuracy than disgust [$M = 0.666$, $SD = 0.026$, $t_{(36)} = 2.34$, $p = 0.023$, $d = 0.780$, 95% *CI* [0.013, 0.166]], and neutral [$M = 0.442$, $SD = 0.036$, $t_{(36)} = 7.45$, $p < 0.001$, $d = 2.48$, 95% *CI* [0.229, 0.397]], disgust identified higher recognition accuracy than neutral [$t_{(36)} = 4.571$, $p < 0.001$, $d = 1.524$, 95% *CI* [0.125, 0.322]]. Furthermore, neutral ($M = 0.700$, $SD = 0.025$) identified higher recognition accuracy than happiness [$M = 0.421$, $SD = 0.042$, $t_{(36)} = 5.167$, $p < 0.001$, $d = 1.722$, 95% *CI* [0.169, 0.389]] and disgust [$M = 0.361$, $SD = 0.032$, $t_{(36)} = 8.692$, $p < 0.001$, $d = 2.897$, 95% *CI* [0.261, 0.418]] under the unattractive faces condition, but no significant differences between happiness and disgust ($p = 0.242$, 95% *CI* [−0.043, 0.164]) (refer to Table 1).

Mean reaction times were submitted to a second repeated measures ANOVA with the same factors described above, outliers (reaction times exceeding the mean of each participant by 1.5 SD) were not included in the analysis. There was no significant main effect of ME, [$F_{(2,\,56)} = 1.661$, $p = 0.199$], and attractiveness, [$F_{(1,\,28)} = 0.453$, $p = 0.507$], no significant interactions between ME and attractiveness, [$F_{(2,\,56)} = 1.363$, $p = 0.264$].

Attractiveness ratings were submitted to a third repeated measures ANOVA with the same factors described above. The results revealed a significant main effect of ME, [$F_{(2,\,74)} = 62.595$, $p < 0.001$, $\eta_p^2 = 0.628$], a significant main effect of attractiveness, [$F_{(1,\,37)} = 64.526$, $p < 0.001$, $\eta_p^2 = 0.636$]. The interactions between ME and attractiveness were significant, [$F_{(2,\,74)} = 7.786$, $p = 0.001$, $\eta_p^2 = 0.174$], indicating that the attractive manipulation of the stimuli used in the current study is effective. Pairwise comparisons with Bonferroni correction show that for ME, the score of attractiveness ratings was significantly higher when responding to happiness compared to disgust ($p <$ 0.001, 95% *CI* [0.500, 0.939]), and neutral ($p < 0.001$, 95% *CI* [0.427, 0.737]), neutral were rated as more attractive than disgust ($p = 0.027$, 95% *CI* [0.013, 0.264]). Further analysis revealed a significant simple main effect of ME under the attractive faces condition, [$F_{(2,\,36)} = 30.378$, $p < 0.001$, $\eta_p^2 = 0.628$], and a significant simple main effect of ME under the unattractive faces condition, [$F_{(2,36)} = 23.264$, $p < 0.001$, $\eta_p^2 = 0.564$]. Under the attractive faces condition, happiness ($M = 4.337$, $SD = 0.164$) were rated with a higher score than disgust [$M = 3.421$, $SD = 0.135$, $t_{(36)} = 7.508$, $p < 0.001$, $d = 2.503$, 95% *CI* [0.668, 1.164]], and neutral [$M = 3.582$, $SD = 0.123$, $t_{(36)} = 7.704$, $p < 0.001$, $d = 2.568$, 95% *CI* [556, 0.954]], disgust were rated with lower score than neutral [$t_{(36)} = 2.439$, $p = 0.020$, $d = 0.813$, 95% *CI* [−0.294, −0.027]]. Under the unattractive faces condition, happiness ($M = 3.361$, $SD = 0.163$) were rated with higher score than disgust [$M = 2.837$, $SD = 0.143$, $t_{(36)} = 6.39$, $p < 0.001$, $d = 2.13$, 95% *CI* [0.358, 0.690]] and neutral [$M = 2.953$, $SD = 0.163$, $t_{(36)} = 5.826$, $p < 0.001$, $d = 1.942$, 95% *CI* [0.266, 0.550]], no significant differences between disgust and neutral [$t_{(36)} = 1.634$, $p = 0.112$, $d = 0.545$, 95% *CI* [−0.260, 0.029]].

In this study, we examine how facial attractiveness influences the processing of ME recognition in static conditions. Analysis of accuracy indicated that the recognition of ME is influenced by attractiveness. Participants categorized attractive faces more accurately than unattractive faces. Specifically, participants showed the happiness superiority effect for the faces with higher attractiveness levels but not for the unattractive ones, the expression of happiness on the attractive faces was the easiest to recognize, followed by neutral, and then disgust.

3. Experiment 2

In Experiment 2, we presented dynamic stimuli to investigate the effects of facial attractiveness on the processing of MEs. We hypothesized that participants could judge attractive faces faster overall in a dynamic context; participants showed the happiness superiority effect for the faces with higher attractiveness levels but not for the unattractive ones.

3.1. Methods

Experiment 2 employed a 2 (ME: happy, disgust) ×2 (Attractiveness: attractive, unattractive) within-subject factors design. The dependent variables were the participants' mean accuracy score (%) and the mean reaction times (ms) for participants to accurately detect MEs. Participants and procedure were the same as in Experiment 1. Based on a *post-hoc* power analysis by using G*Power 3.1 (Faul et al., 2007) and calculating power analysis for the main effect of attractiveness (a partial η^2 equal to 0.436, an alpha of 0.05, and a total sample size of 38), we observed that this sample size generated a high power of 1-β equal to 0.999. To exclude practice effects, we balanced the order of Experiment 1 and Experiment 2 between participants. Thirty-eight participants were randomly divided into two groups (Group A and B), each comprised of 19 participants. Group A completed Experiment 1 follow by Experiment 2, and Group B did the opposite. Also, we used the materials from Experiment 1 to create short video clips. Shen et al. (2012) found a significant difference in recognition accuracy with durations of 40 ms and 120 ms under the METT paradigm condition; however, when the duration was greater than 120 ms, there was no difference in accuracy rate. Thus, we employ the intermediate values with a duration of 80 ms as the target stimulus. Based on the neutral-emotional-neutral paradigm (Zhang et al., 2014), we used neutral as the context expression in this experiment. Zhang et al. (2014) indicated that MEs are contained in the flow of expressions including both neutral and other emotional MEs, considering that a ME is occurred very fast and is always submerged in other MEs, the neutral faces before and after the target ME were presented for 60 ms in order to simulate the real situation in which the ME happened, with happiness or disgust flashed briefly for 80 ms, resulting in a total of 200 ms. Thus, the dynamic stimuli consisted of 20 clips (each clip lasting for 200 ms and showing the same model), comprised of two levels of Attractiveness (attractive and unattractive) and presented as two stimulus types (neutral-happiness-neutral and neutral-disgust-neutral) for each of the 10 models, each clip was shown twice in random order. E-Prime (version 3.0) was used to show the stimuli and collect the data.

3.2. Results and discussion

We launched a 2×2 repeated measures ANOVA with ME (happy, disgust) and Attractiveness (attractive, unattractive) as within-subject factors, and with mean accuracy as dependent variables. The mean accuracy of the two MEs is shown in Figure 3. The results revealed a significant main effect of attractiveness, [$F_{(1,\,37)} = 28.560$, $p < 0.001$, $\eta_p^2 = 0.436$]. The main effect of ME was not significant, [$F_{(1,\,37)} = 0.062$, $p = 0.805$]. The interactions between ME and attractiveness were significant, [$F_{(1,\,37)} = 14.637$, $p < 0.001$, $\eta_p^2 = 0.283$]. A simple main effect of ME was analyzed to examine the interaction between attractiveness and ME. The results revealed a significant simple main effect of ME under the attractive faces condition, [$F_{(1,\,37)} = 5.512$, $p = 0.024$, $\eta_p^2 = 0.130$], and a significant simple main effect of ME under the unattractive faces condition, [$F_{(1,\,37)} = 9.294$, $p = 0.004$, $\eta_p^2 = 0.201$]. Furthermore, happiness ($M = 0.942$, $SD = 0.022$) identified higher recognition accuracy than disgust [$M = 0.732$, $SD = 0.036$, $t_{(37)} = 2.362$, $p = 0.024$, $d = 0.777$, 95% *CI* [0.015, 0.206]] under the attractive faces condition, happiness ($M = 0.832$, $SD = 0.040$) identified lower recognition accuracy than disgust [$M = 0.858$, $SD = 0.021$, $t_{(37)} = 3.073$, $p = 0.004$, $d = 1.010$, 95% *CI* [−0.210, −0.042]] under the unattractive faces condition (refer to Table 2).

Mean reaction times were submitted to a second repeated measures ANOVA with the same factors described above, outliers (reaction times exceeding the mean of each participant by 1.5 SD) were not included in the analysis. There was no significant main effect of ME, [$F_{(1,35)} = 0.218$, $p = 0.644$], or a significant main effect of attractiveness, [$F_{(1,35)} = 2.492$, $p = 0.123$]. Remarkably, the interaction of ME × Attractiveness was significant, [$F_{(1,35)} = 21.245$, $p < 0.001$, $\eta_p^2 = 0.378$]. A follow-up simple effect analysis was employed to investigate the effect of ME within each level of attractiveness.

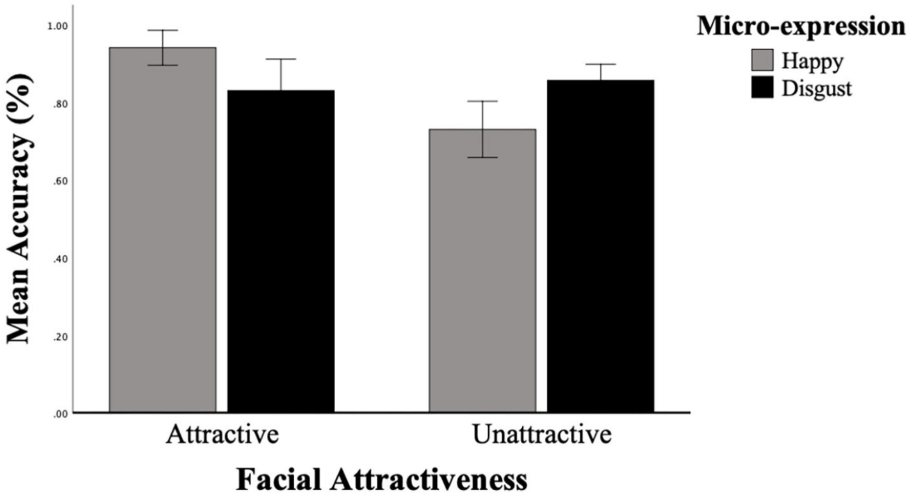


FIGURE 3
Participants' mean accuracy of the dynamic micro-expression recognition task in two facial attractiveness levels (attractive, unattractive). Error bars reflect the 95% CIs for the mean accuracy.

TABLE 2 Mean accuracy of recognition of each Micro-expression in Experience 2.

	Accuracy of recognition (%)	
	Attractive	Unattractive
Micro-expression	M ± SD	M ± SD
Happy	0.942 ± 0.136	0.731 ± 0.221
Disgust	0.731 ± 0.221	0.857 ± 0.127

The results revealed a significant simple main effect of ME under the attractive faces condition, [$F_{(1,37)} = 9.267$, $p = 0.004$, $\eta_p^2 = 0.200$], and a significant simple main effect of ME under the unattractive faces condition, [$F_{(1,37)} = 21.773$, $p < 0.001$, $\eta_p^2 = 0.370$]. Happiness ($M = 758.280$, $SD = 55.873$) identified faster than disgust ($M = 919.013$, $SD = 79.390$) under the attractive faces condition [$t_{(37)} = 3.044$, $p = 0.004$, $d = 1.001$, 95% CI [–267.715, –53.752]], disgust ($M = 821.605$, $SD = 66.602$) identified faster than happiness ($M = 982.400$, $SD = 76.192$) under the unattractive faces condition [$t_{(37)} = 4.666$, $p < 0.001$, $d = 1.534$, 95% CI [–230.616, –90.973]].

Attractiveness ratings were submitted to a third repeated measures ANOVA with the same factors described above. The results revealed a significant main effect of ME, [$F_{(1,\,37)} = 62.947$, $p < 0.001$, $\eta_p^2 = 0.630$], a significant main effect of attractiveness, [$F_{(1,\,37)} = 101.369$, $p < 0.001$, $\eta_p^2 = 0.733$]. The interactions between ME and attractiveness were significant, [$F_{(1,\,37)} = 20.428$, $p < 0.001$, $\eta_p^2 = 0.356$], indicating that the attractive manipulation of the stimuli used in the current study is effective. Further analysis revealed a significant simple main effect of ME under the attractive faces condition, [$F_{(1,\,37)} = 143.607$, $p < 0.001$, $\eta_p^2 = 0.795$], and a significant simple main effect of ME under the unattractive faces condition, [$F_{(1,\,37)} = 29.711$, $p < 0.001$, $\eta_p^2 = 0.445$]. Under the attractive faces condition, happiness ($M = 4.471$, $SD = 0.173$) was rated with a higher score than disgust [$M = 3.195$, $SD = 0.167$, $t_{(37)} = 11.925$, $p < 0.001$, $d = 3.921$, 95% CI [1.061, 1.492]]. Under the unattractive faces condition, happiness ($M = 3.374$, $SD = 0.132$) was rated with a higher score than disgust [$M = 2.682$, $SD = 0.146$, $t_{(37)} = 5.449$, $p < 0.001$, $d = 1.792$, 95% CI [0.435, 0.949]].

In this study, we examine how facial attractiveness influences the processing of ME recognition in dynamic conditions. Analysis of accuracy indicated that attractiveness affects ME recognition. Participants could recognize attractive faces more accurately. Specifically, we observed a higher accuracy rate for happiness than disgust under the attractive faces condition, which supports the assumption that attractiveness could moderate the happiness superiority effect. For the response times, the interaction of Attractiveness × ME was significant, attractive faces were recognized faster than unattractive faces, and happiness was categorized faster than disgust under the attractive face condition whereas this happiness superiority effect did not apply to unattractive faces. According to the results of attractiveness ratings, the advantage of happy faces may be caused by their attractiveness. Overall, participants could identify the happy expression faster and more accurately in higher attractive faces, demonstrating that participants have a stronger ability to identify dynamic expressions that are very attractive.

4. General discussion

Across two experiments, we showed participants static and dynamic faces to recognize MEs. We revealed evidence of the effect of attractiveness on the recognition of ME in either static conditions or dynamically. The results suggest that these two attributes (Attractiveness × ME) are strongly interconnected. Participants showed the happiness superiority effect for the faces with higher attractiveness levels but not for the unattractive ones in both experiments. These findings are in line with the attractiveness stereotype, which defines the phenomena in which individuals correlate physical appearance with a variety of beneficial qualities (Eagly et al. 1991). For instance, attractiveness could boost job interview chances (Watkins and Johnston, 2000). According to the attractiveness stereotype, attractive appearance and good qualities have a strong association with the thoughts of people. Therefore, the identification of attractive faces and positive emotions may be rewarded with an advantage, enhancing their speedy recognition (Golle et al., 2014).

The happiness superiority effect was strengthened by neuroimaging evidence indicating that the medial frontal cortex plays an important role in happy face recognition (Kesler et al., 2001). Ihme et al. (2013) used functional magnetic resonance imaging (fMRI) for the first time to explore the brain mechanism of JACBART and revealed increasing activation with higher performance in the basal ganglia for the negative faces and orbitofrontal areas for happiness and anger. Furthermore, previous research implicated that basal ganglia and orbitofrontal cortex are both involved in the processing of emotional facial expressions. According to O'Doherty et al. (2003), the medial orbitofrontal cortex (OFC) is a region that is known involved in representing stimulus reward value and was shown to be more active when an attractive face was associated with a happy expression, rather than a neutral one. Further studies should find out whether facial attractiveness that correlates with the detection performance of MEs predicts activation in basal ganglia and orbitofrontal cortex.

In general, this study aimed to explore the effects of facial attractiveness on the processing of MEs in static and dynamic experimental conditions. The findings of our study verified and represent an extension of previous research. On one hand, the results show that participants could identify the happy expression quicker in higher attractive faces, which supports the happiness superiority effect and strengthens this theory with more evidence. On the other hand, this research suggests that the moderation of ME recognition is not limited to invariant facial attributes (such as gender and race) but also applies to variable face features such as facial attractiveness. Furthermore, previous studies suggest that ME recognition training has significant effects on the recognition of MEs (Matsumoto and Hwang, 2011). However, the selection of stimulus material in prior research may not address the variations in the attractiveness of the faces representing the various groups. The current findings demonstrate that facial attractiveness is processed quickly enough to influence ME recognition; hence, facial attractiveness should be considered when selecting faces as stimuli for ME recognition training. Also, since individuals can be trained to recognize MEs more accurately and quickly in as little as a few hours, the effects of facial attractiveness on ME recognition may be reduced when individuals receive ME training.

The present experiments entailed several limitations. First, this research only used two basic expressions as experimental materials. It remains unclear whether facial attractiveness affects other MEs (such as a sadness expression) as much as in our research, a wider range of facial expressions should be examined in future research. Second, we used synthetic MEs in the experiences, while natural MEs may be shorter, asymmetrical, and weaker than synthetic MEs, future research could use natural MEs with more ecological validity as research materials. However, this would require a ME database with a rich sample. Third, we employed the Caucasian faces as experimental materials, which were outgroup members to the participants of the current study. However, evidence from cross-cultural studies suggests that the ME recognition process might differ between the ingroup members and outgroup members. For example, Elfenbein and Ambady (2002) suggested that individuals are more accurate at identifying ingroup emotions since they are more familiar with their own race expressions and faces. Therefore, it may be useful to use a wider variety of face types in future studies to evaluate the ingroup advantage in ME recognition-related facial attractiveness in a context of stimulus equivalence. Finally, since a ME is often embedded in the flow of other MEs, we employed 80 ms for target MEs, and the neutral MEs before and after the emotional MEs were only presented for 60 ms to simulate the actual situation in which the ME occurred. This led to the neutral expressions and target ME being combined and the entire duration was examined. Future studies could employ an ERP experiment to investigate the modulation of early visual processing (e.g., P1 and N170) by using natural MEs in order to investigate the neural mechanism for the effect of facial attractiveness on ME. Moreover, this research only examined the presentation time of MEs at 200 ms. Shen et al. (2012) showed that the accuracy of MEs recognition depends on how long they last and reaches a turning point at 200 ms or maybe even less than 200 ms before leveling off. This suggests that the critical time point that differentiates MEs may be 1/5 of a second. Does facial attractiveness have different effects on ME recognition with longer and shorter presentation times? These questions need to be further explored.

5. Conclusion

In conclusion, the current research provides objective evidence that facial attractiveness influences the processing of MEs. Specifically, we observed that attractive happy faces can be recognized faster and more accurately, emphasizing the happiness superiority effect whether in a static condition or dynamically. Moreover, these new results support the assumption that facial attractiveness could moderate emotion perception. Further studies should employ eye tracker technology to detect visual attention mechanisms in MEs processing that is influenced by facial attractiveness.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Institute of Psychology, Chinese Academy of Sciences, Beijing, China. The patients/participants provided their written informed consent to participate in this study.

Author contributions

QL has contributed the main body of text and the main ideas. ZD has contributed to the construction of the text and refinement of ideas and provided extensive feedback and commentary. QZ was responsible for constructing the partial research framework. S-JW led the project and acquired the funding support. All authors contributed to the article and approved the submitted version.

Funding

This article was supported by grants from the National Natural Science Foundation of China (U19B2032).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abbruzzese, L., Magnani, N., Robertson, I. H., and Mancuso, M. (2019). Age and gender differences in emotion recognition. *Front. Psychol.* 10, 2371. doi: 10.3389/fpsyg.2019.02371

Cunningham, M. R. (1986). Measuring the physical in physical attractiveness: quasi-experiments on the sociobiology of female facial beauty. *J. Pers. Soc. Psychol.* 50, 925. doi: 10.1037/0022-3514.50.5.925

Dion, K., Berscheid, E., and Walster, E. (1972). What is beautiful is good. *J. Pers. Soc. Psychol.* 24, 285. doi: 10.1037/h0033731

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., and Longo, L. C. (1991). What is beautiful is good, but…: A meta-analytic review of research on the physical attractiveness stereotype. *Psychol. Bull.* 110, 109. doi: 10.1037/0033-2909.110.1.109

Ekman, P. (1992). An argument for basic emotions. *Cogn. Emot.* 6, 169–200. doi: 10.1080/02699939208411068

Ekman, P. (2009). *Telling lies: Clues to Deceit in the Marketplace, Politics, and Marriage (Revised Edition).* New York, NY: WW Norton & Company.

Ekman, P., and Friesen, W. V. (1969). Nonverbal behavior and clues to deception. *Psychiatry* 32, 88–106. doi: 10.1080/00332747.1969.11023575

Ekman, P., and Friesen, W. V. (1974). "Nonverbal behavior and psychopathology," in *The Psychology of Depression: Contemporary Theory and Research* (Washington, DC), 3–31.

Ekman, P., and Friesen, W. V. (1978). "Facial action coding system," in *Environmental Psychology and Nonverbal Behavior* (Salt Lake City).

Elfenbein, H. A., and Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychol. Bull.* 128, 203–235. doi: 10.1037/0033-2909.128.2.203

Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 175–191. doi: 10.3758/BF03193146

Golle, J., Mast, F. W., and Lobmaier, J. S. (2014). Something to smile about: The interrelationship between attractiveness and emotional expression. *Cogn. Emot.* 28, 298–310. doi: 10.1080/02699931.2013.817383

Hugenberg, K., and Sczesny, S. (2006). On wonderful women and seeing smiles: social categorization moderates the happy face response latency advantage. *Soc. Cogn.* 24, 516–539. doi: 10.1521/soco.2006.24.5.516

Hurley, C. M. (2012). Do you see what i see? learning to detect micro expressions of emotion. *Mot. Emot.*36, 371–381. doi: 10.1007/s11031-011-9257-2

Ihme, K., Lichev, V., Rosenberg, N., Sacher, J., and Villringer, A. (2013). P 59. Which brain regions are involved in the correct detection of microexpressions? preliminary results from a functional magnetic resonance imaging study. *Clin. Neurophysiol.* 124, e92–e93. doi: 10.1016/j.clinph.2013.04.137

Iria, C., Paixão, R., and Barbosa, F. (2019). Facial expression recognition of portuguese using american data as a reference. *J. Psychol. Res.* 2, 700. doi: 10.30564/jpr.v2i1.700

Jackson, L. A., Hunter, J. E., and Hodge, C. N. (1995). Physical attractiveness and intellectual competence: a meta-analytic review. *Soc. Psychol. Q.* 58, 108–122. doi: 10.2307/2787149

Jaensch, M., van den Hurk, W., Dzhelyova, M., Hahn, A. C., Perrett, D. I., Richards, A., et al. (2014). Don't look back in anger: the rewarding value of a female face is discounted by an angry expression. *J. Exp. Psychol.* 40, 2101. doi: 10.1037/a0038078

Kesler, M. L., Andersen, A. H., Smith, C. D., Avison, M. J., Davis, C. E., Kryscio, R. J., et al. (2001). Neural substrates of facial emotion processing using fmri. *Cogn. Brain Res*. 11, 213–226. doi: 10.1016/S0926-6410(00)00073-2

Krumhuber, E., Manstead, A. S., and Kappas, A. (2007). Temporal aspects of facial displays in person and expression perception: the effects of smile dynamics, head-tilt, and gender. *J. Nonverbal. Behav*. 31, 39–56. doi: 10.1007/s10919-006-0019-x

Leppänen, J. M., and Hietanen, J. K. (2004). Positive facial expressions are recognized faster than negative facial expressions, but why? *Psychol. Res*. 69, 22–29. doi: 10.1007/s00426-003-0157-2

Li, J., He, D., Zhou, L., Zhao, X., Zhao, T., Zhang, W., et al. (2019). The effects of facial attractiveness and familiarity on facial expression recognition. *Front. Psychol.* 10, 2496. doi: 10.3389/fpsyg.2019.02496

Liang, J., Yan, W., Wu, Q., Shen, X., Wang, S., and Fu, X. (2013). Recent advances and future trends in micro-expression research. *Bull. Natl. Natural Sci. Foundat. China.* 27, 75–78. doi: 10.16262/j.cnki.1000-8217.2013.02.003

Lindeberg, S., Craig, B. M., and Lipp, O. V. (2019). You look pretty happy: attractiveness moderates emotion perception. *Emotion* 19, 1070. doi: 10.1037/emo0000513

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). "The extended cohn-kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops* (San Francisco, CA: IEEE), 94–101.

Matsumoto, D., and Hwang, H. S. (2011). Evidence for training the ability to read microexpressions of emotion. *Mot. Emot*. 35, 181–191. doi: 10.1007/s11031-011-9212-2

Matsumoto, D., Nezlek, J. B., and Koopmann, B. (2007). Evidence for universality in phenomenological emotion response system coherence. *Emotion* 7, 57. doi: 10.1037/1528-3542.7.1.57

Mertens, A., Hepp, J., Voss, A., and Hische, A. (2021). Pretty crowds are happy crowds: the influence of attractiveness on mood perception. *Psychol. Res*. 85, 1823–1836. doi: 10.1007/s00426-020-01360-x

O'Doherty, J., Winston, J., Critchley, H., Perrett, D., Burt, D. M., and Dolan, R. J. (2003). Beauty in a smile: the role of medial orbitofrontal cortex in facial attractiveness. *Neuropsychologia* 41, 147–155. doi: 10.1016/S0028-3932(02)00145-8

Otta, E., Abrosio, F. F. E., and Hoshino, R. L. (1996). Reading a smiling face: messages conveyed by various forms of smiling. *Percept. Motor Skills* 82(3_Suppl.), 1111–1121. doi: 10.2466/pms.1996.82.3c.1111

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychol. Bull.* 114, 510. doi: 10.1037/0033-2909.114.3.510

Recio, G., Sommer, W., and Schacht, A. (2011). Electrophysiological correlates of perceiving and evaluating static and dynamic facial emotional expressions. *Brain Res*. 1376, 66–75. doi: 10.1016/j.brainres.2010.12.041

Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annu. Rev. Psychol*. 57, 199–226. doi: 10.1146/annurev.psych.57.102904.190208

Sato, W., and Yoshikawa, S. (2007). Spontaneous facial mimicry in response to dynamic facial expressions. *Cognition* 104, 1–18. doi: 10.1016/j.cognition.2006.05.001

Scherer, K. R., Ellgring, H., Dieckmann, A., Unfried, M., and Mortillaro, M. (2019). Dynamic facial expression of emotion and observer inference. *Front. Psychol*. 10, 508. doi: 10.3389/fpsyg.2019.00508

Shen, X., Wu, Q., Zhao, K., and Fu, X. (2016). Electrophysiological evidence reveals differences between the recognition of microexpressions and macroexpressions. *Front. Psychol*. 7, 1346. doi: 10.3389/fpsyg.2016.01346

Shen, X.-B., Wu, Q., and Fu, X.-L. (2012). Effects of the duration of expressions on the recognition of microexpressions. *J. Zhejiang Univ. Sci. B* 13, 221–230. doi: 10.1631/jzus.B1100063

Siciliano, R. E., Madden, D. J., Tallman, C. W., Boylan, M. A., Kirste, I., Monge, Z. A., et al. (2017). Task difficulty modulates brain activation in the emotional oddball task. *Brain Res*. 1664, 74–86. doi: 10.1016/j.brainres.2017.03.028

Takalkar, M. A., Thuseethan, S., Rajasegarar, S., Chaczko, Z., Xu, M., and Yearwood, J. (2021). Lgattnet: automatic micro-expression detection using dual-stream local and global attentions. *Knowl. Based Syst*. 212, 106566. doi: 10.1016/j.knosys.2020.106566

Taylor, A. J., and Bryant, L. (2016). The effect of facial attractiveness on facial expression identification. *Swiss J. Psychol*. 75, 175–181. doi: 10.1024/1421-0185/a000183

Taylor, A. J. G., and Jose, M. (2014). Physical aggression and facial expression identification. *Europes J. Psychol*. 10, 816. doi: 10.5964/ejop.v10i4.816

Watkins, L. M., and Johnston, L. (2000). Screening job applicants: the impact of physical attractiveness and application quality. *Int. J. Select. Assess*. 8, 76–84. doi: 10.1111/1468-2389.00135

Zhang, J., Li, L. U., Ming, Y., Zhu, C., and Liu, D. (2017). The establishment of ecological microexpressions recognition test(emert): an improvement on jacbart microexpressions recognition test. *Acta Psychol. Sin*. 49, 886. doi: 10.3724/SP.J.1041.2017.00886

Zhang, L. L., Wei, B., and Zhang, Y. (2016). Smile modulates the effect of facial attractiveness: an eye movement study. *Psychol. Explor*. 36, 13.

Zhang, M., Fu, Q., Chen, Y.-H., and Fu, X. (2014). Emotional context influences micro-expression recognition. *PLoS ONE* 9, e95018. doi: 10.1371/journal.pone.0095018

Zhang, M., Fu, Q., Chen, Y.-H., and Fu, X. (2018). Emotional context modulates micro-expression processing as reflected in event-related potentials. *Psych J*. 7, 13–24. doi: 10.1002/pchj.196

Zhang, M., Zhao, K., Qu, F., Li, K., and Fu, X. (2020). Brain activation in contrasts of microexpression following emotional contexts. *Front. Neurosci*. 14, 329. doi: 10.3389/fnins.2020.00329

Zhang, Q., Yin, T., and Ran, G. (2015). Psychological and neural mechanisms for the superiority effect of dynamic facial expressions. *Adv. Psychol. Sci*. 23, 1514. doi: 10.3724/SP.J.1042.2015.01514

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 19 December 2022
DOI 10.3389/fnins.2022.1086380

Check for updates

OPEN ACCESS

EDITED BY
Shiqing Zhang,
Taizhou University, China

REVIEWED BY
Xiaohua Huang,
Nanjing Institute of Technology (NJIT), China
Linyuan Wang,
National Digital Switching System Engineering and Technological Research Centre, China
Ke Lu,
University of Chinese Academy of Sciences (CAS), China

*CORRESPONDENCE
Dongmei Jiang
✉ jiangdm@nwpu.edu.cn

SPECIALTY SECTION
This article was submitted to Perception Science, a section of the journal Frontiers in Neuroscience

RECEIVED 01 November 2022
ACCEPTED 02 December 2022
PUBLISHED 19 December 2022

CITATION
Xia X, Zhao Y and Jiang D (2022) Multimodal interaction enhanced representation learning for video emotion recognition. *Front. Neurosci.* 16:1086380. doi: 10.3389/fnins.2022.1086380

Multimodal interaction enhanced representation learning for video emotion recognition

Xiaohan Xia[1], Yong Zhao[1] and Dongmei Jiang[1,2]*

[1]Shaanxi Key Laboratory on Speech and Image Information Processing, National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science, Northwestern Polytechnical University, Xi'an, Shaanxi, China, [2]Pengcheng Laboratory, Shenzhen, Guangdong, China

Video emotion recognition aims to infer human emotional states from the audio, visual, and text modalities. Previous approaches are centered around designing sophisticated fusion mechanisms, but usually ignore the fact that text contains global semantic information, while speech and face video show more fine-grained temporal dynamics of emotion. From the perspective of cognitive sciences, the process of emotion expression, either through facial expression or speech, is implicitly regulated by high-level semantics. Inspired by this fact, we propose a multimodal interaction enhanced representation learning framework for emotion recognition from face video, where a semantic enhancement module is first designed to guide the audio/visual encoder using the semantic information from text, then the multimodal bottleneck Transformer is adopted to further reinforce the audio and visual representations by modeling the cross-modal dynamic interactions between the two feature sequences. Experimental results on two benchmark emotion databases indicate the superiority of our proposed method. With the semantic enhanced audio and visual features, it outperforms the state-of-the-art models which fuse the features or decisions from the audio, visual and text modalities.

KEYWORDS

emotion recognition, representation learning, cross-modal interaction, cross-attention, semantic enhancement

1. Introduction

Automatic emotion recognition, as the first step to enable machines to have emotional intelligence, has been an active research area for the past two decades. Video emotion recognition (VER) refers to predicting the emotional states of the target person by analyzing information from different cues such as facial actions, acoustic characteristics and spoken language (Rouast et al., 2019; Wang et al., 2022). At the heart of this task is how to effectively learn emotional salient representations from multiple modalities including audio, visual, and text.

Previous works usually extract modality-specific features, such as the word-level embeddings from text (Pennington et al., 2014), and frame-level acoustic features from speech (Degottex et al., 2014) or appearance descriptors from face images (Baltrusaitis et al., 2018), then use various fusion strategies to explore the temporal dependencies among the feature sequences of different modalities. For instance, the bidirectional cross-attention proposed by Tsai et al. (2019) to attend interactions between any two pair-wise feature sequences, was extended by Zheng et al. (2022) to implement interactions between three modalities by connecting the cross-attention modules in series. In He et al. (2021), the time squeeze fusion was proposed to model the time-dependent modality-specific interactions. In these works (Tsai et al., 2019; He et al., 2021; Zheng et al., 2022), the audio, visual, and text modalities were treated as three time-series that play the same role. Several works proposed to first fuse the audio and visual feature sequences into a higher level space, then fuse this bimodal feature sequence with the textual feature sequence (Fu et al., 2022; Zhang et al., 2022). Alternatively, text-centered frameworks were designed to explore the cross-modal interactions between textual and non-textual feature sequences (Han et al., 2021; He and Hu, 2021; Wu et al., 2021). In the works above, the textual features are feature sequences composed of the word-level embeddings. In fact, the whole sentence contains more accurate semantics than the word-level embeddings. Accordingly, the challenge is how to effectively leverage textual emotion information while preserving the high-level global semantics. Facing this challenge, Sun et al. (2020) adopted the pre-trained BERT model (Devlin et al., 2019) to obtain global text embeddings and two long-short term memory (LSTM) models to extract sentence-level audio and visual features independently, then modeled the correlations between the outer-product matrices of text-audio and text-visual features to learn the multimodal representations. In Dai et al. (2020), three LSTMs were used to get the global representations of audio, visual, and text modality, respectively. Meanwhile, a set of emotion embeddings was constructed for each modality, representing the semantic meanings for the emotion categories to be recognized. Specifically, the pre-trained GloVe (Pennington et al., 2014) embeddings of emotion category words (happy, sad, etc) were used as textual emotion embeddings, which were mapped to obtain the audio and visual emotion embeddings, respectively, through two learnable mapping functions. Then, the similarity score between the emotion embeddings and the global representation was calculated for each modality separately, and finally fused to get the emotion prediction. This work leveraged the global semantic information, however, the semantics contained in the emotion category words are less goal-oriented toward the target emotion and the important cross-modal feature interactions are ignored.

In fact, as a complex psychological and physiological phenomenon, emotion can be pre- and post-cognitive: initial emotional responses produce thoughts, which produce affect (Lerner and Keltner, 2000). From this perspective, the process of emotional expression, either through facial expression or the way of speaking, is implicitly regulated by the semantic information. Therefore, in this work, we propose a semantically enhanced module for audio or visual encoders, striving to learn more emotion-relevant features from individual video frames or speech segments with the guidance of high-level semantic information from text.

Additionally, in order to capture the temporal dynamics in audio and video signals, sequential learning is usually performed over the unimodal or concatenated features (Dai et al., 2021; Nguyen et al., 2021). However, such approach lacks information exchanging between the audio and visual sequential features. A classical solution is based on the bidirectional cross-attention between the pair-wise modalities (Tsai et al., 2019). Nevertheless, the redundancy that exists in audio and video signals is ignored, moreover, the bidirectional cross-attention leads to additional computational complexity. In the field of video understanding, the Multimodal Bottleneck Transformer (Nagrani et al., 2021; Liu et al., 2022) was recently proposed for audiovisual fusion with the advantage of condensing relevant unimodal information and meanwhile reducing the computational cost. Inspired by this, we adopt the bottleneck Transformer to reinforce the audio and visual features, by leveraging attention bottlenecks as a bridge to explore the temporal interactions between the two modalities. By doing so, our model can simultaneously consider exchanging complementary information and reducing redundancy during the coordinate representation learning process of audio and visual modalities.

Overall, we propose a representation learning approach for video emotion recognition that achieves dual-enhancement through multimodal interactions. First, the encoders of audio and visual modalities are enhanced by the global semantic information in text. Then, the audio and visual feature sequences are reinforced again with the complementary information of each other. Finally, the attentive decision fusion is performed to obtain the final emotion prediction. The effectiveness of the proposed method is verified by extensive experiments on two widely used emotion datasets, i.e., IEMOCAP (Busso et al., 2008) and CMU-MOSEI (Zadeh and Pu, 2018). In summary, the contributions are summarized as follows:

- We propose a semantic enhancement module for the audio and visual feature encoder to enhance the audio and visual features under the guidance of global semantics from the text modality. The enhanced audio and visual features contain more emotion-relevant information.
- To achieve efficient cross-modal interaction between temporal audio and visual feature sequences, the bottleneck Transformer is adopted as the cross-modal encoder. Specifically, the bottleneck Transformer reinforces audio and visual representations by modeling their dynamic

interactions and meanwhile reducing redundancy in the temporal sequences.

- We conduct extensive experiments on two benchmarks and the results demonstrate the superiority of our proposed method for video emotion recognition.

The remainder of this paper is organized as follows. Section 2 reviews the previous related works on video emotion recognition. Section 3 explains our proposed framework in detail. Section 4 reports the experiment results, followed by the conclusions and future work in Section 5.

2. Related works

2.1. Feature representations for video emotion recognition

Extracting effective feature representations is the first and foremost step in video emotion recognition. By considering the heterogeneity of different modalities in the video, separate models are used to extract unimodal features from the raw data of each modality. For text modality, with the advances in natural language processing, pre-trained models such as Word2Vec (Mikolov et al., 2013) and BERT (Devlin et al., 2019) are commonly used for word embedding. As for audio and visual modalities, various hand-crafted features have been designed based on corresponding domain knowledge, such as acoustic descriptors including prosodic and spectral related parameters (Degottex et al., 2014) and visual features based on facial landmarks, facial action units, etc. (Baltrusaitis et al., 2018). Alternatively, benefiting from the development of deep learning, deep-learned feature representations based on the large-scale pre-trained convolutional neural networks (CNN) such as ResNet (He et al., 2016) and VGGish (Hershey et al., 2017) also have been widely used for emotion recognition (Alisamir and Ringeval, 2021; Li and Deng, 2022). Compared with those hand-crafted features, the pre-trained CNN encoders can extract more powerful visual/audio features. However, the general encoding of versatile CNNs does not consider the speciality of emotion and may further limit the emotional representation ability of extracted deep features.

Recently, Nguyen et al. (2021) proposed a two-stream auto-encoder architecture to learn compact yet representative features from audio and visual raw data individually. Then the learned audio and visual features are concatenated and fed into an LSTM for sequential learning and predicting the dimensional emotion scores. In Hazarika et al. (2020), shared-private representations were learned through two separate encoders by projecting each modality to modality-invariant and -specific subspaces, then a Transformer was used to fuse these features into a joint vector for final prediction. By decoupling the common and specific patterns in audio, visual, and text modalities, the learned shared-private representations were highly effective in reducing the modality gap and contributed to significant gains. Self-supervised representation learning also has been adopted for emotion recognition. For instance, Yu et al. (2021) leveraged self-supervised multi-task learning strategy to learn modality-specific representations. Through joint training the multimodal and uni-modal tasks, this model learned the consistency and difference between different modalities simultaneously.

Our work aims at representation learning enhanced with multimodal interactions. Different from previous work, we leverage the high-level global semantics extracted from text modality to guide the representation learning of audio and visual encoders, and therefore the learned audio/visual features could contain more emotion-related information.

2.2. Multimodal fusion for video emotion recognition

Multimodal fusion is another core challenge for video emotion recognition. Early works usually adopted the traditional feature-level or decision-level fusion methods (Ma et al., 2019; Zhang et al., 2019, 2021b; Sharma and Dhall, 2021). With the rise of attention mechanisms, recent works are mostly focusing on cross-modal interactions to explore more effective fusion strategies.

In Tsai et al. (2019), the powerful Transformer network was introduced to multimodal emotion recognition task, to take its advantage of modeling long-term dependencies across modalities. The authors adopted the Transformer decoder-like module to fuse cross-modal information between any two paired modalities by latently adapting one modality to another. To further mine the cross-modal interactions between two or three modalities simultaneously, Zheng et al. (2022) proposed cascade multi-head attention for full fusion of multimodal features by connecting attention modules in series and regarding different modality features as query for different attention modules.

The above-mentioned works focus on exploring the interactions between different modalities by treating audio, visual, and text modalities equally. Another type of representative works argues that text plays a more important role than audio and visual modalities and designs diverse text-centered frameworks for multimodal emotion recognition. In Han et al. (2021), the authors proposed a Transformer-based bi-bimodal fusion network, consisting of two text-related complementing modules, to separately fuse textual feature sequence with audio and visual feature sequences. In Wu et al. (2021), two cross-modal prediction modules, i.e., text-to-visual and text-to-audio models, were designed to decouple the shared and private information of non-textual modalities compared to the textual modality. The shared non-textual information was used to enrich the semantics of textual features and the private non-textual features were later fused with the enhanced textual features through a regression layer for final prediction.

Apart from regarding text as the central modality that plays the most important role among the three modalities, several researchers take into account the difference between audio-visual and text modalities in terms of information granularity. For instance, Fu et al. (2022) proposed a non-homogeneous fusion network by first fusing audio and visual feature sequences through an attention aggregation module and then fusing audio-visual features with textual feature sequence *via* cross-modal attention. Similarly, Zhang et al. (2022) proposed a hierarchical cross-modal encoder module to gradually fuse the modality features. Specifically, an adversarial multimodal refinement module was designed to decompose each modality-specific features to common and private representations. The audio and visual private features were first fused, then this joint audio-visual feature sequence was fused with the textual feature sequence, and finally the fused private features were fused with the common features, resulting in the final joint multimodal representation.

Different from these related works, we are inspired by the emotion expression process that both facial expressions and intonations are implicitly regulated by high-level semantics, and propose a semantic enhancement module to leverage the textual high-level semantics to guide audio and visual representations. In addition, these semantically enhanced audio and visual representations are further reinforced through a multimodal bottleneck Transformer module to exchange their complementary information while reducing redundancy.

3. Proposed method

Figure 1 depicts the architecture of the proposed multimodal emotion recognition (MER) framework with the semantic enhancement module (SEM) and multimodal bottleneck Transformer (MBT), denoted as MER-SEM-MBT. Specifically, we first extract global textual features *via* the textual encoder to represent the high-level semantics, which is used in the SEM to guide the audio/visual encoder to learn emotionally relevant audio/visual features. These semantically enhanced audio and visual feature sequences are sent into the cross-modal encoder to mutually reinforce their representations through cross-modal interaction *via* a bottleneck Transformer. The reinforced audio and visual features are then separately input into a global average pooling (GAP) layer which is followed by a multi-layer perceptron (MLP) to output unimodal decisions. In the meanwhile, the global textual features are fed into another MLP to get the textual decision. Finally, attention-based decision fusion is adopted for the final emotion prediction.

The details are explained in the following subsections.

3.1. Unimodal encoder

For emotion recognition from text, one must analyze the affective state from the complete sentence rather than individual words or phrases. In contrast, regarding the audio and visual modalities, a single video frame or a speech segment longer than 250 ms (Provost, 2013) may contain meaningful emotion information. Therefore, when designing the unimodal encoders, the global semantic features are extracted from the transcripts of the sentences, the audio feature sequence is extracted from the temporal segments, and the visual feature sequence is extracted at the frame level.

3.1.1. Textual encoder

With the advent of Transformer, pre-trained large models such as BERT provided a new paradigm for dynamic text feature encoding based on contextual information with the help of the self-attention mechanism. Therefore, we use the pre-trained BERT model provided in the HuggingFace library (Wolf et al., 2020) as textual encoder. Specifically, the class token ("CLS") of the output layer is adopted as the high-level semantic features $I_t \in \mathbb{R}^{d_t}$, where $d_t = 768$.

3.1.2. Audio encoder

We first calculate the log mel-spectrogram by utilizing 64 Mel filters on the spectrum obtained from the Short-Time Fourier Transform, with a window size of 25 ms and a hop of 10 ms. Then the log mel-spectrogram is split into segments of 960 ms, each of which is fed into the pre-trained VGGish (Hershey et al., 2017) network, outputting a 128-dimensional feature vector from the last fully-connected layer. Therefore, for an audio clip of l s, the audio feature sequence $I_a \in \mathbb{R}^{N_t \times d_a}$ is obtained, with the sequence length $N_t = l/0.96$ and $d_a = 128$.

3.1.3. Visual encoder

The input of visual encoder is a facial image sequence after face alignment. Considering the redundancy between adjacent frames in the face video, we keep consistent with the rate of audio features and randomly sample one frame every 960 ms, forming a face image sequence as input to the visual encoder. For each image, the ResNet18 (He et al., 2016) pre-trained on the AffectNet emotion dataset (Mollahosseini et al., 2017) is adopted as backbone to extract a 512-dimensional spatial feature vector. Correspondingly, for a face video, the visual feature sequence $I_v \in \mathbb{R}^{N_t \times d_v}$ is obtained, with $d_v = 512$.

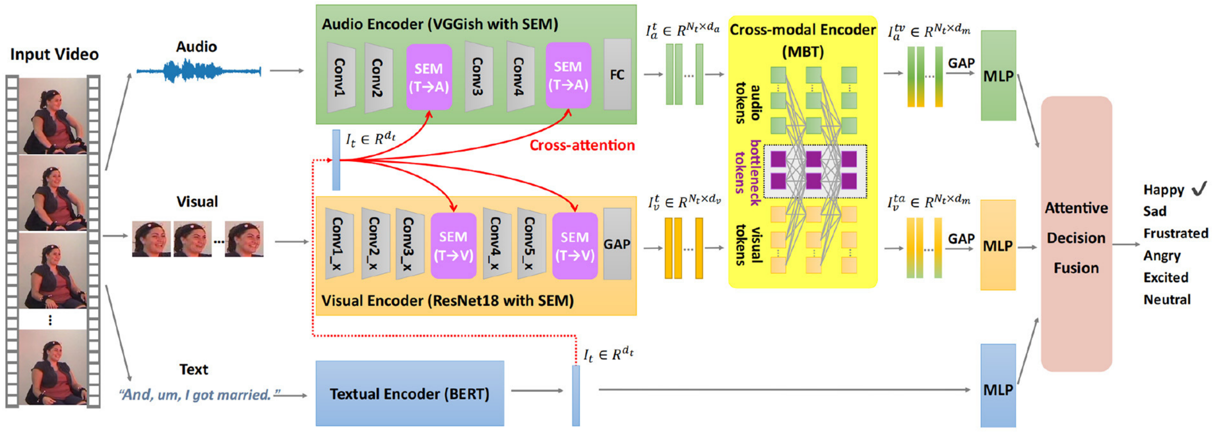


FIGURE 1
The proposed end-to-end multimodal emotion recognition (MER) framework with the semantic enhancement module (SEM) and multimodal bottleneck Transformer (MBT), is denoted as MER-SEM-MBT. Given a facial video clip, the global semantic feature is first extracted through the textual encoder, which is used to guide the audio and visual representation learning through the semantic enhancement module. Then the cross-modal encoder is adopted to reinforce audio and visual representations through temporal cross-modal interaction *via* a multimodal bottleneck Transformer. Lastly, three separate multi-layer perceptrons (MLPs) are implemented to get unimodal decisions from audio, visual, and text modalities, respectively. Attentive fusion is performed to aggregate these decisions for final emotion prediction. The example facial video is from IEMOCAP dataset (Busso et al., 2008).

3.1.4. Semantic enhancement module in audio/visual encoder

In order to guide the audio and visual representation learning, a semantic enhancement module (SEM) is designed to infuse high-level semantic information during audio and visual feature encoding. The implementation of SEM is based on the cross-attention mechanism. As shown in Figure 2, each SEM takes the feature map $F_i \in \mathbb{R}^{C_i \times H_i \times W_i}$ from the middle layer of the audio/visual encoder, as well as the semantic features $I_t \in \mathbb{R}^{d_t}$ from the textual encoder as inputs, then outputs the enriched audio/visual feature map $F_i' \in \mathbb{R}^{C_i \times H_i \times W_i}$ with high-level semantic information. Here, C_i, H_i, and W_i represent the number of channels, the height and width of the feature map after the i^{th} convolution group, respectively.

To retrieve emotion-relevant information from the semantic features to guide audio/visual representation learning, we use the input audio/visual feature map F_i as query Q_f and the input semantic features I_t as key K_t and value V_t during the cross-attention computation, implying a latent adaption from text to audio/visual modality. Formally, the query, key, and value are computed as follows:

$$Q_f = Conv_q(F_i) \in \mathbb{R}^{C_i \times H_i \times W_i}; \quad K_t = Conv_k(I_t) \in \mathbb{R}^{C_i}; \quad V_t = Conv_v(I_t) \in \mathbb{R}^{C_i} \tag{1}$$

where $Conv_q$, $Conv_k$, and $Conv_v$ are projection functions with 1×1 convolution operation. Next we compute the dot products of Q_f with K_t, divided by $\sqrt{C_i}$, and then apply a softmax operator to obtain the weights on V_t. Note that Q_f is first flattened to unroll the spatial dimensions of feature map for proper calculation, yielding $Q_f' \in \mathbb{R}^{C_i \times H_i W_i}$. The output matrix is formulated as:

$$E_{att} = \text{softmax}\left(\frac{{Q_f'}^T K_t}{\sqrt{C}}\right) {V_t}^T \in \mathbb{R}^{H_i W_i \times C_i} \tag{2}$$

Then, the attention map E_{att} is reshaped to the same size of the input audio/visual feature map through the unflatten and transpose operations, yielding $E_{att}' \in \mathbb{R}^{C_i \times H_i \times W_i}$. Finally, the enriched feature map F_i' is output with semantic guided information as follows:

$$F_i' = \text{ReLU}\left(Conv_o\left(F_i + \text{LN}(E_{att}')\right)\right) \in \mathbb{R}^{C_i \times H_i \times W_i} \tag{3}$$

where $Conv_o$ denotes 1×1 convolution operation, LN represents layer normalization (Ba et al., 2016), and ReLU is the nonlinear activation function.

Conventionally, the audio encoder backbone VGGish contains four convolution groups, and the visual backbone ResNet18 contains five convolution groups, as shown in Figure 1. We empirically insert the semantic enhancement module after the second and last convolution group (conv2 and conv4) of VGGish, and the third and last convolution group (conv3_x and conv5_x) of ResNet18, respectively. The effect of the numbers of SEM in audio/visual encoder will be discussed in the Section 4.

Equipped with SEM, the feature sequences output from the audio and visual encoders are enhanced by the high-level semantic information from the text modality, denoted as I_a^t and I_v^t, respectively.

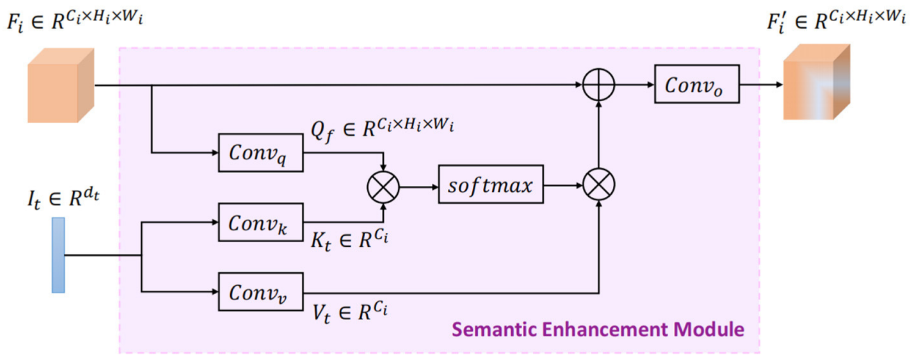

$F_i \in R^{C_i \times H_i \times W_i}$
$F_i' \in R^{C_i \times H_i \times W_i}$
$I_t \in R^{d_t}$
$Conv_q$
$Conv_k$
$Conv_v$
$Conv_o$
$Q_f \in R^{C_i \times H_i \times W_i}$
$K_t \in R^{C_i}$
$V_t \in R^{C_i}$
softmax
Semantic Enhancement Module

FIGURE 2
The semantic enhancement module (SEM) in audio/visual encoder.

3.2. Cross-modal encoder

After obtaining the semantically enhanced audio and visual feature sequences through the above-mentioned unimodal encoders, a cross-modal encoder is required to model the cross-modality relationship between audio and visual modalities. The classical approach is to apply the pair-wise bidirectional cross-attention (Tsai et al., 2019). In the case of considering two modalities (audio and visual), this approach needs four cross-modal Transformer branches, which greatly increases the computational cost. Therefore, we borrow the solution of multimodal bottleneck Transformer (MBT) (Liu et al., 2022) from the field of video understanding, to implement the cross-modal encoder with efficient interactions between audio and visual feature sequences.

As shown in Figure 3, the MBT architecture contains two parallel Transformer branches, serving audio and visual feature sequences for temporal modeling, respectively. The attention bottlenecks are used as the information bridge to exchange complementary information and remove redundancy between audio and visual modalities. Accordingly, the audio and visual feature sequences are mutually reinforced through audio-visual temporal interaction.

Specifically, linear projection is first performed to map the audio/visual features into the identical dimension d_m. Then, a set of bottleneck tokens $\{b_i\}_{i=1}^{N_b}$ are introduced to aggregate audiovisual temporal information. Following Liu et al. (2022), we use the same two-stage cross-modal interaction through feature compression and expansion.

The first interaction stage implies a process of feature compression using a multi-head attention (MHA) layer in the audio and visual Transformer branch, respectively. By treating bottleneck tokens as *query* and audio/visual tokens as $key-value$ pairs, the emotional-relevant multimodal information is condensed into the corresponding audio/visual/bottleneck tokens. Through summing up these three tokens, the multimodal information is aggregated into $\{b_i'\}_{i=1}^{N_b}$.

Subsequently, the second interaction stage is propagating the aggregated multimodal emotional information to the target audio/visual modality through another multi-head attention layer in the audio and visual Transformer branch, respectively. Different from feature compression, the bottleneck tokens are treated as $key - value$ and audio/visual tokens as *query* during this process of feature expansion. Through this two-stage cross-modal attention, audio and visual representations are reinforced with complementary information through interaction with another modality and different time stamps.

Next, the audio and visual features are separately fed into a feed-forward network (FFN) layer to further increase non-linearity, resulting in the reinforced audio and visual feature sequences, denoted as I_a^{tv} and I_v^{ta}, respectively.

3.3. Attentive decision fusion

Finally, the mutually enhanced audio and visual feature sequences are separately input into a global average pooling (GAP) layer and an MLP to obtain unimodal decisions $S_x \in \mathbb{R}^M$, where M represents the number of emotion categories and $x \in \{a, v\}$ represents the audio or visual modality. Meanwhile, the semantic feature vector I_t is input into another MLP to get the textual decision $S_t \in \mathbb{R}^M$.

When fusing these unimodal emotion decisions, we perform attention-based decision fusion to assign higher weights to emotionally salient modality. The unimodal decisions are first concatenated as $S_{con} = [S_a; S_v; S_t] \in \mathbb{R}^{M\times 3}$. Then, the attention weights are calculated as:

$$S' = \tanh\left(W_1 S_{con}\right) \quad (4)$$

$$\alpha_{att} = \text{softmax}\left(W_2^T S'\right) \quad (5)$$

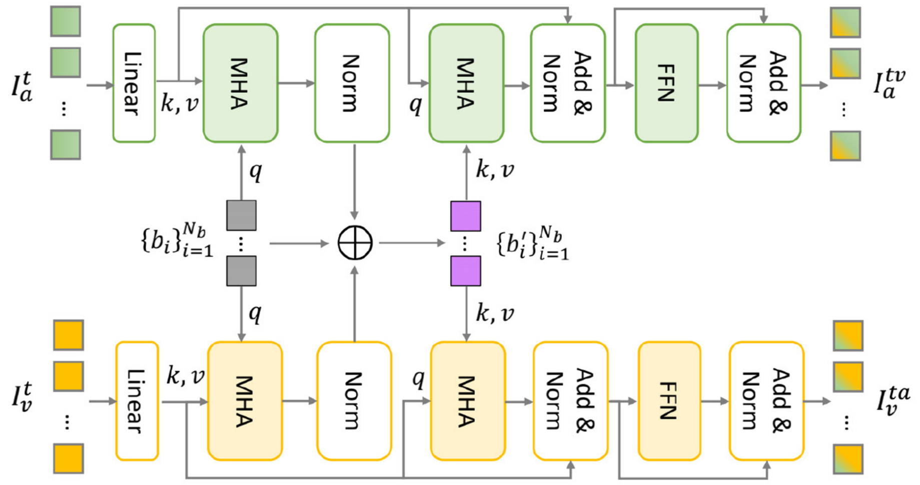


FIGURE 3
The multimodal bottleneck transformer (MBT) architecture (Liu et al., 2022).

where $W_1 \in \mathbb{R}^{M\times M}$ and $W_2 \in \mathbb{R}^{M\times 3}$ are both trainable parameters, and the attention weight $\alpha_{att} \in \mathbb{R}^{1\times 3}$. Finally, the emotion prediction is output after attentive weighted fusion:

$$\text{output} = S_{con}\alpha_{att}^T \tag{6}$$

4. Experiments

4.1. Datasets

To validate the effectiveness of our proposed method, we conduct experiments on two popular video emotion recognition benchmarks, including the Interactive Emotional Dyadic Motion Capture dataset (IEMOCAP) (Busso et al., 2008) and the CMU Multimodal Opinion Sentiment and Emotion Intensity dataset (CMU-MOSEI) (Zadeh and Pu, 2018):

- IEMOCAP consists of 10 performers, five males and five females, who conduct dialogues in pairs to record 151 videos. These videos are segmented into 10,039 utterances and annotated at the utterance level. Six categorical emotions are considered in this work, namely happiness, sadness, angry, frustrated, excited and neutral.
- CMU-MOSEI contains 3,228 video monologs of 1,000 speakers collected from the YouTube website. Annotation of discrete emotion is performed on 23,453 video clips with a total of six emotion categories: anger, disgust, fear, happiness, sadness, and surprise.

For a fair comparison, we use the raw data reorganized by Dai et al. (2021) to implement fully end-to-end training. Specifically, the train/valid/test set of IEMOCAP includes 5,162, 737, and 1,481 samples, respectively, and the train/valid/test split of CMU-MOSEI dataset corresponds to 14,524, 1,765, and 4,188 video clips, respectively. Note that both datasets are multi-labeled at the utterance level and the statistics are shown in Table 1.

4.2. Evaluation metrics

We use the same metrics adopted in Dai et al. (2021): the average binary accuracy (Avg. Acc) and the average F_1 (Avg. F_1) for IEMOCAP, and the average binary weighted accuracy (Avg. WA) and the average F_1 for CMU-MOSEI. These metrics can be formulated as follows:

$$\text{Avg. Acc} = \frac{1}{C}\sum_{i=1}^{C}\text{Acc}_i \tag{7}$$

$$\text{Avg. WA} = \frac{1}{C}\sum_{i=1}^{C}\text{WA}_i \tag{8}$$

$$\text{Avg.}F_1 = \frac{1}{C}\sum_{i=1}^{C}F_{1i} \tag{9}$$

where C is the number of emotion categories, Acc_i, WA_i, and F_{1i} denotes the binary accuracy, binary weighted accuracy and F_1 score of the i^{th} emotion category, respectively:

$$\text{Acc}_i = \frac{TP}{P+N} \tag{10}$$

$$\text{WA}_i = \frac{TP\times N/P + TN}{2N} \tag{11}$$

TABLE 1 Statistics of the IEMOCAP and CMU-MOSEI datasets used in this work.

	IEMOCAP					
	Happiness	Anger	Excited	Frustrated	Sadness	Neutral
Train	398	757	736	1,298	759	1,214
Valid	62	112	92	180	118	173
Test	135	234	213	371	207	321
	CMU-MOSEI					
	Happiness	Anger	Disgust	Surprise	Sadness	Fear
Train	7,587	3,267	2,738	1,465	4,026	1,263
Valid	945	318	273	197	509	169
Test	2,220	1,015	744	393	1,066	371

$$F_{1i} = \frac{2TP}{2TP + FP + FN} \tag{12}$$

In which P and N denote the total number of positive and negative samples, respectively, TP/TN denotes the number of positive/negative samples that are correctly predicted, FP/FN is the number of negative/positive samples that are incorrectly predicted.

Considering the unbalanced distribution of emotion categories, the Avg. F_1 is used as the main evaluation indicator during the training process.

4.3. Implementation details

Data preprocessing: For the input audio, log mel-spectrogram is first calculated by using 64 mel-spaced frequency bins on the spectrum obtained from a short-time Fourier transform applying 25 ms windows every 10 ms. The log mel-spectrogram is divided into non-overlapping 960 ms segments that form the input to the audio encoder. The OpenFace (Baltrusaitis et al., 2018) toolkit is utilized to perform face detection and alignment from original videos. After obtaining the facial image sequence from OpenFace, we consider the redundancy between adjacent frames and randomly sample one frame within every 960ms-long duration for each video, yielding the input to the visual encoder. In addition, this sampling operation enables audio and visual features to be temporally aligned at the video level.

Network parameters: For the audio encoder backbone VGGish, the output feature dimension is $d_a = 128$. The output feature dimension of visual encoder backbone ResNet18 is $d_v = 512$. The pre-trained BERT (*bert-base-uncased*) provided in the HuggingFace library (Wolf et al., 2020) is used as textual encoder. The base BERT model contains 12 layers with a hidden dimension of 768, therefore the semantic feature I_t (i.e., the class token "CLS" of the output layer) is a 768-dimensional vector. For cross-modal encoder, the number of bottleneck tokens of MBT is insensitive and set to $N_b = 4$ according to the conclusions in Liu et al. (2022), the number of attention heads in multi-head attention layers is 8, the hidden dimension is $d_m = 64$ and sine-cosine positional encoding is used to preserve the temporal information in the audio/visual feature sequence. The number of floating point operations per second (FLOPs) is 7.22×10^9, the number of parameters is 173M, and the recognition time of one video is around 0.2 s.

Training parameters: Regarding the loss function, since both IEMOCAP and CMU-MOSEI datasets are multi-labeled, video emotion recognition is regarded as a multi-label binary classification task in this work, and the binary cross-entropy loss is adopted and weighted by the ratio of the number of positive and negative samples to alleviate the problem of unbalanced sample distribution. Adam optimizer is adopted with a mini-batch size of 8 and the initial learning rate is 1e-4 with early-stopping to prevent overfitting. For the audio and visual encoder backbones, we freeze the first two convolution groups of VGGish and the first three convolution groups of ResNet18, and use a smaller learning rate 1e-5 to fine-tune the rest parameters. The whole framework is implemented using PyTorch on one NVIDIA TITAN RTX GPU.

4.4. Results and analysis

4.4.1. Comparison with the state-of-the-art

We compare our model with the following state of the art (SOTA) works where the audio, visual and text modalities are considered: (1) Late Fusion LSTM (LF-LSTM), where each modality uses an individual LSTM to extract global features followed by an MLP for unimodal decision, and the final prediction is obtained by weighted fusion; (2) Late Fusion Transformer (LF-TRANS) which is similar to LF-LSTM except that the Transformer models are used instead of LSTMs to model the temporal dependency for each modality; (3) EmoEmbs (Dai et al., 2020) where three LSTMs are

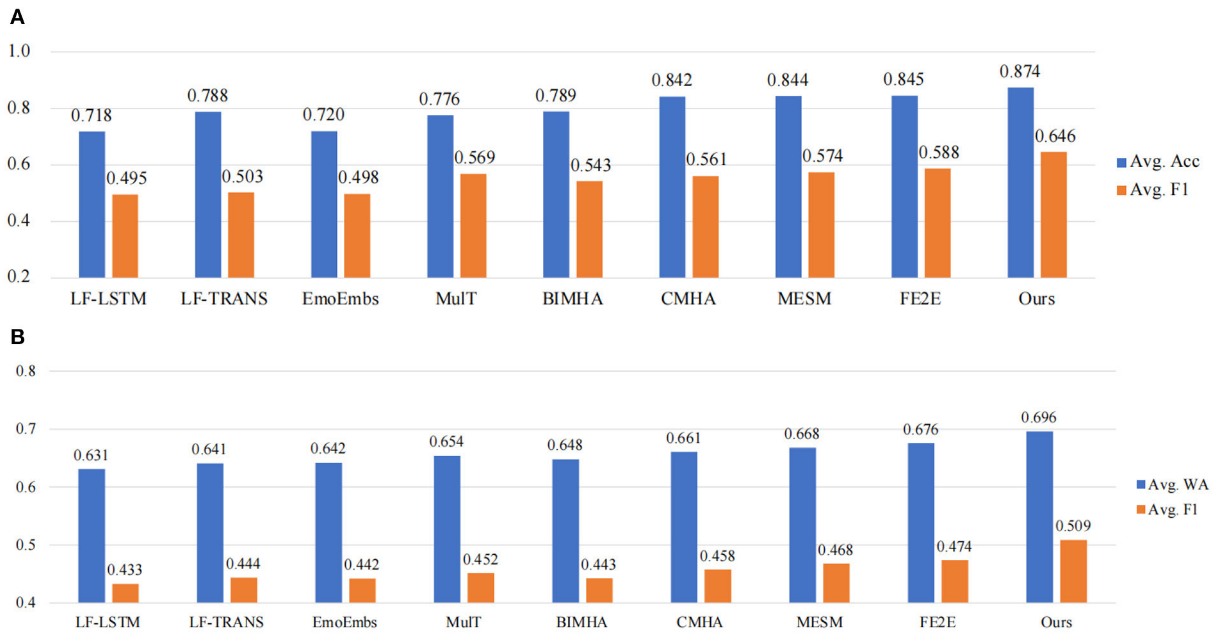


FIGURE 4
Comparison with state-of-the-art methods on two benchmark datasets. **(A)** Comparison results on IEMOCAP. **(B)** Comparison results on CMU-MOSEI.

adopted to obtain the global features for each modality and generates modality-specific emotion embeddings through mapping the GloVe textual emotion embeddings to the non-textual modalities respectively, and finally the similarity scores between the emotion embedding and the global features are calculated and fused to get the final prediction; (4) MulT (Tsai et al., 2019) that employs six cross-modal attention modules for any two pairs of the three modalities, and then three self-attention modules to collect temporal information within each modality. Finally the concatenated features are passed through the fully-connected layers to make predictions; (5) BIMHA (Wu et al., 2022) mainly consists of two parts: inter-modal interaction and inter-bimodal interaction, where the outer product is first used to represent three pairs of bimodal global features and then the bimodal attention is calculated *via* an extended multi-head attention mechanism; (6) CMHA (Zheng et al., 2022) where the core is connecting multiple multi-head attention modules in series, to model the interactions between two unimodal feature sequences first and then with the third one. Additionally, the sequential order of modality fusion is considered, resulting in three similar fusion modules but in different orders of fusion; (7) FE2E (Dai et al., 2021) which is a fully end-to-end framework, where the textual features are extracted from a pre-trained ALBERT model and the audio and visual features are extracted from two pre-trained CNNs, each followed by a Transformer to encode the sequential representations, and then three MLPs are adopted to make unimodal decision and weighted fusion is performed to output predictions; (8) MESM (Dai et al., 2021) which is similar to FE2E, except that the original CNN layers are replaced with cross-modal sparse CNN blocks to reduce the computational overhead.

The results are shown in Figure 4. Note that all the SOTA results are based on tri-modal decisions from audio, visual and text. It should also be mentioned that, the first five methods (LF-LSTM, LF-TRANS, EmoEmbs, MulT, and BIMHA) are based on hand-crafted features, where 142-dimensional audio features are extracted using the DisVoice toolkit (Vasquez-Correa et al., 2019), 35-dimensional visual features are extracted *via* the OpenFace toolkit (Baltrusaitis et al., 2018), and 300-dimensional word embeddings are extracted using the pre-trained GloVe Pennington et al. (2014). To evaluate the significance of our experimental results, following (Zhang et al., 2021a), the paired *t*-test is performed with a default significance level of 0.05. As it can be seen, our proposed model outperforms all the SOTA works on both IEMOCAP and CMU-MOSEI datasets. The average accuracy reaches 0.874 and the average F_1 is 0.646 on IEMOCAP dataset. On CMU-MOSEI dataset, our model also achieves the highest average weighted accuracy of 0.696 and an average F_1 of 0.509. In addition, the end-to-end methods achieve superior recognition results compared to the two-stage methods based on hand-crafted features, indicating that joint optimization of unimodal feature extraction and multimodal fusion helps improve the performance of video emotion recognition. It should also be mentioned that MESM

(Dai et al., 2021) was equipped with cross-modal attention in the feature encoding stage with the aim to make CNN encoders sparse, however, modeling the emotion dependency between audio-video sequences, as a key for multimodal emotional representation learning, was neglected in their whole framework. Compared with MESM, our proposed MER-SEM-MBT obtains better performance due to additional audio-visual temporal interaction.

We also list the binary classification results regarding each emotion category to make a deeper comparison. The detailed results are listed in Table 2, and the best results are bolded. One can notice that our proposed MER-SEM-MBT model achieves the best results on majority emotion category. In addition, we verify a variation of the proposed model by removing the textual decision and the corresponding results are listed in the last row. Under this circumstance, our proposed method, equipped with SEM and MBT modules, still obtains a comparative performance without a textual decision.

4.4.2. Ablation study

4.4.2.1. Effect of SEM and MBT

To evaluate the contribution of each design module, we further carry out experiments on different model variants by ablating either SEM or MBT, corresponding to MER-MBT (without SEM in unimodal audio/visual encoder) and MER-SEM (without MBT as the cross-modal encoder) respectively. The results are shown in Table 3, where MER stands for a baseline model with unimodal encoders and late attentive fusion. As we can see, either MER-SEM or MER-MBT yields a sub-optimal performance on both IEMOCAP and CMU-MOSEI datasets. Specifically, when MBT is removed, meaning there is no temporal interactions between audio and visual feature sequences, the modal variant MER-SEM obtains an average F_1 of 0.636 on IEMOCAP dataset with a decrease of 1% compared with our full model MER-SEM-MBT, but still 2.2% better than the baseline MER model benefiting from the semantic guidance from SEM. Similarly, when SEM is removed, the model variant MER-MBT achieves an average F_1 of 0.633 on IEMOCAP, which is 1.3% lower than the full model. Furthermore, if both SEM and MBT modules are removed, i.e., the baseline MER model, the average F_1 only reaches 0.614 on IEMOCAP, which is 3.2% lower than our proposed full model MER-SEM-MBT. This may be due to the fact that the baseline model MER only adopts attentive fusion to aggregate the individual audio and visual decisions without interaction across different modalities. Similar conclusions can also be drawn from the reuslts on the CMU-MOSEI dataset.

4.4.2.2. Effectiveness of SEM in audio/visual encoder

We further analyze the effectiveness of SEM on audio and visual representation learning for audio and visual emotion recognition, respectively. For convenience, we denote the audio emotion recognition as SER and visual emotion recognition as FER. Note that the textual decision is not used in the following experiments. As listed in Table 4, the first/third row represents the SER/FER results from the CNN-Transformer-MLP framework without SEM, where the CNN encoder (VGGish for audio and ResNet18 for video) is for feature extraction from raw data, Transformer is for temporal modeling, and MLP is for classification. The second/fourth row shows the results of SEM being inserted in the unimodal CNN encoder for SER/FER. It can be seen that when SEM is inserted to guide the audio/visual encoder to learn the emotional representation from the semantics, the performances are greatly improved. For SER, the average Acc improves from 0.752 to 0.839 on IEMOCAP dataset with a gain of 8.7% after SEM is used to enhance the representation learning of audio encoder. For FER, the average Acc also achieves a gain of 4.4% in terms of Avg. WA on CMU-MOSEI dataset.

4.4.2.3. Effect of the number of SEMs

As described in Section 3.1.4, SEM is empirically inserted after the second and last (fourth) convolution group for audio encoder backbone VGGish, and the third and last (fifth) convolution group for visual encoder backbone ResNet18, respectively. Here, we conduct experiments on IEMOCAP dataset to explore the effect of different numbers of SEMs in audio/visual encoder, the results are shown in Figure 5. Taking SER for example, the default setting is inserting two SEMs after the second and the fourth convolutional group, respectively. From Figure 5A, we can see that when adding another SEM after the third convolution group of VGGish, the result is close to that of the default setting, and further adding another SEM after the first convolution group results in a significant drop in performance. Similar conclusion can be drawn from Figure 5B for visual encoder. This is probably because the feature maps output from the earlier convolution group mainly contain low-level information, while those from the deeper layers with high-order features are more relevant to emotions, therefore the semantics can better adapt the high-level audio/visual feature maps with emotion-related information.

4.4.2.4. Performance comparison of different cross-modal encoders

To validate the effectiveness of adopting MBT as cross-modal encoder in our proposed framework, we perform audio-visual multi-modal emotion recognition (MER) experiments using different cross-modal encoders. Note that all the methods in this comparative experiment use the same audio and visual encoders, i.e., VGGish for audio and ResNet18 for video (without using semantic information for enhancement), and the same attentive decision fusion as described in Section 3.3. The results are shown in Table 5.

Concretely, three typical attention-based solutions are compared: (1) joint attention (JointAtt), where the audio and

TABLE 2 Binary classification results of each emotion category on IEMOCAP and CMU-MOSEI datasets.

Models	IEMOCAP											
	Happiness		Anger		Sadness		Excited		Frustrated		Neutral	
	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1
LF-LSTM†	0.672	0.376	0.712	0.494	0.782	0.540	0.793	0.572	0.682	0.515	0.665	0.470
LF-TRANS†	0.852	0.376	0.819	0.507	0.874	0.574	0.853	0.573	0.605	0.493	0.724	0.497
EmoEmbs (Dai et al., 2020)†	0.696	0.383	0.659	0.489	0.808	0.530	0.735	0.583	0.685	0.520	0.736	0.487
MulT (Tsai et al., 2019)†	0.800	0.468	0.779	0.607	0.835	0.654	0.769	0.580	0.724	0.570	0.749	0.537
BIMHA (Wu et al., 2022)††	0.834	0.432	0.772	0.576	0.838	0.637	0.783	0.561	0.739	0.542	0.764	0.509
CMHA (Zheng et al., 2022)††	0.890	0.458	0.886	0.611	0.883	0.616	0.879	0.605	0.751	0.563	0.765	0.512
MESM (Dai et al., 2021)†	0.895	0.473	0.882	0.628	0.886	0.622	0.883	0.612	0.749	0.584	0.770	0.520
FE2E (Dai et al., 2021)†	**0.900**	0.448	0.887	0.639	0.891	0.657	0.891	0.619	0.712	0.578	0.791	0.584
MER-SEM-MBT (Our full model)	0.891	**0.577**	**0.894**	**0.665**	**0.924**	**0.721**	**0.905**	**0.677**	**0.797**	**0.613**	**0.832**	**0.623**
MER-SEM-MBT (Ours w/o textual decision)	0.889	0.546	0.893	0.662	0.918	0.701	0.892	0.643	0.794	0.602	0.827	0.613
Models	CMU-MOSEI											
	Happiness		Sadness		Anger		Surprise		Fear		Disgust	
	WA	F_1	WA	F_1	WA	F_1	WA	F_1	WA	F_1	WA	F_1
LF-LSTM†	0.613	0.732	0.634	0.472	0.645	0.471	0.571	0.206	0.617	0.222	0.705	0.498
LF-TRANS†	0.606	0.729	0.601	0.455	0.653	0.477	0.621	0.242	0.621	0.240	0.744	0.519
EmoEmbs (Dai et al., 2020)†	0.612	0.719	0.605	0.475	0.668	0.494	0.633	0.240	0.638	0.234	0.696	0.487
MulT (Tsai et al., 2019)†	0.672	**0.754**	0.640	0.483	0.649	0.475	0.614	0.256	0.629	0.253	0.716	0.493
BIMHA (Wu et al., 2022)††	0.658	0.721	0.626	0.479	0.653	0.474	0.625	0.249	0.618	0.247	0.705	0.489
CMHA (Zheng et al., 2022)††	0.652	0.721	0.642	0.467	0.659	0.491	0.645	0.266	0.634	0.273	0.736	0.532
MESM (Dai et al., 2021)†	0.641	0.723	0.630	0.466	0.668	0.493	0.657	0.272	0.658	0.289	0.756	0.564
FE2E (Dai et al., 2021)†	0.654	0.726	0.652	0.490	0.670	**0.496**	0.667	0.291	0.638	0.268	0.777	0.571
MER-SEM-MBT (Our full model)	**0.673**	0.753	**0.668**	**0.538**	**0.687**	0.495	0.676	**0.330**	**0.672**	**0.319**	**0.802**	**0.616**
MER-SEM-MBT (Ours w/o textual decision)	0.672	0.749	0.655	0.531	0.673	0.491	0.660	0.328	0.659	0.312	0.787	0.612

$P < 0.05$ for paired t-test. † denotes the results are from Dai et al. (2021), and †† means our reproduction using the same data split as other experiments. The bold values are indicated to highlight the best results.

TABLE 3 Ablation study results on IEMOCAP and CMU-MOSEI datasets.

Models	SEM	MBT	LF	IEMOCAP		CMU-MOSEI	
				Avg. Acc	Avg. F_1	Avg. WA	Avg. F_1
MER	-	-	✓	0.855	0.614	0.682	0.496
MER-SEM	✓	-	✓	0.871	0.636	0.691	0.506
MER-MBT	-	✓	✓	0.868	0.633	0.688	0.504
MER-SEM-MBT	✓	✓	✓	**0.874**	**0.646**	**0.696**	**0.509**

The bold values are indicated to highlight the best results.

TABLE 4 Unimodal audio/visual emotion recognition results with and without SEM.

Methods		IEMOCAP		CMU-MOSEI	
		Avg. Acc	Avg. F_1	Avg. WA	Avg. F_1
SER	w/o SEM	0.752	0.463	0.628	0.424
	w/ SEM	**0.839**	**0.560**	**0.659**	**0.450**
FER	w/o SEM	0.796	0.512	0.631	0.429
	w/ SEM	**0.828**	**0.553**	**0.675**	**0.456**

SER refers to speech emotion recognition, and FER denotes facial expression recognition. All frameworks follow the CNN-Transformer-MLP architecture, the difference is whether SEM is used in the CNN encoder. The bold values are indicated to highlight the best results.

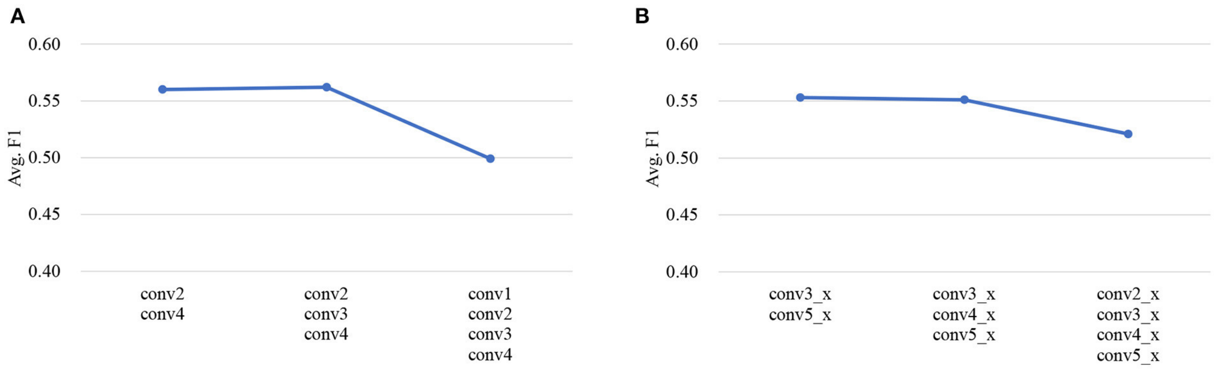


FIGURE 5
The effect of different numbers of SEMs in audio and visual encoder, respectively. The results are shown in terms of Avg. F_1 on IEMOCAP dataset. **(A)** Audio encoder (VGGish with SEM). **(B)** Visual encoder (ResNet18 with SEM).

TABLE 5 Audio-visual emotion recognition results using different cross-modal encoders.

Cross-modal Encoder	IEMOCAP		CMU-MOSEI	
	Avg. Acc	Avg. F_1	Avg. WA	Avg. F_1
JointAtt (Vaswani et al., 2017)	0.846	0.582	0.667	0.487
Bi-CrossAtt (Tsai et al., 2019)	0.842	0.571	0.671	0.473
MBT (Liu et al., 2022)	**0.859**	**0.592**	**0.676**	**0.491**

The bold values are indicated to highlight the best results.

visual feature sequences are temporally concatenated and then input into a vanilla Transformer (Vaswani et al., 2017), therefore the information within these two modalities can be fully communicated; (2) bidirectional cross-attention (Bi-CrossAtt) (Tsai et al., 2019), where two cross-modal Transformer branches are employed, each serves to reinforce a target modality with the features from the other modality *via* learning the attention across the audio and visual feature sequences; (3) multimodal

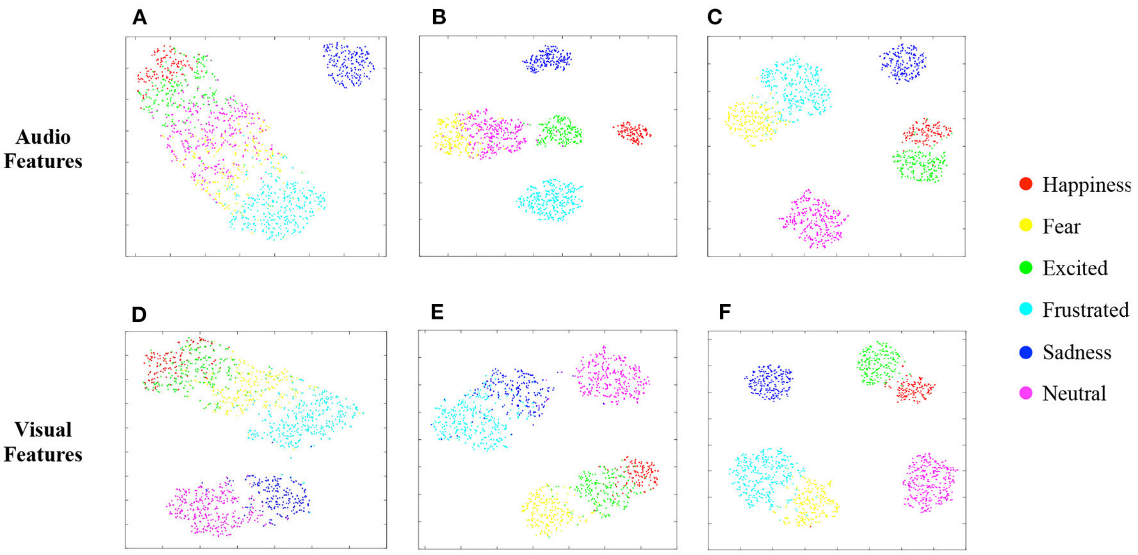


FIGURE 6
Visualization of audio and visual feature distribution on IEMOCAP. **(A)** SER w/o SEM, w/o MBT. **(B)** SER w/SEM, w/o MBT. **(C)** MER w/SEM, w/MBT. **(D)** FER w/o SEM, w/o MBT. **(E)** FER w/SEM, w/o MBT. **(F)** MER w/SEM, w/o MBT.

bottleneck attention (MBT) (Liu et al., 2022), which introduces bottleneck tokens as the bridge connecting two Transformer branches, to exchange essential information from one modality to the other through a two-stage cross-modal interaction.

It can be seen that the cross-modal interaction with MBT achieves the highest recognition results on both datasets, indicating that attention bottlenecks, with the advantage of exchanging audio-visual complementary information and reducing redundancy, further enhance the representation learning of audio/visual modalities.

4.4.3. Visualization

We also perform t-SNE (Van der Maaten and Hinton, 2008) to visualize the learned audio and visual features, under three different settings, from the penultimate layer of their MLPs, respectively. Note that the textual decision is not used in the involved models here. Figures 6A, D represents the audio/visual features learned by the unimodal SER/FER model without SEM and MBT, which corresponds to the results in the first/third row of Table 4. As we can see, the learned audio/visual features can not distinguish different emotions well in the absence of additional information from other modalities. When SEM is added in the audio/visual encoder for SER/FER, the enhanced audio/visual features of different emotion categories, as shown in Figures 6B, E, are more discriminatively distributed, which help to improve the emotion recognition performance as compared in Table 4. In addition, when MBT is further added, achieving cross-modal interaction between audio and visual representations, the dually reinforced audio/visual features (corresponds to Figures 6C, F) are more distinguishable, contributing to the best performance.

5. Conclusions

In this work, we proposed a multimodal interaction enhanced representation learning method targeting video emotion recognition. The high-level semantic information extracted from the text modality is utilized to enhance audio and visual feature encoding, and the bottleneck Transformer is adopted to further reinforce audio and visual feature sequences through exchanging complementary information while reducing redundancy. Finally, audio, visual, and textual unimodal decisions are fused using attention weights to output the final emotion prediction. Experiments and visualization show that the proposed method achieves state-of-the-art video emotion recognition results. In the future, we are interested to leverage self-supervised learning methods to learn better emotional-salient representations by exploring the correlations among audio, visual, and text modalities.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://sail.usc.edu/iemocap/ and http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/.

Author contributions

XX, YZ, and DJ contributed to conception and design of the study. XX wrote the first draft of the manuscript. YZ and DJ revised the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 62236006), the Key Research and Development Program of Shaanxi (Grant No. 2022ZDLGY06-03), and the Shaanxi Provincial International Science and Technology Collaboration Project (Grant No. 2017KW-ZD-14).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alisamir, S., and Ringeval, F. (2021). On the evolution of speech representations for affective computing: a brief history and critical overview. *IEEE Signal Process. Mag.* 38, 12–21. doi: 10.1109/MSP.2021.3106890

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*. doi: 10.48550/arXiv.1607.06450

Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L.-P. (2018). "Openface 2.0: facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018)* (Xi'an: IEEE), 59–66.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., et al. (2008). Iemocap: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluat.* 42, 335–359. doi: 10.1007/s10579-008-9076-6

Dai, W., Cahyawijaya, S., Liu, Z., and Fung, P. (2021). "Multimodal end-to-end sparse model for emotion recognition," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Mexico City: Association for Computational Linguistics), 5305–5316.

Dai, W., Liu, Z., Yu, T., and Fung, P. (2020). "Modality-transferable emotion embeddings for low-resource multimodal emotion recognition," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing* (Suzhou: Association for Computational Linguistics), 269–280.

Degottex, G., Kane, J., Drugman, T., Raitio, T., and Scherer, S. (2014). "COVAREP–a collaborative voice analysis repository for speech technologies," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Florence: IEEE), 960–964.

Devlin, J., Chang, M. -W., Lee, K., and Toutanova, K. (2019). "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapolis, MN: Association for Computational Linguistics), 4171–4186.

Fu, Z., Liu, F., Xu, Q., Qi, J., Fu, X., Zhou, A., et al. (2022). "NHFNET: a non-homogeneous fusion network for multimodal sentiment analysis," in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (Taipei: IEEE), 1–6.

Han, W., Chen, H., Gelbukh, A., Zadeh, A., Morency, L. -P., and Poria, S. (2021). "Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis," in *Proceedings of the 2021 International Conference on Multimodal Interaction* (Montreal, QC: Association for Computing Machinery), 6–15.

Hazarika, D., Zimmermann, R., and Poria, S. (2020). "MISA: Modality-invariant and -specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA), 1122–1131.

He, J., and Hu, H. (2021). Mf-bert: multimodal fusion in pre-trained bert for sentiment analysis. *IEEE Signal Process. Lett.* 29, 454–458. doi: 10.1109/LSP.2021.3139856

He, J., Mai, S., and Hu, H. (2021). A unimodal reinforced transformer with time squeeze fusion for multimodal sentiment analysis. *IEEE Signal Process. Lett.* 28, 992–996. doi: 10.1109/LSP.2021.3078074

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," |*PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV: IEEE), 770–778.

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., et al. (2017). "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (New Orleans, LA: IEEE), 131–135.

Lerner, J. S., and Keltner, D. (2000). Beyond valence: toward a model of emotion-specific influences on judgement and choice. *Cogn. Emot.* 14, 473–493. doi: 10.1080/026999300402763

Li, S., and Deng, W. (2022). Deep facial expression recognition: a survey. *IEEE Trans. Affect. Comput.* 13, 1195–1215. doi: 10.1109/TAFFC.2020.2981446

Liu, Y., Li, S., Wu, Y., Chen, C. -W., Shan, Y., and Qie, X. (2022). "UMT: Unified multi-modal transformers for joint video moment retrieval and highlight detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans, LA: IEEE), 3042–3051.

Ma, Y., Hao, Y., Chen, M., Chen, J., Lu, P., and Košir, A. (2019). Audio-visual emotion fusion (avef): a deep efficient weighted approach. *Inf. Fusion* 46, 184–192. doi: 10.1016/j.inffus.2018.06.003

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. doi: 10.48550/arXiv.1301.3781

Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2017). Affectnet: a database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* 10, 18–31. doi: 10.1109/TAFFC.2017.2740923

Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., and Sun, C. (2021). "Attention bottlenecks for multimodal fusion," in *Advances in Neural Information Processing Systems, Vol. 34* (Curran Associates), 14200–14213.

Nguyen, D., Nguyen, D. T., Zeng, R., Nguyen, T. T., Tran, S. N., Nguyen, T., et al. (2021). Deep auto-encoders with sequential learning for multimodal dimensional emotion recognition. *IEEE Trans. Multimedia* 24, 1313–1324. doi: 10.1109/TMM.2021.3063612

Pennington, J., Socher, R., and Manning, C. D. (2014). "Glove: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha: IEEE), 1532–1543.

Provost, E. M. (2013). "Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (Vancouver, BC: IEEE), 3682–3686.

Rouast, P. V., Adam, M. T., and Chiong, R. (2019). Deep learning for human affect recognition: Insights and new developments. *IEEE Trans. Affect. Comput*. 12, 524–543. doi: 10.1109/TAFFC.2018.2890471

Sharma, G., and Dhall, A. (2021). "A survey on automatic multimodal emotion recognition in the wild," in *Advances in Data Science: Methodologies and Applications*, eds G. Phillips-Wren, A. Esposito, and L. C. Jain (Cham: Springer), 35–64. doi: 10.1007/978-3-030-51870-7_3

Sun, Z., Sarma, P., Sethares, W., and Liang, Y. (2020). Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. *Proc. AAAI Conf. Artif. Intell*. 34, 8992–8999. doi: 10.1609/aaai.v34i05.6431

Tsai, Y. -H. H., Bai, S., Liang, P. P., Lolter, J. Z., Morency, L. -P., and Salakhutdinov, R. (2019). "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence: Association for Computational Linguistics), 6558–6569.

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res*. 86, 2579–2605. Available online at: http://jmlr.org/papers/v9/vandermaaten08a.html

Vasquez-Correa, J. C., Klumpp, P., Orzco_Arroyave, J. R., and Noth, E. (2019). "Phonet: A tool based on gated recurrent neural networks to extract phonological posteriors from speech," in *Proc Interspeech 2019* (Graz), 549–553.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems* (Long Beach, CA), 5998–6008.

Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., et al. (2022). A systematic review on affective computing: emotion models, databases, and recent advances. *Inf. Fusion*. 83–84, 19–52. doi: 10.1016/j.inffus.2022.03.009

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020). "Transformers: state-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.

Wu, T., Peng, J., Zhang, W., Zhang, H., Tan, S., Yi, F., et al. (2022). Video sentiment analysis with bimodal information-augmented multi-head attention. *Knowledge Based Syst*. 235, 107676. doi: 10.1016/j.knosys.2021.107676

Wu, Y., Lin, Z., Zhao, Y., Qin, B., and Zhu, L.-N. (2021). "A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (Bangkok), 4730–4738.

Yu, W., Xu, H., Yuan, Z., and Wu, J. (2021). Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *Proc. AAAI Conf. Artif. Intell*. 35, 10790–10797. doi: 10.1609/aaai.v35i12.17289

Zadeh, A., and Pu, P. (2018). "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 (Long Papers)* (Melbourne, VIC), 2236–2246.

Zhang, K., Li, Y., Wang, J., Cambria, E., and Li, X. (2021a). Real-time video emotion recognition based on reinforcement learning and domain knowledge. *IEEE Trans. Circ. Syst. Video Technol*. 32, 1034–1047. doi: 10.1109/TCSVT.2021.3072412

Zhang, K., Li, Y., Wang, J., Wang, Z., and Li, X. (2021b). Feature fusion for multimodal emotion recognition based on deep canonical correlation analysis. *IEEE Signal Process. Lett*. 28, 1898–1902. doi: 10.1109/LSP.2021.3112314

Zhang, Y., Chen, M., Shen, J., and Wang, C. (2022). Tailor versatile multi-modal learning for multi-label emotion recognition. *Proc. AAAI Conf. Artif. Intell*. 36, 9100–9108. doi: 10.1609/aaai.v36i8.20895

Zhang, Y., Wang, Z.-R., and Du, J. (2019). "Deep fusion: an attention guided factorized bilinear pooling for audio-video emotion recognition," in *2019 International Joint Conference on Neural Networks (IJCNN)* (Budapest: IEEE), 1–8.

Zheng, J., Zhang, S., Wang, Z., Wang, X., and Zeng, Z. (2022). Multi-channel weight-sharing autoencoder based on cascade multi-head attention for multimodal emotion recognition. *IEEE Trans. Multimedia*. 1–13. doi: 10.1109/TMM.2022.3144885

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 06 January 2023
DOI 10.3389/fnins.2022.1107284

OPEN ACCESS

EDITED BY
Xiaopeng Hong,
Harbin Institute of Technology, China

REVIEWED BY
Lang He,
Xi'an University of Posts
and Telecommunications, China
Yong Li,
Nanjing University of Science
and Technology, China

*CORRESPONDENCE
Hongsheng Lu
✉ luhs@tzc.edu.cn

SPECIALTY SECTION
This article was submitted to
Perception Science,
a section of the journal
Frontiers in Neuroscience

RECEIVED 24 November 2022
ACCEPTED 13 December 2022
PUBLISHED 06 January 2023

CITATION
Zhao X, Liao Y, Tang Z, Xu Y, Tao X, Wang D, Wang G and Lu H (2023) Integrating audio and visual modalities for multimodal personality trait recognition *via* hybrid deep learning.
Front. Neurosci. 16:1107284.
doi: 10.3389/fnins.2022.1107284

Integrating audio and visual modalities for multimodal personality trait recognition *via* hybrid deep learning

Xiaoming Zhao[1], Yuehui Liao[1,2], Zhiwei Tang[1], Yicheng Xu[3], Xin Tao[1], Dandan Wang[1], Guoyu Wang[1] and Hongsheng Lu[1]*

[1]Taizhou Central Hospital (Taizhou University Hospital), Taizhou University, Taizhou, Zhejiang, China, [2]School of Computer Science, Hangzhou Dianzi University, Hangzhou, China, [3]School of Information Technology Engineering, Taizhou Vocational and Technical College, Taizhou, Zhejiang, China

Recently, personality trait recognition, which aims to identify people's first impression behavior data and analyze people's psychological characteristics, has been an interesting and active topic in psychology, affective neuroscience and artificial intelligence. To effectively take advantage of spatio-temporal cues in audio-visual modalities, this paper proposes a new method of multimodal personality trait recognition integrating audio-visual modalities based on a hybrid deep learning framework, which is comprised of convolutional neural networks (CNN), bi-directional long short-term memory network (Bi-LSTM), and the Transformer network. In particular, a pre-trained deep audio CNN model is used to learn high-level segment-level audio features. A pre-trained deep face CNN model is leveraged to separately learn high-level frame-level global scene features and local face features from each frame in dynamic video sequences. Then, these extracted deep audio-visual features are fed into a Bi-LSTM and a Transformer network to individually capture long-term temporal dependency, thereby producing the final global audio and visual features for downstream tasks. Finally, a linear regression method is employed to conduct the single audio-based and visual-based personality trait recognition tasks, followed by a decision-level fusion strategy used for producing the final Big-Five personality scores and interview scores. Experimental results on the public ChaLearn First Impression-V2 personality dataset show the effectiveness of our method, outperforming other used methods.

KEYWORDS

multimodal personality trait recognition, hybrid deep learning, convolutional neural networks, bi-directional long short-term memory network, Transformer, spatiotemporal

1. Introduction

In personality psychology, researchers believe that human personality is innate, and have developed various theoretical methods to understand and measure a person's personality. Costa and McCrae (1998) proposed a personality trait theory, in which personality characteristic were referred to as the main factors affecting the characteristics of individual behaviors, the critical factor in forming personality traits, and the basic unit for measuring personality traits. In Vinciarelli and Mohammadi (2014) personality is defined as: "personality is a psychological construct that can explain the diversity of human behaviors on the basis of a few, stable and measurable individual characteristics." At present, researchers have used psychological scales to establish various personality traits models, including Big-Five (McCrae and John, 1992), Cattell sixteen personality factor (16PF) (Karson and O'Dell, 1976), Myers-Briggs type indicators (MBTI) (Furnham, 1996), Minnesota multiple personality inventory (MMPI) (Bathurst et al., 1997), and so on. Among them, the Big-Five model has become the most fashionable measure model for automatic personality trait recognition. In particular, the Big-Five model, also known as the OCEAN model, aims to measure a person's personality through five dipolar scales: openness (O), conscientiousness (C), extroversion (E), agreeableness (A), and neuroticism (N). In affective neuroscience, the neural mechanisms of emotion expression are investigated by means of combining neuroscience with the psychological study of personality, emotion, and mood (Montag and Davis, 2018; Wang and Zhao, 2022; Zhang et al., 2022).

In recent years, researchers have employed computational techniques such as machine learning and deep learning methods (Gao et al., 2020; Liang et al., 2021; Wang and Deng, 2021; Yan et al., 2021; Ye et al., 2021) to model and measure human personality from the first impression behavior data, which is called personality computing (Junior et al., 2019). One of the most important research subject in personality computing is automatic personality trait recognition, which aims to identify people's first impression behavior data by computer and then analyze people's psychological characteristics (Zhao et al., 2022). Personality trait recognition has significant applications to human emotional behavior analysis, human-computer interaction, and interview recommendation. For example, Zhao et al. (2019) explored the influence of personality on emotional behavior by means of a hypergraph learning framework. When an enterprise recruits, human resource department can leverage personality trait recognition techniques to analyze personality characteristics of the job seekers by collecting their first-impression behavior data, and then select employees who can better meet the needs of the enterprise. To advance the development of personality trait recognition, the 2016 European Conference on Computer Vision (ECCV) released a publicly available personality dataset, i.e., ChaLearn-2016, and organized an academic competition of personality trait recognition (Ponce-López et al., 2016). Since 2016, personality trait recognition has become a hot research topic in psychology, affective neuroscience, and artificial intelligence.

In a basic personality trait recognition system, two important steps are involved: feature extraction and personality trait classification or prediction (Zhao et al., 2022). Feature extraction aims to derive appropriate feature parameters related to the expression of personality traits from the acquired first impression behavioral data. Personality trait classification or prediction aims to employ machine learning methods to conduct personality classification or prediction. The conventional classifiers or regressors such as support vector machines (SVM) and linear regressors can be adopted for personality trait classification or prediction. This paper will focus on feature extraction in a personality trait recognition system.

According to the types of extracted features characterizing personality traits, personality trait recognition techniques can be divided into hand-crafted based methods and deep learning based methods. Based on the extracted hand-crafted or deep learning features, previous works (Zhao et al., 2022) focus on performing personality trait recognition from single modality, such as audio-based personality trait recognition (Mohammadi and Vinciarelli, 2012), visual-based personality trait recognition (Gürpınar et al., 2016), etc. Although these works based on single modality have achieved good performance, there are still two limitations for them. First, the people's first impression behavior data in real-world scenery are often multimodal rather than single-modal for characterizing personality traits. For instance, both verbal and non-verbal information such as audio and visual modality are highly correlated with personality traits. In this case, it is thus necessary to adopt multiple input modalities for personality trait recognition. Second, although deep learning methods have been fashionable for personality trait recognition, each of them has its advantages and disadvantages. Therefore, integrating the advantages of different deep learning methods may further improve the performance of personality trait recognition, which will be investigated in this work.

To address these two issues above-mentioned, this paper proposes a multimodal personality trait recognition method integrating audio and visual modalities based on a hybrid deep learning framework. As depicted in **Figure 1**, the proposed method combines three different deep models, including convolutional neural networks (CNN) (LeCun et al., 1998; Krizhevsky et al., 2012), bi-directional long short-term memory network (Bi-LSTM) (Schuster and Paliwal, 1997), recently emerged Transformer (Vaswani et al., 2017), to learn high-level audio-visual feature representations, followed by a decision-level fusion strategy for final personality trait recognition. In particular, for audio feature extraction, the pre-trained deep audio CNN model called VGGish (Hershey et al., 2017) is used to learn high-level segment-level audio features. For visual feature extraction, the pre-trained deep face CNN model called VGG-Face (Parkhi et al., 2015) is leveraged to separately learn high-level frame-level global scene image features and local facial image features from each frame in dynamic video sequences. Then, these extracted deep audio-visual features are fed into a Bi-LSTM and a Transformer network (Vaswani et al., 2017) to individually capture long-term temporal dependency, thereby producing the final global audio and visual features for downstream tasks. Finally, a linear regression method is employed to conduct the single audio-based and visual-based personality trait recognition tasks, and yield six independent personality trait prediction scores. A decision-level fusion strategy is adopted to merge these personality trait prediction scores and output the final Big-Five personality scores and interview scores. Extensive experiments is conducted on the public ChaLearn First Impressions-V2 dataset (Escalante et al., 2017), and demonstrate the effectiveness of the proposed method on personality trait recognition tasks.

The main contributions of this paper are summarized as follows:

(1) This paper proposes a multimodal personality trait recognition method integrating audio and visual modalities based on a hybrid deep learning framework, in which CNN, Bi-LSTM, and Transformer are combined to capture high-level audio-visual spatio-temporal feature representations for personality trait recognition.

(2) Extensive experiments are performed on the public ChaLearn First Impressions-V2 dataset and experimental results show that the proposed method outperforms other comparing methods on personality trait recognition tasks.

2. Related work

The majority of prior works for personality trait recognition concentrates on single modality such as audio or visual cues, as described below.

2.1. Audio-based personality trait recognition

In early works, the conventional extracted hand-crafted audio features are low-level descriptor (LLD) features including intensity, pitch, formants, Mel-Frequency Cepstrum Coefficients (MFCCs), and so on. Mohammadi and Vinciarelli (2012) derived the LLD features like intensity, pitch, and formants, and then employed a logistic regression to predict the Big-five personality traits in audio clips. An et al. (2016) extracted the typical Interspeech-2013 ComParE feature set (Schuller et al., 2013) and fed them into a SVM classifier to conduct the Big-Five personality trait recognition.

In recent years, researchers have tried to leverage deep learning (LeCun et al., 2015) models with a multilayer network structure to learn high-level audio feature representations for promoting the performance of personality trait recognition. Among them, the representative deep learning methods are CNN (LeCun et al., 1998; Krizhevsky et al., 2012), recurrent neural networks (RNN) (Elman, 1990) and its variants called long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997), etc. Hayat et al. (2019) proposed an audio personality feature extraction method based on CNN. They fine-tuned the pre-trained CNN model called AudioSet in the first-impression behavior dataset and extracted high-level audio features for Big-Five personality prediction, demonstrating the advantages of CNN-based learned features compared with hand-crafted features. Zhu et al. (2018) presented a method of automatic perception of speakers' personality from speech in Mandarin. They developed a new skip-frame LSTM system to learn personality information from frame-level descriptor like MFCCs instead of hand-crafted prosodic features.

2.2. Visual-based personality trait recognition

In terms of the input type of visual data, visual-based personality trait recognition can be divided into two groups: static images-based and dynamic video sequences-based personality trait recognition.

For static images-based personality trait recognition, the extracted visual features mainly come from facial features, since facial morphology provides explicit cues for personality trait recognition. In early works, the commonly used hand-crafted facial features are color histograms, local binary patterns (LBP), global descriptor, aesthetic features, etc. Guntuku et al. (2015) extracted low-level hand-crafted features of facial images, including color histograms, LBP, global descriptor, and aesthetic features, and then employed the lasso regressor to predict the Big-five personality traits of users in self-portrait images. Recently, deep learning methods have been applied for static images-based personality trait recognition. Xu et al. (2021) explored the relationship between self-reported personality characteristics and static facial images. They investigated the performance of several deep learning models pre-trained on the ImageNet data, such as MobileNetv2, ResNeSt50, and the designed personality prediction neural network based on soft thresholding (S-NNPP) by means of fine-tuning them on the self-constructed dataset composed of facial images and personality characteristics.

For dynamic video sequences-based personality trait recognition, dynamic video sequences contain temporal information related to facial activity statistics, thereby providing

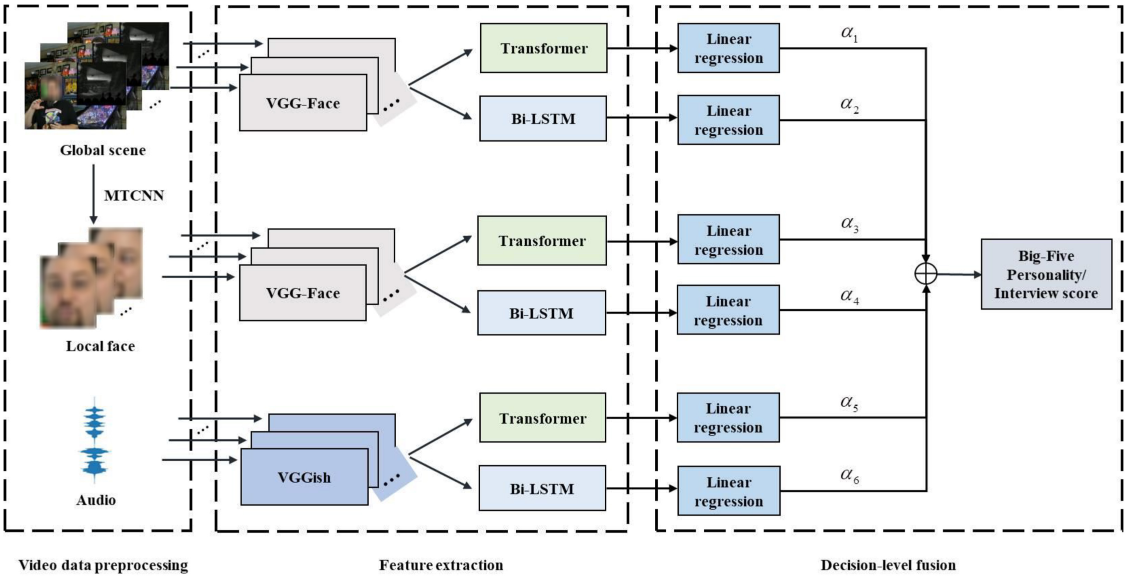


FIGURE 1
The flowchart of the proposed multimodal personality trait recognition method integrating audio and visual modalities based on a hybrid deep learning framework.

useful and complementary cues for personality trait recognition (Junior et al., 2019). In early works, the hand-crafted video features related to facial activity statistics were usually adopted for personality trait recognition. Teijeiro-Mosquera et al. (2014) exploited the relationships between facial expressions in dynamic video sequences and personality impressions of the Big-Five traits. To characterize facial activity statistics, they extracted four kinds of behavioral cues for personality trait recognition, including statistic-based cues, Threshold (THR) cues, Hidden Markov Models (HMM) cues, and Winner Takes All (WTA) cues. Likewise, several recently developed deep learning methods have been employed for dynamic video sequences-based personality trait recognition. Gürpınar et al. (2016) extracted deep facial and scene feature representations in dynamic video sequences by fine-tuning a pre-trained VGG-19 model, and then input them into a kernel extreme learning machine to perform the prediction of Big-Five personality traits. Beyan et al. (2021) presented a classification method of perceived personality traits on the basis of novel deep visual activity (VA)-based features derived only from key-dynamic images in dynamic video sequences. They adopted a dynamic image construction, which aimed to learn long-term VA with CNN + LSTM, and detect spatiotemporal saliency to decide key-dynamic images.

3. The proposed method

To alleviate the problem of single modality based personality trait recognition, this paper proposes a multimodal personality trait recognition method integrating audio and visual modalities based on a hybrid deep learning framework. Figure 1 depicts the flowchart of the proposed method. As depicted in Figure 1, the proposed method adopts two modalities as its input: one is the audio signals, the other is the visual signals including the global scene images and facial images. The used hybrid deep learning framework comprises of three different deep learning models like CNN, Bi-LSTM, and Transformer, which are used for high-level feature learning tasks. The proposed method consists of three key steps: video data preprocessing, audio-visual feature extraction, and decision-level fusion, as described below.

3.1. Video data preprocessing

For audio signals in the video data, we use the pre-trained VGGish model (Hershey et al., 2017) to extract high-level audio segment-level features. It is noted that the length of speech segments as input of VGGish is required to be 0.96 s. To this end, the original audio signals in the video data are divided into to a certain number of adjacent segments which last a time period of 0.96 s.

For visual signals in the video data, two preprocessing tasks are implemented. For global scene images in a video, 100 scene images are selected at equal intervals form each original video sample. Then, the resolution of each global scene image is resampled from the original 1280 × 720 pixels to 224 × 224 as inputs of VGG-Face model (Parkhi et al., 2015). For local face images in a video, we employ the popular Multi-Task Convolutional Neural Network (MTCNN) (Zhang et al., 2016) to conduct face detection tasks. The resolution of face image detected in each frame is sampled to 224 × 224. Since some videos are affected by environmental factors such as illumination, MTCNN may detect face images with a low accuracy. As a tradeoff, 30 frames of detected face images are selected at equal intervals from the original video. For the video with less than 30 frames of detected face images, the first and last face images are repeatedly until the frame number of face video is 30.

3.2. Audio-visual feature extraction

Audio-visual feature extraction aims to learn the local and global feature representations from original audio and visual signals in a video for personality trait recognition, as described below.

3.2.1. Audio-visual local feature extraction

For the divided audio segment with 0.96 s, we leverage the VGGish model (Hershey et al., 2017) pre-trained on the AudioSet dataset (Gemmeke et al., 2017) to capture high-level segment-level deep audio features. The used VGGish model consists of 6 convolutional layers, 4 pooling layers, and 3 fully connected layers. The kernel size of convolutional layers and pooling layers is 3 × 3 and 2 × 2, respectively. Since the neuron number of the last fully connected layer in the VGGish network is 128, the learned audio features by the VGGish model are 128-dimension.

For each scene and face image in a video, we employ the VGG-Face model (Parkhi et al., 2015) pre-trained on the ImageNet dataset (Deng et al., 2009) to learn high-level frame-level deep visual feature representations for downstream scene and face global feature learning tasks, respectively. The VGG-Face model includes 13 convolution layers, 5 pooling layers, and 2 fully connected layers. Since the neuron number of the last full connection layer in the VGG-Face network is 4096, the dimension of visual frame-level features obtained by VGG-Face network is 4096.

Given i-th input video clip a_i $(i = 1, 2, \cdots N)$ and its corresponding Big-Five personality score y_i, we fine-tune the pre-trained VGGish network (Hershey et al., 2017) to obtain deep segment-level audio feature representations, as described below:

$$\min_{W^{VG},\theta^{VG}} \sum_{i=1}^{N} L(\text{sigmoid}(W^{VG}\eta^{VG}(a_i;\theta^{VG})), y_i) \tag{1}$$

where $\eta^{VG}(a_i;\theta^{VG})$ represents the output of the last full connected layer in the VGGish network. θ^{VG} and W^{VG} separately denotes the network parameters of the VGGish network and the weights of the sigmoid layer. The cross-entropy loss function L is defined as:

$$L(VG, y) = -\sum_{j=1}^{N} y_j \log(y_j^p) \tag{2}$$

where y_j is the j-th ground-truth Big-Five personality score, and y_j^p is represented by the predicted Big-Five personality score.

For deep visual scene and face feature extraction on each frame of video, we fine-tune the pre-trained VGG-Face network (Parkhi et al., 2015) to learn high-level visual feature representations. The process of fine-tuning the pre-trained VGG-Face network is similar to the above-mentioned Eqs 1, 2.

3.2.2. Audio-visual global feature extraction

After completing the local audio and visual feature extraction tasks, it is necessary to individually learn the global audio features, visual scene features, and visual face features from the entire videos so as to conduct personality trait prediction tasks. To this end, we adopt the Bi-LSTM (Schuster and Paliwal, 1997) and recently emerged Transformer (Vaswani et al., 2017) to independently model long-term dependencies of temporal dynamics in video sequences, as described below.

Given an input sequence e_t, the learning process of the Bi-LSTM network is:

$$E = \text{Bi} - \text{LSTM}(W_{Bi-LSTM}, e_t) \tag{3}$$

where $E \in \mathbb{R}^{1\times d}$ is the learned temporal features, and $W_{Bi-LSTM}$ is weight parameters of Bi-LSTM.

The original Transformer (Vaswani et al., 2017) is developed based on self-attention mechanisms like a Multi-Head attention without any recurrent structures and convolutions. A Multi-Head attention module consists of several Scaled Dot-Product Attention (SDPA) modules in parallel and then their outputs are concatenated as an input of a linear layer. Given the input query (Q), key (K), and value (V), the output of each SDPA module is defined as:

$$\text{Attention}\,(Q, K, V) = \text{soft}\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

where d_k is the feature dimension of the key matrix K.

3.3. Decision-level fusion

After obtaining audio-visual global features extracted by a Bi-LSTM model and a Transformer model, we adopt a linear regression layer to predict the Big-Five personality and interview scores. The linear regression layer is calculated as follows:

$$f_i(x) = x_i w_i + b \tag{5}$$

where x_i, w_i, and b represent the i-th input sample, the corresponding weight value, and bias, respectively. $f_i(x)$ is the i-th prediction score value.

As shown in Figure 1, when using the learned audio features, visual scene features, and visual face features as inputs of a linear regression layer, we can obtain six different recognition results. To effectively fuse these six different recognition results, a weighted decision-level fusion strategy is employed, as described below:

$$\widetilde{f}(x) = \sum_{i=1}^{6} \alpha_i f_i(x) \tag{6}$$

where α_i is the weight value, $f_i(x)$ is the predicted value of each type of features, and $\sum_{i=1}^{6} \alpha_i = 1$. The mean squared error (MSE) loss is computed as follows:

$$\text{MSE}(\widetilde{f}(X)) = E[(\widetilde{f}(X) - Y)^2] = E\left[\left(\sum_{i=1}^{6} \alpha_i (f_i(X) - Y)\right)^2\right] \tag{7}$$

where Y is the ground-truth score. Our goal is to minimize the MSE loss subject to $\sum_{i=1}^{6} \alpha_i = 1$. To this end, the Lagrangian expression of this problem is expressed as:

$$L(X, \lambda) = \text{MSE}(\widetilde{f}(X)) - \lambda\left(\sum_{i=1}^{6} \alpha_i - 1\right) \tag{8}$$

where λ is the Lagrange multiplier.

Then, we calculate the partial derivation of Eq. 8 based on α_m for $m = 1, 2, \cdots 6$, as defined as:

$$\frac{\partial L(X, \lambda)}{\partial \alpha_m} = E\left[2\sum_{i=1}^{6} \alpha_i (f_i(X) - Y)(f_m(X) - Y)\right] - \lambda \tag{9}$$

We set the gradient to be 0, and get:

$$2\sum_{i=1}^{6} \alpha_i E[(f_i(X) - Y)(f_m(X) - Y)] - \lambda = 0, m = 1, 2, \cdots 6 \tag{10}$$

Let $\alpha = [\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6]^T$, $\Omega = [w_{ij}] = E[(f_i(X) - Y)(f_j(X) - Y)]$, Eq. 10 can be transformed as:

$$\Omega\alpha = \frac{\lambda}{2} 1 \tag{11}$$

Then, the optimal weight vector $\boldsymbol{\alpha}$ can be obtained by:

$$\alpha = \frac{\Omega^{-1} 1}{1^T \Omega^{-1} 1} \tag{12}$$

4. Experiments

4.1. Dataset

To verify the effectiveness of the proposed method, the public ChaLearn First Impression-V2 (Escalante et al., 2017) is employed for personality and interview prediction. This dataset contains 10,000 video clips collected from more than 3,000 different YouTube videos. The language involved in video participants is English. The resolution of the video is 1280 × 720, and the duration of each video clip is about 15 s. This dataset annotates the "Interview" scene labels for interview analysis. The divided training set, testing set and validation set in this dataset contain 6,000, 2,000, 2,000 video clips, respectively. In this work, we use the training and validation sets for experiments because the testing set is only open to competitors. Each video in this dataset is labeled by using the Big-Five personality score [0,1]. Figure 2 shows several image samples from the ChaLearn First Impression-V2 dataset.

4.2. Implementation details

When training all used deep learning models, the batch size is set to 32, and the initial learning rate is $1 \times e^{-4}$. After each epoch, the learning rate will become a half of the original learning rate. The maximum epoch number of is 30, and the Adam optimizer is used. The MSE loss function is adopted. The experimental platform is NVIDIA GPU Quadro M6000 with 24 GB memory. In order to improve the generalization performance of trained deep learning models and avoid overfitting, the early stopping strategy (Prechelt, 1998) is used.

In this work, we choose a two-layer Bi-LSTM to capture temporal dynamics related to video sequences. The number of neurons in each layer of Bi-LSTM is 2048. The number of encoding layer in the Transformer model is 6 for its best performance, and its last layer output 1024-dimension features. To compare with these deep learning models, several classical regression models such as Support Vector Regression (SVR) with polynomial (poly), radial basis function (RBF), and linear kernel functions, Decision Tree Regression (DTR) are employed. In the SVR model, the degree of polynomial kernel function is 3, the penalty factor "*C*" of radial basis kernel function is 2, and the parameter "gamma" is 0.5. The DTR model is implemented for its default parameters, such as the splitting policy "split = best" at each node, "min _ samples _ split = 2" for splitting an internal node. For these classical regression models, a simple average-pooling strategy is conducted on these extracted audio-visual local features so as to produce the global features as their inputs.

The evaluation metric for evaluating the predicted personality trait or interview scores is defined as:

$$S = 1 - \sum_{j=1}^{N} \frac{\left|y_j^p - y_j\right|}{N} \tag{13}$$

where N is the number of samples, y_j^p is the predicted value, and y_j is the ground-truth value. The higher the value S is, the better the obtained performance on personality or interview prediction tasks is.

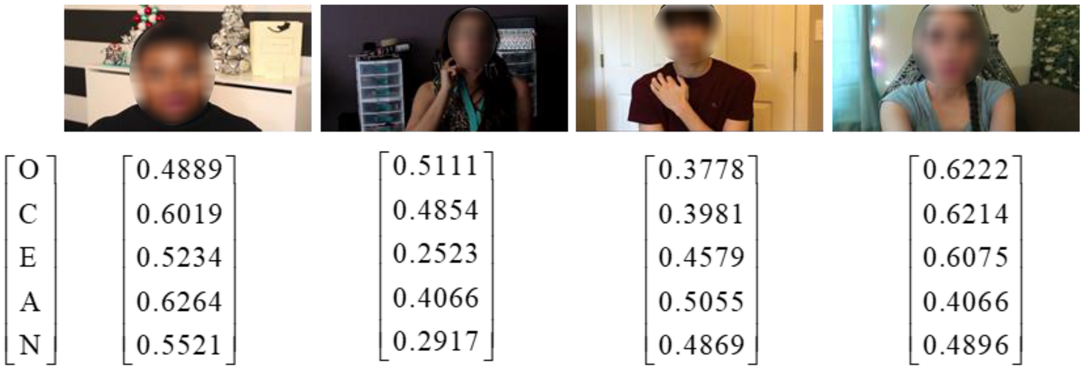


FIGURE 2
Image samples with the labeled Big-Five personality score from the ChaLearn First Impression-V2 dataset.

TABLE 1 Prediction results of deep audio features extracted by the pre-trained VGGish for different methods.

Models	O	C	E	A	N	Average score	Interview score
SVR (poly)	0.8540	0.8329	0.8624	0.8402	0.8744	0.8528	0.8319
SVR (rbf)	0.8967	0.8844	0.8932	0.9012	0.8906	0.8932	0.8920
SVR (linear)	0.8980	0.8846	0.8935	0.9025	0.8920	0.8941	0.8945
DTR	0.8541	0.8411	0.8542	0.8610	0.8453	0.8511	0.8511
Transformer	0.8972	0.8814	0.8920	0.9035	0.8907	0.8930	0.8915
Bi-LSTM	0.8986	0.8834	0.8932	0.9045	0.8928	0.8945	0.8947
Transformer + Bi-LSTM	**0.8989**	**0.8847**	**0.8938**	**0.9048**	**0.8935**	**0.8952**	**0.8953**

Bold values denote the highest performance.

4.3. Experimental results and analysis

In this section, two groups of experiments are carried out on the ChaLearn First Impression-V2 data set to verify the effectiveness of all used methods. One is the single-modal personality trait recognition, the other is multi-modal personality trait recognition.

4.3.1. Results of single-modal personality trait recognition

For single-modal personality recognition, we present the experiment results and analysis based on the single extracted audio features, visual scene features, and visual face features by using the pre-trained deep models.

Table 1 shows the prediction results of deep audio features extracted by the pre-trained VGGish for different methods. “Transformer + Bi-LSTM” denotes that the learned features with Transformer and Bi-LSTM are directly concatenated to form a whole feature vector as inputs of the latter linear regression layer for prediction. It can be seen from Table 1 that Transformer + Bi-LSTM performs best based on deep audio features. More specially, the average Big-Five personality prediction score is 0.8952 and the corresponding interview prediction score of 0.8953, thereby outperforming other used methods. The ranking order for other used methods is Bi-LSTM, SVR (linear), SVR (rbf), Transformer, SVR (poly), and DTR. This shows the advantages of Transformer + Bi-LSTM on audio personality trait recognition tasks. It is noted that Transformer + Bi-LSTM performs better than Transformer and Bi-LSTM, indicating that there is a certain complementarity between Transformer and Bi-LSTM.

Tables 2, 3 separately present personality prediction results of deep visual scene features and deep visual face features extracted by the pre-trained VGG-Face for different methods. It can be observed from Tables 2, 3 that Transformer + Bi-LSTM still obtains better performance other methods. In particular, Transformer + Bi-LSTM employs deep visual scene features and face features to produce the average Big-Five personality prediction scores of 0.9039 and 0.9124, respectively, and the interview prediction scores of 0.9057 and 0.9163, respectively. The ranking order for other used methods is Bi-LSTM, Transformer, SVR (poly), SVR (linear), SVR (rbf), and DTR. This shows the superiority of Transformer + Bi-LSTM on deep visual (scene and face) personality trait recognition tasks. The visual face images outperforms the visual scene images on personality trait recognition tasks. This may be because face images are more correlated with personality traits than scene images.

TABLE 2 Prediction results of deep visual scene features extracted by the pre-trained VGG-Face for different methods.

Models	O	C	E	A	N	Average score	Interview score
SVR (poly)	0.8921	0.8896	0.8896	0.8962	0.8850	0.8905	0.8890
SVR (rbf)	0.8841	0.8736	0.8804	0.8963	0.8780	0.8825	0.8818
SVR (linear)	0.8896	0.8872	0.8867	0.8922	0.8809	0.8873	0.8865
DTR	0.8636	0.8607	0.8627	0.8711	0.8586	0.8633	0.8639
Transformer	0.8941	0.8844	0.8909	0.9021	0.8884	0.8920	0.8920
Bi-LSTM	0.9042	0.9013	0.9012	0.9091	0.8993	0.9030	0.9050
Transformer + Bi-LSTM	**0.9043**	**0.9025**	**0.9035**	**0.9093**	**0.9000**	**0.9039**	**0.9057**

Bold values denote the highest performance.

TABLE 3 Prediction results of deep visual face features extracted by the pre-trained VGG-Face for different methods.

Models	O	C	E	A	N	Average score	Interview score
SVR (poly)	0.8871	0.8922	0.8923	0.8980	0.8855	0.8910	0.8963
SVR (rbf)	0.8841	0.8736	0.8804	0.8963	0.8780	0.8825	0.8818
SVR (linear)	0.8953	0.8922	0.8986	0.8974	0.8913	0.8950	0.8960
DTR	0.8714	0.8683	0.8702	0.8760	0.8674	0.8706	0.8721
Transformer	0.9023	0.9000	0.9029	0.9068	0.8968	0.9017	0.9017
Bi-LSTM	0.9103	**0.9155**	0.9129	0.9135	0.9085	0.9121	0.9161
Transformer + Bi-LSTM	**0.9110**	0.9148	**0.9130**	**0.9143**	**0.9087**	**0.9124**	**0.9163**

Bold values denote the highest performance.

TABLE 4 Comparisons of recognition results obtained by different methods.

Modality	O	C	E	A	N	Average score	Interview score
A	0.8989	0.8847	0.8938	0.9048	0.8935	0.8952	0.8953
S	0.9043	0.9025	0.9035	0.9093	0.9000	0.9039	0.9057
F	0.9110	0.9148	0.9130	0.9143	0.9087	0.9124	0.9153
A + S + F (EF)	0.9145	**0.9176**	0.9171	0.9158	0.9121	0.9154	0.9178
A + S + F (MF)	0.9151	0.9172	0.9156	0.9150	0.9123	0.9150	0.9180
A + S + F (LF)	**0.9167**	0.9163	**0.9176**	**0.9177**	**0.9150**	**0.9167**	**0.9200**

A, audio; S, scene; F, face; EF, early fusion; MF, model-level fusion; LF, late fusion. Bold values denote the highest performance.

In summary, the results in Tables 1–3 demonstrate that for single-modal personality recognition the visual face features perform best on personality trait and interview prediction tasks, followed by deep visual scene features and deep audio features. This shows that the facial images related to facial expression contain more discriminant information for personality trait recognition.

4.3.2. Results of multimodal personality trait recognition

For multimodal personality recognition tasks, we compare the performance of three typical multimodal information fusion methods, such as feature-level fusion, decision-level fusion, and model-level fusion. In feature-level fusion, the audio-visual global features learned by Bi-LSTM and Transformer networks, are concatenated into a whole feature vector as input of the linear regression layer for personality trait prediction. In this case, feature-level fusion is also called early fusion (EF). In model-level fusion (MF), the concatenated audio-visual global features are fed into a 4-layer full-collection layer network (1024-512-256-128) for personality trait prediction. In decision-level fusion, we adopt Eq. 12 to obtain the analytical solution of the optimal weight values in Eq. 6. In this case, decision-level fusion is also called late fusion (LF).

Table 4 presents the comparisons of recognition results obtained by different fusion methods such as EF, MF, and LF, as

TABLE 5 Comparisons with other existing methods.

References	Modality	Feature extraction	Fusion methods	Average score
Güçlütürk et al., 2016	Audio, visual	Audio:ResNet-17 Visual:ResNet-17	EF	0.9109
Güçlütürk et al., 2017	Audio, visual, text	Audio:ResNet-17 Visual:ResNet-17 Text:skip-thought vectors	EF	0.9118
Wei et al., 2017	Audio, visual	Audio:MFCCs Visual:DAN	LF	0.9130
Principi et al., 2021	Audio, visual	Audio:1D CNN Visual:ResNet-50	MF	0.9160
Escalante et al., 2022	Audio, visual, text	Audio:ResNet-18 Visual:ResNet-18 Text: skip-thought vectors	LF	0.9161
Ours	Audio, visual	Audio:VGGish Visual:VGG-Face	LF	**0.9167**

EF, early fusion; MF, model-level fusion; LF, late fusion. Bold values denote the highest performance.

well as the single modality methods. From the results in **Table 4**, we can see that: (1) among three used fusion methods, the used LF method combining audio, scene, and face obtains the best performance with an average score of 0.9167 on personality trait recognition tasks, and an average score of 0.9200 on interview prediction tasks. For personality trait recognition, the used EF method slightly outperforms the MF method, yielding an average score of 0.9154. By contrast, the used MF method slightly outperforms the EF method on interview prediction tasks. In particular, the MF method gives an average interview score of 0.9180. (2) All used fusion methods such as LF, MF, and EF provide superior performance to the single modality methods. This indicates the complementarity to some extent among audio, scene, and face modality on target recognition tasks.

4.3.3. Comparisons with other existing methods

To further verify the effectiveness of the proposed method, **Table 5** presents the comparisons of different used methods. **Table 5** shows that the proposed method obtains an average score of 0.9167, which is better than other reported results obtained by audio, visual, and text modalities. This demonstrates the advantage of our method on personality trait recognition tasks. These comparing works are described as follows.

Güçlütürk et al. (2016) provided an audio-visual personality trait recognition based on 17-layer deep residual networks (ResNet-17). They concatenated the learned features of audio-visual streams at feature-level as an input of a fully connected layer and reported an average score of 0.9109 for final personality trait prediction. In this case, the used network does not need any feature engineering or visual analysis like face detection, face landmark alignment. Similarly, they also presented an multimodal personality trait analysis integrating audio, visual, and text modalities by using the 17-layer deep residual networks (Güçlütürk et al., 2017). Here, they extracted skip-thought vectors as text features. They fused these modalities at feature-level and reported an average score of 0.9118. Wei et al. (2017) presented a deep bimodal regression method of personality traits on short video sequences. For audio modality, they extracted MFCCs and logfbank features. For visual modality, they employed a modified CNN model called Descriptor Aggregation Network (DAN) to extract visual features. Finally, they fused these predicted regression scores of audio-visual modalities at decision-level, and reported an average score of 0.9130. Principi et al. (2021) presented a multimodal deep learning method integrating the raw audio and visual modalities for personality trait prediction. For audio modality, a 14-layer 1D CNN was used for audio feature extraction. For visual modality, they employed a pre-trained ResNet-50 network for visual feature extraction. Finally, they employed a fully connected layer to jointly learn audio-visual feature representations at model-level for final personality trait recognition, and achieved an average score of 0.9160. Escalante et al. (2022) proposed a multimodal deep personality trait recognition method based on audio, visual, and text modalities. They adopted a ResNet-18 to extract audio and visual features, and skip-thought vectors as text features. Then, a late fusion strategy was utilized to fuse all three modalities, and yielded an average score of 0.9161.

5. Conclusion

This paper presents a multimodal personality trait recognition method based on CNN + Bi-LSTM + Transformer network. In this work, CNN, Bi-LSTM, and Transformer are combined to capture high-level audio-visual spatio-temporal feature representations for personality trait recognition. Finally, we compare multimodal personality prediction results based on three different fusion methods such as feature-level fusion, model-level fusion, and decision-level fusion. Experiments on the public ChaLearn First Impression-V2 dataset show that decision-level fusion achieves the best multimodal personality trait recognition results with an average score of 0.9167, outperforming other existing methods.

It is noted that this work only focuses on integrating audio and visual modalities for multimodal personality trait recognition. Considering the diversity of modal information related to the expression of personality traits, it is interesting to combine current audio-visual modalities with other modalities such as physiological signals, text cues, etc., to further improve the performance of personality trait recognition. In addition, exploring a more advanced deep learning model for personality trait recognition is also an important direction in our future work.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://chalearnlap.cvc.uab.cat/dataset/24/description/.

Author contributions

XZ contributed to the writing and drafted the article. YL, ZT, YX, XT, DW, and GW contributed to the data preprocessing and analysis, software and experiment simulation. HL contributed to the project administration and writing—reviewing and editing. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by National Natural Science Foundation of China (NSFC) and Zhejiang Provincial Natural Science Foundation of China under Grant Nos. 61976149, 62276180, LZ20F020002, and LQ21F020002.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

An, G., Levitan, S. I., Levitan, R., Rosenberg, A., Levine, M., and Hirschberg, J. (2016). "Automatically classifying self-rated personality scores from speech," in *Proceedings of the INTERSPEECH Conference 2016*, (Incheon: ISCA), 1412–1416. doi: 10.21437/Interspeech.2016-1328

Bathurst, K., Gottfried, A. W., and Gottfried, A. E. (1997). Normative data for the MMPI-2 in child custody litigation. *Psychol. Assess.* 9:205. doi: 10.1037/1040-3590.9.3.205

Beyan, C., Zunino, A., Shahid, M., and Murino, V. (2021). Personality traits classification using deep visual activity-based nonverbal features of key-dynamic images. *IEEE Trans. Affect. Comput.* 12, 1084–1099. doi: 10.1109/TAFFC.2019.2944614

Costa, P. T., and McCrae, R. R. (1998). "Trait theories of personality," in *Advanced Personality*, eds D. F. Barone, M. Hersen, and V. B. Hasselt (Cham: Springer), 103–121. doi: 10.1007/978-1-4419-8580-4_5

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "Imagenet: a large-scale hierarchical image database," in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, 248–255. doi: 10.1109/CVPR.2009.5206848

Elman, J. L. (1990). Finding structure in time. *Cogn. Sci.* 14, 179–211. doi: 10.1207/s15516709cog1402_1

Escalante, H. J., Guyon, I., Escalera, S., Jacques, J., Madadi, M., Baró, X., et al. (2017). "Design of an explainable machine learning challenge for video interviews," in *Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN)*, (Piscataway, NJ: IEEE), 3688–3695. doi: 10.1109/IJCNN.2017.7966320

Escalante, H. J., Kaya, H., Salah, A. A., Escalera, S., Güçlütürk, Y., Güçlü, U., et al. (2022). Modeling, recognizing, and explaining apparent personality from videos. *IEEE Trans. Affect. Comput.* 13, 894–911. doi: 10.1109/TAFFC.2020.2973984

Furnham, A. (1996). The big five versus the big four: the relationship between the Myers-Briggs Type Indicator (MBTI) and NEO-PI five factor model of personality. *Pers. Individ. Diff.* 21, 303–307. doi: 10.1016/0191-8869(96)00033-5

Gao, J., Li, P., Chen, Z., and Zhang, J. (2020). A survey on deep learning for multimodal data fusion. *Neural Comput.* 32, 829–864. doi: 10.1162/neco_a_01273

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., et al. (2017). "Audio set: an ontology and human-labeled dataset for audio events," in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Piscataway, NJ: IEEE), 776–780. doi: 10.1109/ICASSP.2017.7952261

Güçlütürk, Y., Güçlü, U., Baro, X., Escalante, H. J., Guyon, I., Escalera, S., et al. (2017). Multimodal first impression analysis with deep residual networks. *IEEE Trans. Affect. Comput.* 9, 316–329. doi: 10.1109/TAFFC.2017.2751469

Güçlütürk, Y., Güçlü, U., Van Gerven, M. A., and Van Lier, R. (2016). "Deep impression: audiovisual deep residual networks for multimodal apparent personality trait recognition," in *Proceedings of the European Conference on Computer Vision*, (Cham: Springer), 349–358. doi: 10.1007/978-3-319-49409-8_28

Guntuku, S. C., Qiu, L., Roy, S., Lin, W., and Jakhetiya, V. (2015). "Do others perceive you as you want them to? Modeling personality based on selfies," in *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia, Association for Computing Machinery*, (New York, NY), 21–26. doi: 10.1145/2813524.2813528

Gürpınar, F., Kaya, H., and Salah, A. A. (2016). "Combining deep facial and ambient features for first impression estimation," in *Proceedings of the European Conference on Computer Vision*, (Berlin: Springer), 372–385. doi: 10.1007/978-3-319-49409-8_30

Hayat, H., Ventura, C., and Lapedriza, À (2019). On the use of interpretable CNN for personality trait recognition from audio. *CCIA* 319, 135–144.

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., et al. (2017). "CNN architectures for large-scale audio classification," in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Piscataway, NJ: IEEE), 131–135. doi: 10.1109/ICASSP.2017.7952132

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Junior, J. C. S. J., Güçlütürk, Y., Pérez, M., Güçlü, U., Andujar, C., Baró, X., et al. (2019). First impressions: a survey on vision-based apparent personality trait analysis. *IEEE Trans. Affect. Comput.* 13, 75–95. doi: 10.1109/TAFFC.2019.2930058

Karson, S., and O'Dell, J. W. (1976). *A Guide to The Clinical Use of the 16 PF.* Chandigarh: Inst for Personality & Ability Test.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, eds I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et al. (Cambridge, MA: MIT Press), 1097–1105.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Liang, Y., Li, S., Yan, C., Li, M., and Jiang, C. (2021). Explaining the black-box model: a survey of local interpretation methods for deep neural networks. *Neurocomputing* 419, 168–182. doi: 10.1016/j.neucom.2020.08.011

McCrae, R. R., and John, O. P. (1992). An introduction to the five-factor model and its applications. *J. Personal.* 60, 175–215. doi: 10.1111/j.1467-6494.1992.tb00970.x

Mohammadi, G., and Vinciarelli, A. (2012). Automatic personality perception: prediction of trait attribution based on prosodic features. *IEEE Trans. Affect. Comput.* 3, 273–284. doi: 10.1109/T-AFFC.2012.5

Montag, C., and Davis, K. L. (2018). Affective neuroscience theory and personality: an update. *Personal. Neurosci.* 1:e12. doi: 10.1017/pen.2018.10

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). "Deep face recognition," in *Proceedings of the British Machine Vision Conference (BMVC)*, Aberystwyth, 411–412. doi: 10.5244/C.29.41

Ponce-López, V., Chen, B., Oliu, M., Corneanu, C., Clapés, A., Guyon, I., et al. (2016). "Chalearn lap 2016: first round challenge on first impressions-dataset and results," in *Proceedings of the European Conference on Computer Vision*, (Berlin: Springer), 400–418. doi: 10.1007/978-3-319-49409-8_32

Prechelt, L. (1998). Automatic early stopping using cross validation: quantifying the criteria. *Neural Netw.* 11, 761–767. doi: 10.1016/S0893-6080(98)00010-0

Principi, R. D. P., Palmero, C., Junior, J. C., and Escalera, S. (2021). On the effect of observed subject biases in apparent personality analysis from audio-visual signals. *IEEE Trans. Affect. Comput.* 12, 607–621. doi: 10.1109/TAFFC.2019.2956030

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., et al. (2013). "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings of the INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*, Lyon. doi: 10.21437/Interspeech.2013-56

Schuster, M., and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* 45, 2673–2681. doi: 10.1109/78.650093

Teijeiro-Mosquera, L., Biel, J.-I., Alba-Castro, J. L., and Gatica-Perez, D. (2014). What your face vlogs about: expressions of emotion and big-five traits impressions in YouTube. *IEEE Trans. Affect. Comput.* 6, 193–205. doi: 10.1109/TAFFC.2014.2370044

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems*, eds I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et al. (Cambridge, MA: MIT Press), 5998–6008.

Vinciarelli, A., and Mohammadi, G. (2014). A survey of personality computing. *IEEE Trans. Affect. Comput.* 5, 273–291. doi: 10.1109/TAFFC.2014.2330816

Wang, D., and Zhao, X. (2022). Affective video recommender systems: a survey. *Front. Neurosci.* 16:984404. doi: 10.3389/fnins.2022.984404

Wang, M., and Deng, W. (2021). Deep face recognition: a survey. *Neurocomputing* 429, 215–244. doi: 10.1016/j.neucom.2020.10.081

Wei, X. S., Zhang, C. L., Zhang, H., and Wu, J. X. (2017). Deep bimodal regression of apparent personality traits from short video sequences. *IEEE Trans. Affect. Comput.* 9, 303–315. doi: 10.1109/TAFFC.2017.2762299

Xu, J., Tian, W., Lv, G., Liu, S., and Fan, Y. (2021). Prediction of the big five personality traits using static facial images of college students with different academic backgrounds. *IEEE Access* 9, 76822–76832. doi: 10.1109/ACCESS.2021.3076989

Yan, A., Chen, Z., Zhang, H., Peng, L., Yan, Q., Hassan, M. U., et al. (2021). Effective detection of mobile malware behavior based on explainable deep neural network. *Neurocomputing* 453, 482–492. doi: 10.1016/j.neucom.2020.09.082

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., and Hoi, S. C. (2021). Deep learning for person re-identification: a survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intellig.* 44, 2872–2893. doi: 10.1109/TPAMI.2021.3054775

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* 23, 1499–1503. doi: 10.1109/LSP.2016.2603342

Zhang, S., Zhao, X., and Tian, Q. (2022). Spontaneous speech emotion recognition using multiscale deep convolutional LSTM. *IEEE Trans. Affect. Comput.* 13, 680–688. doi: 10.1109/TAFFC.2019.2947464

Zhao, S., Gholaminejad, A., Ding, G., Gao, Y., Han, J., and Keutzer, K. (2019). Personalized emotion recognition by personality-aware high-order learning of physiological signals. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 1–18. doi: 10.1145/3233184

Zhao, X., Tang, Z., and Zhang, S. (2022). Deep personality trait recognition: a survey. *Front. Psychol.* 13:839619. doi: 10.3389/fpsyg.2022.839619

Zhu, M., Xie, X., Zhang, L., and Wang, J. (2018). "Automatic personality perception from speech in mandarin," in *Proceedings of the 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Taipei, 309–313. doi: 10.1109/ISCSLP.2018.8706692

TYPE Original Research
PUBLISHED 09 January 2023
DOI 10.3389/fpsyg.2022.1075624

Speech emotion recognition based on improved masking EMD and convolutional recurrent neural network

Congshan Sun, Haifeng Li* and Lin Ma

Faculty of Computing, Harbin Institute of Technology, Harbin, China

Speech emotion recognition (SER) is the key to human-computer emotion interaction. However, the nonlinear characteristics of speech emotion are variable, complex, and subtly changing. Therefore, accurate recognition of emotions from speech remains a challenge. Empirical mode decomposition (EMD), as an effective decomposition method for nonlinear non-stationary signals, has been successfully used to analyze emotional speech signals. However, the mode mixing problem of EMD affects the performance of EMD-based methods for SER. Various improved methods for EMD have been proposed to alleviate the mode mixing problem. These improved methods still suffer from the problems of mode mixing, residual noise, and long computation time, and their main parameters cannot be set adaptively. To overcome these problems, we propose a novel SER framework, named IMEMD-CRNN, based on the combination of an improved version of the masking signal-based EMD (IMEMD) and convolutional recurrent neural network (CRNN). First, IMEMD is proposed to decompose speech. IMEMD is a novel disturbance-assisted EMD method and can determine the parameters of masking signals to the nature of signals. Second, we extract the 43-dimensional time-frequency features that can characterize the emotion from the intrinsic mode functions (IMFs) obtained by IMEMD. Finally, we input these features into a CRNN network to recognize emotions. In the CRNN, 2D convolutional neural networks (CNN) layers are used to capture nonlinear local temporal and frequency information of the emotional speech. Bidirectional gated recurrent units (BiGRU) layers are used to learn the temporal context information further. Experiments on the publicly available TESS dataset and Emo-DB dataset demonstrate the effectiveness of our proposed IMEMD-CRNN framework. The TESS dataset consists of 2,800 utterances containing seven emotions recorded by two native English speakers. The Emo-DB dataset consists of 535 utterances containing seven emotions recorded by ten native German speakers. The proposed IMEMD-CRNN framework achieves a state-of-the-art overall accuracy of 100% for the TESS dataset over seven emotions and 93.54% for the Emo-DB dataset over seven emotions. The IMEMD alleviates the mode mixing and obtains IMFs with less noise and more physical meaning with significantly improved efficiency. Our IMEMD-CRNN framework significantly improves the performance of emotion recognition.

KEYWORDS

speech emotion recognition, empirical mode decomposition, mode mixing, convolutional neural networks, bidirectional gated recurrent units

OPEN ACCESS

EDITED BY
Xiaobai Li,
University of Oulu,
Finland

REVIEWED BY
Maximos Kaliakatsos-Papakostas,
Athena Research Center,
Greece
Serap Aydin,
Hacettepe University,
Turkey

*CORRESPONDENCE
Haifeng Li
✉ lihaifeng@hit.edu.cn

SPECIALTY SECTION
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Psychology

RECEIVED 20 October 2022
ACCEPTED 16 December 2022
PUBLISHED 09 January 2023

CITATION
Sun C, Li H and Ma L (2023) Speech emotion recognition based on improved masking EMD and convolutional recurrent neural network.
Front. Psychol. 13:1075624.
doi: 10.3389/fpsyg.2022.1075624

1. Introduction

Emotion is a kind of physiological and psychological state (Liu Z. et al., 2022). Physiological stimulation, subjective experience, and facial and behavioral expression all work together to form a complete emotional process (Nitsche et al., 2012; Lu et al., 2021). Basic emotional states comprise anger, disgust, fear, happiness, sadness, and surprise (Ekman and Friesen, 1971). The remaining emotions are combinations of these basic emotions, such as excitement, embarrassment, and contempt (Krishnan et al., 2021). Reliable analysis, recognition, understanding, and expression of emotions are significant for communicating and understanding information between humans and computers.

Attempts utilizing separate modalities have been made to recognize emotions (Aydın et al., 2018; Dominguez-Jimenez et al., 2020; Li et al., 2020a,b). Accumulating evidence have proved the efficiencies of EEG and other physiological signals (such as electrocardiograph, galvanic skin response, and respiration) in emotion recognition (Quan et al., 2021; Chen et al., 2022). In these experiments, physiological signals were simultaneously recorded while subjects were presented with diversified emotional stimulus materials (such as static pictures, facial expressions, video film clips, and acoustic music clips) that induced specific emotions, among which the parameters of these stimulus materials would also influence the intensity of induced emotions (Kılıç and Aydın, 2022). For emotion recognition, emotional features of EEG signals usually include power spectrum density (PSD), differential entropy (DE), rational asymmetry (RASM), differential entropy asymmetry (DASM), phase locking value (PLV), and phase lag index (PLI; Lu et al., 2021). For other physiological signals, some statistical features based on temporal or frequency-domain information are usually extracted for emotion recognition (Picard et al., 2001; Goshvarpour et al., 2017).

Speech is one of the most natural and intuitive ways of emotional communication, which contains rich emotions while conveying information (Li et al., 2020a). Speech emotion recognition (SER) is a computer simulation of human speech emotion perception and understanding, a key prerequisite for human-computer interaction. There are three main methods for emotional corpora collection: collecting natural speech from the real world (natural speech database), collecting audio recordings of subjects acting based on pre-decided affect-related scripts (actor-based speech database), and collecting corpora from the speaker by creating an artificial emotional situation (elicited emotional speech database; Basu et al., 2017). Emotional features of speech signals include prosody features, spectral features, and timbre features (Li et al., 2020b). The current SER is mainly supervised pattern recognition. Commonly used machine learning algorithms include k-nearest neighbor (KNN), support vector machine (SVM), linear discriminative analysis (LDA), Gaussian naive Bayes, and artificial neural network (ANN).

With the development of deep learning, SER based on deep neural networks (DNNs) has begun to attract attention. These methods train deep-learning models for speech emotion recognition by taking the original emotional speech or hand-crafted features as the inputs and have achieved fruitful results (Anvarjon et al., 2020). Sarma et al. (2018) identified emotions from raw speech signals using an interleaving time-delay neural network (TDNN) with unidirectional long short-term memory (LSTM) and time-restricted attention mechanisms (TDNN-LSTM-attention). The results outperformed previously reported results on the IEMOCAP dataset (Busso et al., 2008). Wang et al. (2021) proposed a novel end-to-end SER architecture that stacked multiple transformer layers and used log Mel-filterbank energy features as the input. This method outperformed prior methods by a relatively 20% improvement on the IEMOCAP dataset. Deschamps-Berger et al. (2021) presented an end-to-end temporal CNN-BiLSTM network and extracted the spectrogram by short-term Fourier transform (STFT) as the input of the network. This method was evaluated on the IEMOCAP and CEMO datasets and obtained good results. Kim and Saurous (2018) used two CNN layers for local and global convolution, two LSTM layers for sequence learning, and 20 features from eGeMAPs (containing rhythmic, spectral, and timbre features) as inputs to the model. On the Emo-DB dataset, an unweighted accuracy of 88.9% was achieved. Wang et al. (2022) extracted traditional hand-crafted features from GeMAPS and deep automatic features from the VGGish model. Then, they proposed a multi-feature fusion and Multi-lingual fusion speech emotion recognition algorithm based on the recurrent neural network (RNN) with an improved local attention mechanism. The speech emotion recognition accuracy is improved when the dataset is small. Hou et al. (2022) proposed a collective multi-view relation network (CMRN) based on bidirectional gate recurrent units (Bi-GRU) and the attention mechanism. In the CMRN, Mel-frequency cepstral coefficients (MFCCs), log Mel-frequency spectral coefficients (MFSCs), and prosody features are collected as multi-view representations. The proposed method performs better than the state-of-the-art methods on Emo-DB and IEMOCAP datasets.

For actual voice, automatic feature learning methods using deep networks can effectively learn the underlying patterns in the data. However, it is not easy to interpret the information obtained from these deep networks (Bhattacharjee et al., 2020). On the other hand, hand-crafted features used in deep-learning methods are mainly extracted based on the STFT. In practical applications, speech signals are non-stationary amplitude modulated-frequency modulated (AM-FM) signals with rich frequency components and temporal rhythm variations (Hsieh and Liu, 2019). The nonlinear features of speech emotion are variable, complex, and subtly changing (Kerkeni et al., 2019). However, limited by the fundamental uncertainty principle, the STFT cannot get good resolution in both time and frequency, and the non-linearity issue remains problematic (Kerkeni et al., 2019). Meanwhile, the STFT method requires pre-set basis functions and lacks adaptiveness in analyzing non-stationary speech (Yang et al., 2018). Therefore, reliable recognition of emotions from speech remains challenging.

More recently, empirical mode decomposition (EMD), a decomposition method for non-stationary AM-FM signals, has been used to analyze emotional speech signals. EMD adaptively decomposes a non-stationary signal into a finite number of intrinsic mode functions (IMFs) without losing the original properties of signals (Huang et al., 1998). IMFs have been shown to manifest the vocal tract structure and the glottal source information (Sharma et al., 2018; Karan et al., 2020). At the same time, experimental studies have shown that variations in the physiological properties of the vocal folds vary significantly across emotional patterns (Yao et al., 2020). Therefore, good results are obtained for speech emotion recognition based on EMD. Based on empirical mode decomposition (EMD) and Teager-Kaiser energy operator (TKEO), Kerkeni et al. (2019) extracted two new types of features. Combining these two feature sets with cepstral features, the unweighted accuracy using the support vector machine (SVM) on the Emo-DB dataset is 86.22%. Vieira et al. (2020) presented a novel Hilbert–Huang–Hurst coefficient (HHHC) feature based on the ensemble EMD (EEMD) to represent the emotional states. Experiments on different emotional datasets showed that HHHC led to significant classification improvements compared to the baseline acoustic features. Krishnan et al. (2021) extracted entropy features from principal IMFs based on EMD for recognizing emotions on the TESS dataset and the linear discriminant analysis (LDA) classifier presented a peak balanced accuracy of 93.3%. However, EMD and EEMD suffer from the mode mixing problem, which makes the physical meaning of IMF unclear (Rilling and Flandrin, 2008), thus reducing the performance of EMD-based methods for speech emotion recognition. Researchers have proposed several improvement methods for the mode mixing problem, such as the masking signal-based EMD (MSEMD; Deering and Kaiser, 2005), improved complete ensemble EMD with adaptive noise (ICEEMDAN; Colominas et al., 2014), uniform phase EMD (UPEMD; Wang et al., 2018), and robust EMD (REMD; Liu P. et al., 2022). Although these methods alleviate the modal aliasing problem to some extent, there are still problems in that the method parameters cannot be determined adaptively, there is residual noise in the IMFs, and the time complexity of the algorithm is high.

It is still challenging for computers to accurately capture emotional information in speech (Anvarjon et al., 2020). Therefore, this paper focuses on exploring and proposing an effective SER method to help computers develop advanced emotional intelligence. In this paper, we present a novel framework, named IMEMD-CRNN, to address the above challenges and improve speech-based emotion recognition performance.

The contributions of this work are three-fold: (i) We propose an improved version of the masking signal-based EMD (IMEMD). In the IMEMD, the parameters of masking signals are adaptively derived from the natures of the original signals. IMEMD obtains IMFs with less noise and more physical meaning with significantly improved efficiency. (ii) We use IMEMD to extract the timbre features proposed in our previous work (Li et al., 2020b) and Mel-frequency cepstral coefficients based on the reconstructed signal (SMFCC; Kerkeni et al., 2019) as the features used in the IMEMD-CRNN to characterize speech emotions. These are important speech emotion features (Guidi et al., 2019; Kerkeni et al., 2019). (iii) We feed the timbre features based on IMEMD into a convolutional recurrent neural network (CRNN) to recognize emotions. In the CRNN, we first use 2D CNN layers to capture nonlinear local temporal and frequency information of the emotional speech. Then, the outputs of the CNN module are fed to bidirectional gated recurrent units (BiGRU) layers to learn the temporal context information further. In the experimental part, we first demonstrated the advantages of IMEMD for decomposing non-stationary signals through the performance of the different improved algorithms for EMD in simulated and real speech emotion signals. Then experiments on two popular standard speech emotion datasets showed the significance and the robustness of our proposed IMEMD-CRNN framework for speech emotion recognition.

2. Materials and methods

In this section, our proposed IMEMD-CRNN to predict emotion is introduced. Figure 1 shows the framework of IMEMD-CRNN. As illustrated, IMEMD-CRNN consists of three modules: IMEMD-based emotional speech signal decomposition, extraction of time-frequency features from IMFs, and speech emotion recognition based on CRNN. Arano et al. (2021) show that effective hand-crafted features, compared to sophisticated deep-learning feature sets, can still have better performance. Therefore, we combine IMEMD-based features with CRNN network in order to improve the robustness and accuracy of the speech emotion recognition system. The framework of IMEMD-CRNN is shown in Figure 1. Design details of the three modules are introduced below.

2.1. Improved masking empirical mode decomposition

This part begins with a brief introduction to EMD and MSEMD, and the causes of mode mixing problems are analyzed. Then, we describe our proposed IMEMD.

2.1.1. The masking signal-based EMD

The EMD decomposes a non-stationary signal into a finite and often small number of IMFs and a residue (Huang et al., 1998). The IMFs contain progressively lower frequency components of the signal. The given signal $x(t)$ can be reconstructed as:

$$x(t) = \sum_{k=1}^{n_{imf}} c_k(t) + r_{es}(t) \qquad (1)$$

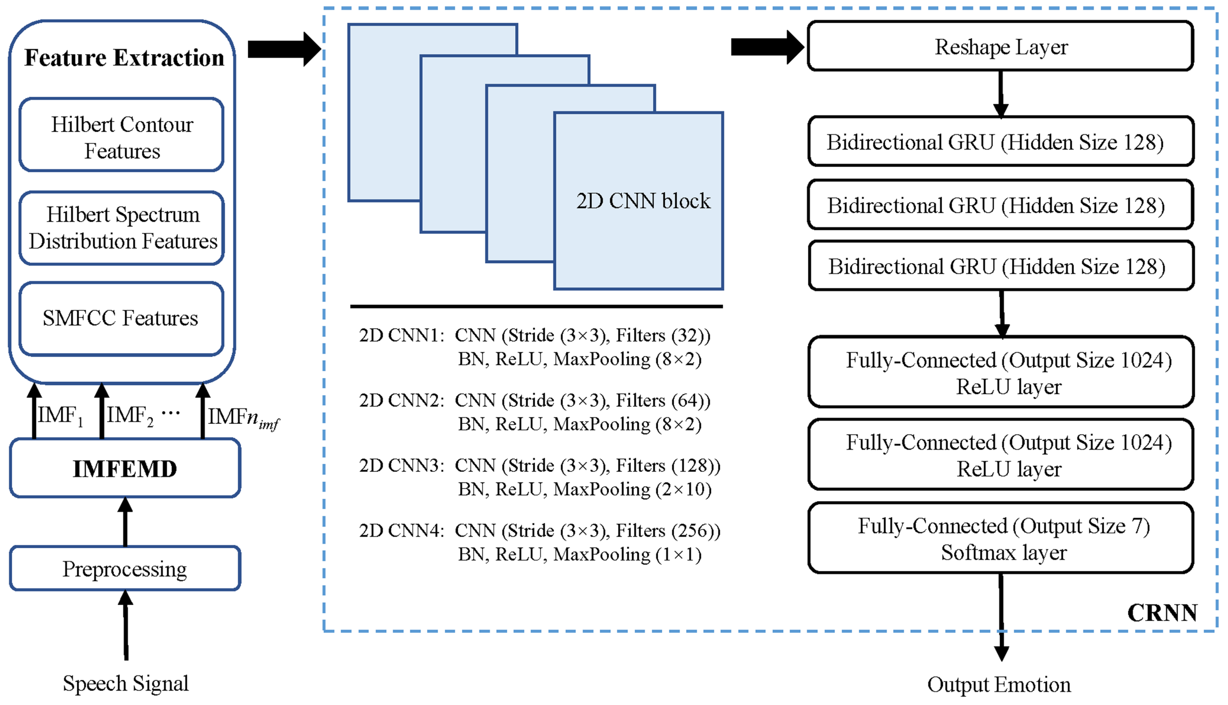

Feature Extraction
Hilbert Contour Features
Hilbert Spectrum Distribution Features
SMFCC Features
IMF$_1$
IMF$_2$ ···
IMFn_{imf}
IMFEMD
Preprocessing
Speech Signal
2D CNN block
2D CNN1: CNN (Stride (3×3), Filters (32)) BN, ReLU, MaxPooling (8×2)
2D CNN2: CNN (Stride (3×3), Filters (64)) BN, ReLU, MaxPooling (8×2)
2D CNN3: CNN (Stride (3×3), Filters (128)) BN, ReLU, MaxPooling (2×10)
2D CNN4: CNN (Stride (3×3), Filters (256)) BN, ReLU, MaxPooling (1×1)
Reshape Layer
Bidirectional GRU (Hidden Size 128)
Bidirectional GRU (Hidden Size 128)
Bidirectional GRU (Hidden Size 128)
Fully-Connected (Output Size 1024) ReLU layer
Fully-Connected (Output Size 1024) ReLU layer
Fully-Connected (Output Size 7) Softmax layer
CRNN
Output Emotion

FIGURE 1
Overall scheme of the IMEMD-CRNN. In the figure, every 2D CNN block (2D CNN) has 4 parts: a 2D CNN layer, a batch normalization layer (BN), a ReLU layer (ReLU), and a 2D max pooling layer (MaxPooling).

where $c_k(t)$ ($k = 1, \ldots, n_{imf}$) represents the kth IMF and $r_{es}(t)$ indicates the residue of the signal $x(t)$. The sifting process of EMD to obtain an IMF from $x(t)$ is as follows (Huang et al., 1998):

Step 1. Initialize $r(t) = x(t)$.

Step 2. Compute all local maxima and minima of $r(t)$.

Step 3. Interpolating the local maxima (minima) by the cubic spline to obtain the upper (lower) envelope $e_u(t)$ ($e_l(t)$) of $r(t)$.

Step 4. Compute the local mean envelope $e(t) = [e_u(t) + e_l(t)]/2$.

Step 5. Subtract $e(t)$ from $r(t)$ and update $r(t) = r(t) - e(t)$.

Step 6. Repeat steps 2 to 5 until $r(t)$ meets the conditions of IMF.

The mode mixing is that the IMF may contain widely distributed scales (Wu and Huang, 2009). Figures 2C–F show the mode mixing. The mode mixing is mainly caused by the following two situations: (i) intermittency caused by intermittent signal, pulse interference, and noise and (ii) different frequency components of the signal lying within an octave (Deering and Kaiser, 2005; Rilling and Flandrin, 2008). Therefore, many improved algorithms for EMD have been proposed to solve the mode mixing problem. Deering et al. first proposed using masking signals to resolve the mode mixing in EMD (Deering and Kaiser, 2005). The method is called the masking signal-based EMD (MSEMD), which uses a sinusoid signal $x_m(t)$ as the masking signal. The process of obtaining an IMF by MSEMD is shown in Algorithm 1 (Shown in Table 1). Let EMD$_k$ (•) be the operator, which produces the kth IMF using EMD. The β, f_w, and θ represent the amplitude, frequency, and phase of the masking signal, respectively. Their detailed computational process is shown in reference (Deering and Kaiser, 2005). MSEMD has high computational efficiency and can solve mode mixing to some extent, but the parameter selection methods of the masking signal need to be further improved.

2.1.2. The proposed IMEMD

In this section, we propose a novel method to construct masking signals to alleviate mode mixing. Since our proposed method is an improved version of the MSEMD, it is called improved masking EMD (IMEMD). In IMEMD, obtaining the highest frequency component of the original signal is as follows: First, a masking signal whose frequency is higher than the highest frequency component of the original signal is added to the original signal. Next, the signal is decomposed by EMD, and the first IMF obtained contains the highest frequency component and the masking signal. Then, the masking signal is removed from this IMF to obtain the highest frequency component. The proposed IMEMD is given in Algorithm 2 (Table 2). The value of ε_1 ($\varepsilon_1 = 30$ dB) is referred to as reference (Liu et al., 2017), where ε_1 is the decomposition stop threshold.

In Section 2.1.1, we analyze two main reasons for mode mixing: the intermittent components in the signal and the components whose frequencies are within an octave. By adding an appropriate sinusoidal signal (The duration is equal to the original signal) to the original signal, the extrema of the new signal are more uniformly distributed. Thus, the mode mixing due to intermittent components can be alleviated (Wang et al., 2018). At the same time, adding the sinusoidal signal improves the

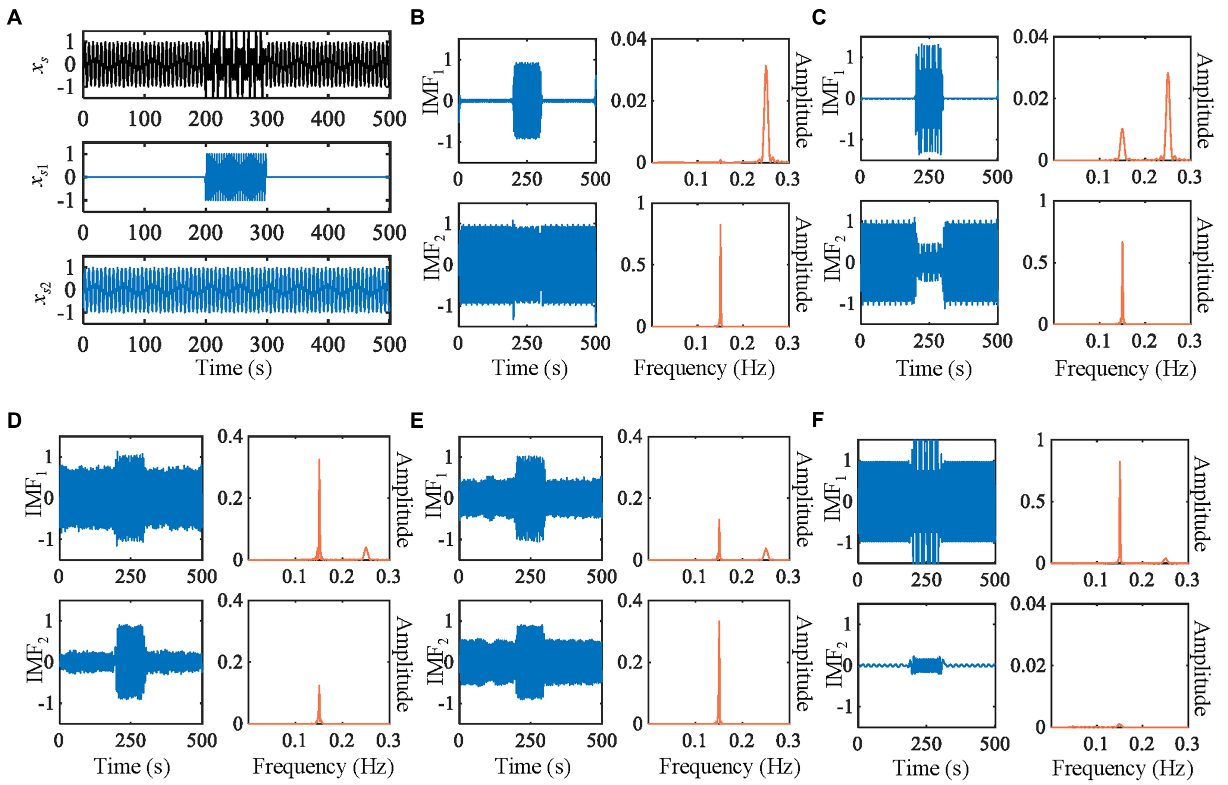


FIGURE 2
Decomposition of the synthetic signal by five methods. **(A)** The waveforms of synthetic signals. **(B)** IMEMD. **(C)** UPEMD. **(D)** EEMD. **(E)** ICEEMDAN. **(F)** REMD. In each subgraph of **(B–F)**, the left is waveforms of IMFs, the right is power spectra of IMFs.

TABLE 1 The algorithm to obtain an IMF by MSEMD.

Algorithm 1 Obtaining an IMF by MSEMD	
Function: $c(t) = \text{MSEMD}\ (x(t))$	
Input: $x(t)$	
Output: $c(t)$	
1:	Construct a masking signal $x_{ms}(t) = \beta \sin\ (2\pi f_w t + \theta)$
2:	Compute $c_+(t) = \text{EMD}_1(x(t) + x_{ms}\ (t))$
3:	Compute $c_-(t) = \text{EMD}_1(x(t) - x_{ms}\ (t))$
4:	$c(t) = (c_+(t) + c_-(t))/2$

TABLE 2 The algorithm of IMEMD.

Algorithm 2 IMEMD	
Function: $\{c_k(t)\} = \text{IMEMD}\ (x(t))$	
Input: $x(t)$	
Output: $\{c_k(t)\}$	
1:	Initialize: n_p is the number of phases, $r_0(t) = x(t)$, $k = 1$
2:	While $\int \lvert x(t) \rvert^2 dt \,/\, \int \lvert r_{k-1}(t) \rvert^2 dt < \varepsilon_1$ and $r_{k-1}(t)$ is not monotonic Do
3:	$c_k(t) = \left(\sum_{i=1}^{n_p} \text{EMD}_1\left(r_{k-1}(t) + v_{ki}(t)\right) \right) / n_p$
4:	$r_k(t) = r_{k-1}(t) - c_k(t)$
5:	$k \leftarrow k + 1$
6:	End while
7:	$r_{es}(t) = r_{k-1}(t)$

filtering characteristics of the EMD for separating components whose frequencies lie in an octave (Xu et al., 2009). How to construct an appropriate masking signal is shown below:

Our proposed masking signal $v_{ki}(t)$ is represented as follows:

$$v_{ki}(t) = \xi_k \sin\left(2\pi \overline{f_k} t + 2\pi \frac{i-1}{n_p}\right) \tag{2}$$

where ξ_k and $\overline{f_k}$ are the amplitude and frequency of the kth masking signal $v_k(t)$, respectively. The parameter n_p is the number of phases ($n_p \in N$, $n_p > 1$) and $i = 1, 2, \ldots, n_p$.

In the proposed IMEMD, ξ_k and $\overline{f_k}$ are determined adaptively according to the nature of the signal, and they are calculated as follows:

$$\xi_k = \xi_0 \cdot \frac{\sum_{t=1}^{T} A_k(t)}{T} \tag{3}$$

$$f_k = \frac{\sum_{t=1}^{T} A_k(t) \cdot F_k(t)}{\sum_{t=1}^{T} A_k(t)} \tag{4}$$

$$\overline{f_k} = \begin{cases} f_k + f_k, k = 1 \\ f_{k-1} + f_k, k > 1 \end{cases} \tag{5}$$

where $A_k(t)$ and $F_k(t)$ are the instantaneous amplitude and frequency of the IMF obtained by $\text{EMD}_1(r_{k\text{-}1}(t))$, respectively. T is the duration of the signal and $f_{k-1} > f_k$. Following Huang et al. (1998), $A_k(t)$ and $F_k(t)$ are defined as

$$y_k(t) = \frac{1}{\pi} P \int_{-\infty}^{+\infty} \frac{c_k(\tau)}{\tau - t} d\tau \tag{6}$$

$$A_k(t) = \sqrt{c_k^2(t) + y_k^2(t)} \tag{7}$$

$$F_k(t) = \frac{1}{2\pi} \cdot \frac{d}{dt} \left(\arctan \frac{y_k(t)}{c_k(t)} \right) \tag{8}$$

where P indicates the Cauchy principal value integral, and $y_k(t)$ is the Hilbert transform (HT) of the kth IMF, $c_k(t)$.

Equations 2–8 describe the calculation of the frequency, amplitude, and phase of the mask signal in the proposed IMEMD. For the masking frequency, studies have shown that two components with a frequency ratio between 0.5 and 2 can be separated when the frequency of the mask signal is higher than the frequency of the high-frequency component (Senroy et al., 2007; Rilling and Flandrin, 2008). For signal $x(t)$, when its two adjacent frequency components $f_{\text{tr},k}$ and $f_{\text{tr},k+1}$ satisfy

$$1 < \frac{f_{tr,k}}{f_{tr,k+1}} < 2 \tag{9}$$

and the mode mixing occurs after the $\text{EMD}_1(x(t))$ operation, $f_k > f_{tr,k+1}\left(k = 1\right)$, hence $f_{tr,k} < 2\ f_k$. When $k > 1$ and the mode mixing occurs after the $\text{EMD}_1(r_{k\text{-}1}(t))$ operation, $f_{k-1} > f_{tr,k}$ $(k > 1)$, hence $f_{tr,k} < f_k + f_{k-1}$. So, the masking frequency $\overline{f_k}$ in Equation 5 still satisfies that the frequency of the mask signal is higher than the frequency of the high-frequency component. Wang et al. (2018) prove that the residual noise can be reduced by using a few sinusoidal signals with uniform phase distribution as masking signals. Therefore, in obtaining the kth IMF by IMEMD, we construct n_p mask signals whose phases are uniformly distributed over the 2π space. Then, the new signals after adding these n_p mask signals are decomposed by EMD, respectively, to obtain n_p IMFs. The mean of these n_p IMFs is used as the final kth IMF, which can reduce the residual of the mask signals in the decomposition results and decrease the decomposition error. The effect of the different number of phases on the signal reconstruction error is experimentally analyzed in Section 3.3.1. In the power quality detection task, the appropriate masking amplitude can be determined based on the amplitude of the frequency component obtained by fast Fourier transform (FFT) (Wu et al., 2014). Inspired by this, we use instantaneous amplitudes obtained based on the HT to construct masking amplitude. Since the HT-based instantaneous amplitudes are time-varying, we average all instantaneous amplitudes during T. In Equations 2, 3, the values of n_p ($n_p = 64$) and ξ_0 ($\xi_0 = 1.5$) are empirical. In Section 3.3, we test the effect of different values of n_p and ξ_0 on the IMF estimation.

2.2. Feature extraction based on IMEMD

In this section, we extract two feature sets for SER using IMEMD. The first feature set is the timbre features proposed in our previous work (Li et al., 2020b). Timbre features are proven to be essential features for SER (Guidi et al., 2019). The other feature set is the Mel-frequency cepstral coefficients based on the reconstructed signal (SMFCC), which has been proven effective in distinguishing different speech emotions (Kerkeni et al., 2019). The following are details of two feature sets used in IMEMD-CRNN. Table 3 shows the details of these two feature sets.

2.2.1. Timbre features based on IMEMD

IMEMD method is first adopted in this section to extract the intrinsic mode functions of speech. Then, the timbre feature sets, including the Hilbert spectrum distribution features and Hilbert contour features, are extracted.

TABLE 3 The feature sets extracted by IMEMD for SER.

Category	Feature name	Dimensions
Timbre features	Hilbert spectrum distribution features (*SC*, *SP*, *SK*, *SU*)	4
	Hilbert contour features (*SE*, Δ*SE*, Δ^2*SE*)	3
Spectral features	SMFCC	12
	First derivative of SMFCC (Δ*SMFCC*)	12
	Second derivative of SMFCC (Δ^2*SMFCC*)	12

For each frame of the signal, Hilbert spectrum distribution features are calculated as follows

$$SC = \frac{\sum_{k=1}^{n_{imf}} F_{ce}[k] \cdot E_{me}[k]}{\sum_{k=1}^{n_{imf}} E_{me}[k]} \quad (10)$$

$$SP = \sqrt{\frac{\sum_{k=1}^{n_{imf}} E_{me}[k] \cdot \left(F_{ce}[k] - SC\right)^2}{\sum_{k=1}^{n_{imf}} E_{me}[k]}} \quad (11)$$

$$SK = \frac{\sum_{k=1}^{n_{imf}} E_{me}[k] \cdot \left(F_{ce}[k] - SC\right)^3}{SP^3 \sum_{k=1}^{n_{imf}} E_{me}[k]} \quad (12)$$

$$SU = \frac{\sum_{k=1}^{n_{imf}} E_{me}[k] \cdot \left(F_{ce}[k] - SC\right)^4}{SP^4 \sum_{k=1}^{n_{imf}} E_{me}[k]} \quad (13)$$

where $F_{ce}[k]$ is the centroid frequency calculated for the instantaneous frequency of one frame in the *k*th IMF. $E_{me}[k]$ is the mean value of the instantaneous amplitude of one frame in the *k*th IMF.

For each frame of the signal, Hilbert contour features are calculated as follows:

$$SE = \max\left(E_{me}[k]\right) \quad (14)$$

$$\Delta SE(\varphi) = \begin{cases} SE(\varphi+1) - SE(\varphi), 1 \le \varphi \le Q \\ \dfrac{\sum_{q=1}^{Q} q\left(SE(\varphi+q) - SE(\varphi-q)\right)}{\sqrt{2\sum_{q=1}^{Q} q^2}}, Q < \varphi \le \Phi - Q \\ SE(\varphi) - SE(\varphi-1), \Phi - Q < \varphi \le \Phi \end{cases} \quad (15)$$

where Φ is the total number of frames of the signal. The second derivative $\Delta^2 SE$ can be solved by replacing the *SE* in the above equation with ΔSE where *Q* is the time difference of the first derivative, which is usually taken as 2.

2.2.2. Spectral features based on IMEMD

We extract the Mel-frequency cepstral coefficients based on the reconstructed signal (*SMFCC*) (Kerkeni et al., 2019) as the features to characterize speech emotions. The reconstructed signal is obtained by IMEMD. In order to improve the accuracy of speech emotion recognition, we also extract the first derivative of *SMFCC* ($\Delta SMFCC$) and the second derivative of *SMFCC* ($\Delta^2 SMFCC$). Because derivative features contain some temporal information, research show that this information is essential for speech emotion recognition (Kerkeni et al., 2019).

First, we use the zero-crossing rate detection method to find the signal trend $x_{tr}(t)$, as shown in Equation 16.

$$x_{tr} = \sum_{k} c_k(t), \text{if } \frac{ZeroCross_{c_k}(t)}{ZeroCross_{c_1}(t)} \left(k = 2,3,..,n_{imf}\right) \quad (16)$$

where $ZeroCross_{c_k}(t)$ is the zero-crossing rate. Then, $x_{tr}(t)$ is subtracted from the original signal, and the rest of the signal is used to reconstruct the original signal. The *SMFCC* is obtained by calculating the MFCCs with 12 orders of the reconstructed signal. Thus, for the reconstructed signal, the number of *SMFCC* coefficients returned per frame is 12; that is, the dimension of *SMFCC* features is 12.

The $\Delta SMFCC$ and $\Delta^2 SMFCC$ describe the trajectories of *SMFCC* over time. When the number of frames of the reconstructed signal is Φ, the first derivative of φ th frame $\Delta SMFCC(\varphi)$ is calculated as follows:

$$\Delta SMFCC(\varphi) = \begin{cases} SMFCC(\varphi+1) - SMFCC(\varphi), 1 \le \varphi \le Q \\ \dfrac{\sum_{q=1}^{Q} q \begin{pmatrix} SMFCC(\varphi+q) \\ -SMFCC(\varphi-q) \end{pmatrix}}{\sqrt{2\sum_{q=1}^{Q} q^2}}, Q < \varphi \le \Phi - Q \\ SMFCC(\varphi) - SMFCC(\varphi-1), \Phi - Q < \varphi \le \Phi \end{cases} \quad (17)$$

where *Q* is the time difference of the first derivative, which is usually taken as 2. The second derivative is calculated in the same way, but it is calculated from $\Delta SMFCC(\varphi)$, not *SMFCC*. Thus, the number of dimensions of $\Delta SMFCC$ and $\Delta^2 SMFCC$ features is also 12.

2.3. Convolutional recurrent neural network

The architecture of CRNN in this paper is based on Adavanne et al. (2019) and Cao et al. (2019). The CRNN contains three parts. The first part includes four 2D CNN blocks and a reshape layer. Each of these 2D CNN blocks consists of a batch normalization layer (BN), a ReLU layer (ReLU), and a 2D max pooling layer (MaxPooling). The second part has three bidirectional GRUs. The third part has three fully connected layers. The output layer uses the softmax activation function. The cross-entropy loss is used to train the network and is optimized using an Adam optimizer. We train the network for 60 epochs with a mini-batch size of 512. The initial learning rate η_0 is 0.001. The architectural details of CRNN are shown in Figure 1.

3. Results and discussion

3.1. Datasets

3.1.1. Synthetic signals

The synthetic signals to evaluate the performance of our IMEMD is a classical mode mixing example (shown in Figure 2). The synthetic signal $x_s(t)$ consists of a sustained pure tone $x_{s1}(t)$ and a gapped one $x_{s2}(t)$ with a higher frequency, where their frequencies lie within an octave. The data $x_s(t) = x_{s1}(t) + x_{s2}(t)$ is sampled at 1 Hz rate, $0 \le t \le 500$, with

$$x_{s1}(t) = \begin{cases} \sin\left(2\pi \cdot 0.25 \cdot (t-201)\right), 201 \le t \le 300 \\ 0, t < 201\, or\, t > 300 \end{cases} \tag{18}$$

$$x_{s2}(t) = \sin\left(2\pi \cdot 0.15 \cdot (t-1)\right), 0 \le t \le 500 \tag{19}$$

3.1.2. Public datasets

The IMEMD-CRNN system is validated on the Berlin Emotional Database (Emo-DB; Burkhardt et al., 2005) and Toronto Emotional Speech Set (TESS; Pichora-Fuller and Dupuis, 2020). They are the most popularly used databases for emotion recognition (Deb and Dandapat, 2019). Both datasets were approved by ethical committees. The Emo-DB dataset includes 535 audio files simulated by 10 actors on 10 German utterances. All files are in 16-bit stereo wave sampled at 16 kHz and labeled with one of the 7 emotions. The average duration of the utterances in this dataset is 3.5 s, and the approximate duration of the utterances is 3 s to 5 s. The number of emotional labels across the dataset is anger (127), anxiety/fear (69), boredom (81), disgust (46), happiness (71), neutral (79), and sadness (62). Audio files in the Emo-DB are single-channel audio.

The TESS database is recorded by two actresses aged 26 and 64. Both actresses speak English as their first language. There are 2,800 audio samples in the database, including seven different emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. There are 400 data samples for each emotion. The sampling rate is 24.414 kHz and is saved in WAV format with all audio samples between 2 s and 3 s in length. Audio files in the TESS are single-channel audio.

3.2. Preprocessing and evaluation metrics

Utterances in TESS and Emo-DB datasets are recorded in a noise-less environment; therefore, there is no need to filter and denoise the data (Krishnan et al., 2021). Utterances in the two datasets are split into equal-length segments of 3 s, and zero padding is used for utterances with a duration of less than 3 s (Chen et al., 2018). Each utterance is normalized by dividing the time-domain signal by its maximum value. For each utterance (sampling rate: 16 kHz for Emo-DB, 24.414 kHz for TESS), the frame size is uniformly set to 25 ms, and the hop size is 10 ms. To improve the performance of our IMEMD-CRNN architecture, we use data augmentation techniques to enlarge the size of the Emo-DB dataset, and every file is enlarged to 60 augmentations. We enlarge the Emo-DB dataset with three data enhancement methods: pitch shifting, time shifting, and noise addition. For pitch shifting, the range of pitch shift in semitones is [−2, 2]. The range of time shift in seconds is [−0.4, 0.4]. We use the Gaussian white noise addition, and the range of noise SNR in dB is [−20, 40]. Each audio is normalized by dividing the time-domain signal by its maximum value.

When evaluating our proposed IMEMD, the reconstruction error of the reconstructed signal $\tilde{x}$ relative to the original signal x is measured by the relative root mean square error (RRMSE), and the calculation formula is as follows:

$$\text{RRMSE}_x(\tilde{x}) = \frac{\sqrt{\sum_{n=1}^{N}(\tilde{x}_n - x_n)^2}}{\sqrt{\sum_{n=1}^{N} x_n{}^2}} \tag{20}$$

To compare with the state-of-the-art SER methods, we use unweighted accuracy (UA) to evaluate the performance of different SER methods (Zhong et al., 2020).

3.3. Performance of IMEMD

3.3.1. Simulations and comparisons

We compare the results of IMEMD with those of EEMD, UPEMD, ICEEMDAN, and REMD in Figure 2 through the decomposition of the artificial signal. We only show the first two IMFs of these methods as the mode mixing mainly occurs in the first two modes of the artificial signal. We set the noise standard deviation to 0.4, the ensemble size to 100, and phase number to 16 for EEMD, UPEMD, and ICEEMDAN, which are similar to those in Colominas et al. (2014) and Wang et al. (2018). For IMEMD, we set $n_p = 64$ and $\xi_0 = 1.5$ through experiments. The number of IMF obtained by IMEMD, REMD, UPEMD, ICEEMDAN, EEMD, and EMD is 2, 3, 8, 12, 14, and 14, respectively. In Figure 2, when separating components whose frequencies lie within an octave, the separation degree of each method from high to low is IMEMD > UPEMD > ICEEMDAN > EEMD > REMD> EMD. IMEMD substantially reduces the mode mixing. The proper value of ξ_0 greatly impacts the performance of IMEMD and in this work, ξ_0 is empirical. In Figure 3, three case studies are performed to show the effect of ξ_0 on mode estimation by IMEMD. The values of other parameters are the same as in Figure 2. Figures 3A–C show the decomposition of the synthetic signal by IMEMD when ξ_0 is taken as the most appropriate value, ξ_0 increase to a large value, and ξ_0 increase to a small value, respectively. As shown in Figure 3B, when the value of ξ_0 is too small, there are $x_{s1}(t)$ and

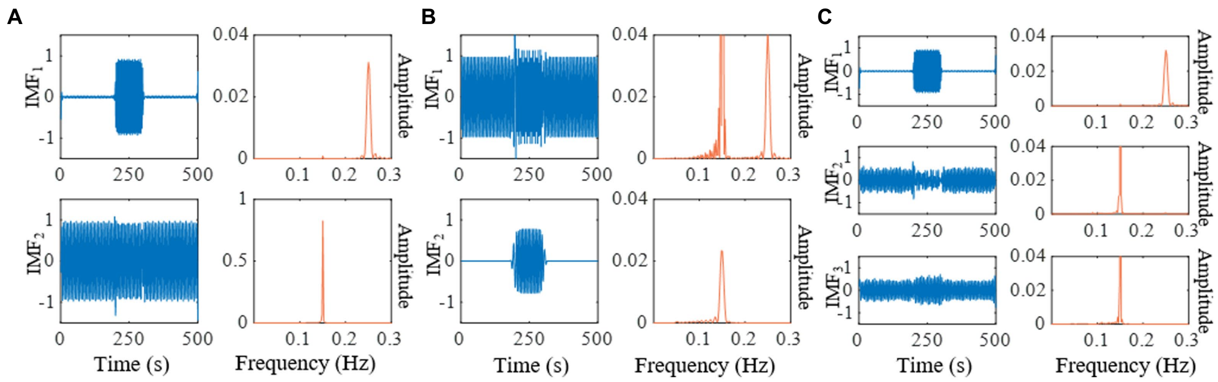


FIGURE 3
Decomposition of the synthetic signal by IMEMD. **(A)** The ξ_0 =1.5. **(B)** The ξ_0 =0.1. **(C)** The ξ_0 =3.

$x_{s2}(t)$ in IMF1. In Figure 3C, when the value of ξ_0 is too large, $x_{s2}(t)$ appears in IMF2 and IMF3. The results in Figures 3B,C are mode mixing. These mean that inappropriate values of ξ_0 can cause mode mixing problems.

In order to better compare the reconstruction errors of different methods in a different number of trials (the results are shown in Figure 4), we set the frequency of $x_{s2}(t)$ to 0.07. So, frequencies of $x_{s1}(t)$ and $x_{s2}(t)$ do not lie within an octave. Assisted signals with an amplitude of 0.2 are utilized for EEMD, ICEEMD, and UPEMD (Wang et al., 2018). Ensemble sizes of EEMD and ICEEMDAN are set to $I = 50$, 100, 200, 400, 600, and 800 (Wu and Huang, 2009; Colominas et al., 2014). Masking signals with phase numbers $n_p = 2$, 4, 8, 16, 32, and 64 are used in UPEMD and IMPEMD (Wang et al., 2018) to explore the effect of phase numbers on the decomposition results of the algorithms. Moreover, 10 sifting iterations are used to extract IMFs for all methods. In order to quantify the performance of the methods, all methods are decomposed 100 times to obtain the statistical average results (shown in Figure 4). Figure 4 shows that when $n_p > 32$, reconstruction errors (the value is 7.25×10^{-17}) of $x_s(t)$ by IMEMD are smaller than those of ICEEMDAN (the value is 7.38×10^{-17}). For all values of n_p, reconstruction errors of $x_s(t)$ by IMEMD are about one-tenth of the reconstruction errors of $x_s(t)$ by UPEMD. When the number of phases n_p ranges from 2 to 64, the reconstruction errors of $x_{s1}(t)$ and $x_{s2}(t)$ reconstructed by IMEMD have little changes, and the reconstruction errors of $x_s(t)$ reconstructed by IMEMD decrease. When $n_p = 64$, the reconstruction errors of these signals decomposed by IMEMD are small enough and smaller than these of the compared algorithms. Moreover, the time complexity of IMEMD is increasing as n_p increases, so we set the value of n_p in the IMEMD to 64. Reconstruction errors of $x_s(t)$ using EEMD are greater than 0.07. This may be because the signal contains a lot of residual noise. Therefore, the results of EEMD are not drawn in Figure 4A. Figures 4B,C plot errors of recovering $x_{s1}(t)$ and $x_{s2}(t)$, respectively. As shown in Figure 4, IMEMD is better than

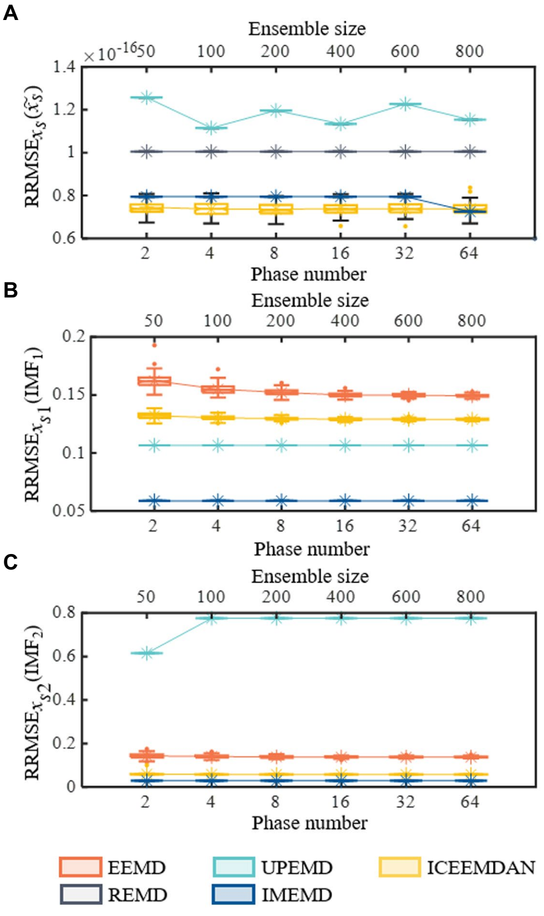


FIGURE 4
Performances of recovering known components on synthetic signal $x_s(t)$. All five methods are decomposed 100 times to obtain the statistical average results and shown using boxplots. **(A)** Reconstruction errors of synthetic signal $x_s(t)$. **(B)** Performances of recovering $x_{s1}(t)$. **(C)** Performances of recovering $x_{s2}(t)$. In each subgraph of **(A–C)**, the symbol "*" represents the mean value of the corresponding 100 decomposition results.

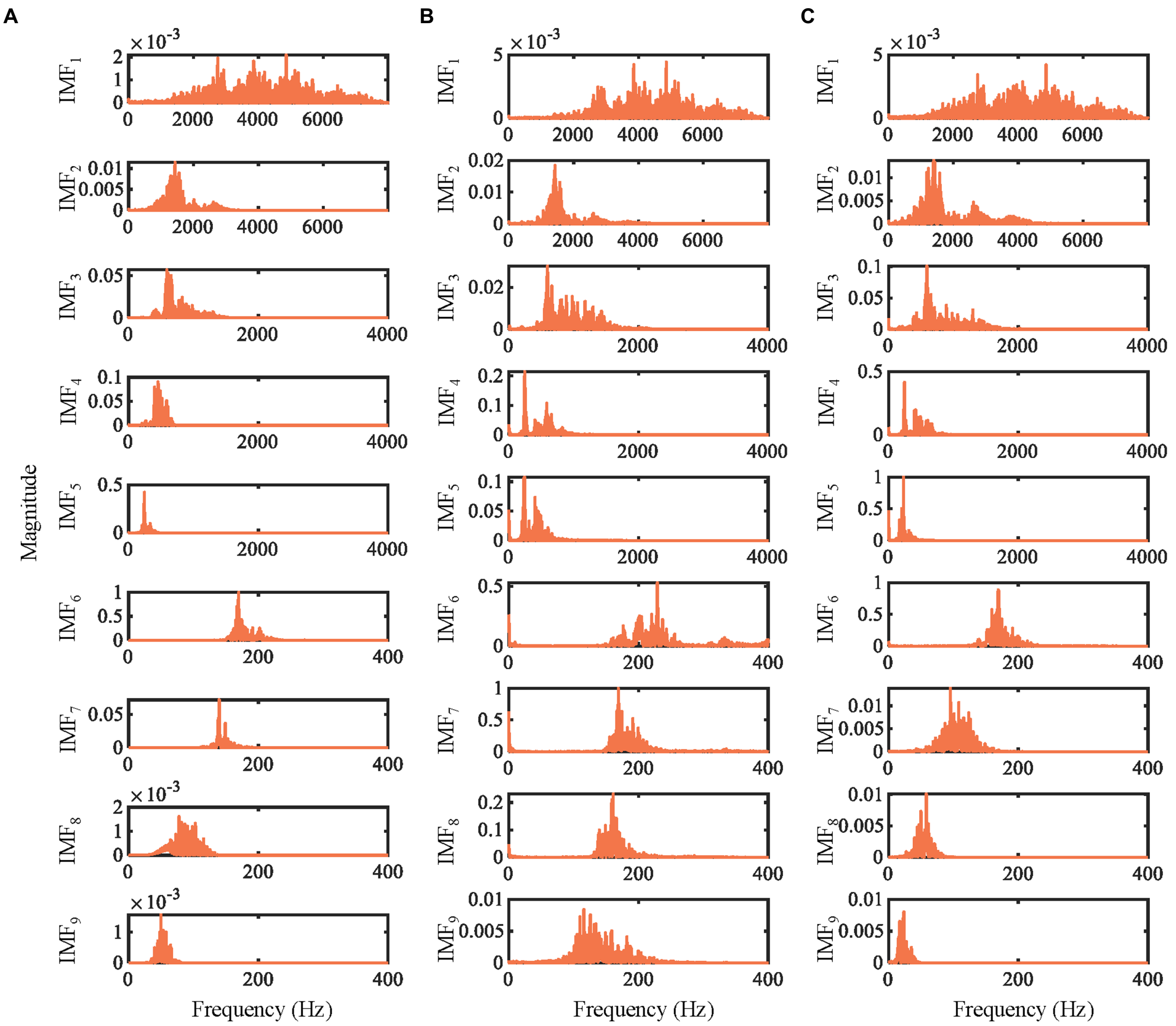


FIGURE 5
The power spectra of the first 9 IMFs obtained by decomposing the emotional speech signal by different methods. **(A)** IMEMD. **(B)** ICEEMDAN. **(C)** UPEMD.

the other methods. The reconstruction error of $x_{s1}(t)$ and $x_{s2}(t)$ obtained by REMD is the largest among all the compared algorithms. For $x_{s1}(t)$, the reconstruction error obtained by REMD is more than 12 times higher than that of EEMD, which has the second-highest reconstruction error. For $x_{s2}(t)$, the reconstruction error obtained by REMD is more than 1.2 times higher than that of UPEMD, which has the second-highest reconstruction error. Therefore, the results of REMD are not drawn in Figures 4B, 3C. The boxplots in Figure 4 show that the distribution of results obtained by IMEMD and UPEMD is more concentrated than that obtained by ICEEMDAN and EEMD. This is because perturbations used by IMPEMD and UPEMD are deterministic, while ICEEMDAN and EMD use random noise. So IMEMD and UPEMD can obtain reproducible decompositions. In conclusion, the IMEMD proposed in this paper reduces the mode mixing effect, provides reproducible decompositions, and has less computational time.

3.3.2. Emotional speech and comparisons

IMEMD is applied to real emotional speech (from the Emo-DB dataset) shown in Figure 5. Figure 5 shows the power spectra of the first 9 IMFs. The spectra of IMFs by each algorithm are normalized by dividing the spectra by their maximum magnitudes. As shown in Section 3.3.1, the reconstruction errors of EEMD and REMD are large. Therefore, IMEMD is only compared with UPEMD and ICEEMDAN. The phase number of $n_p = 64$ is used in IMEMD and UPEMD. The ensemble size of ICEEMDAN is $I = 100$. We set $\xi_0 = 1.5$ for IMEMD and the amplitude of assisted signals to 0.2 for UPEMD and ICEEMDAN.

In Figure 5, the mode mixing of IMEMD is less than that of other methods. For ICEEMDAN and UPEMD, there is mode

TABLE 4 Comparison of different SER methods on the EMO-DB dataset.

Methods	Input feature	UA (%)
Deb and Dandapat (2019)	MFCCs and their first- and second-order difference	85.10
Suganya and Charles (2019)	Raw audio recording	85.62
Kerkeni et al. (2019)	Modulation spectral and modulation frequency features based on EMD and TKEO, and cepstral features.	86.22
Chen et al. (2018)	Log Mel-spectrogram	87.81
Muppidi and Radfar (2021)	RGB Mel-spectrogram	88.78
Kim et al. (2018)	20 features in the eGeMAPS	88.90
Mustaqeem and Kwon (2021)	Raw audio recording	89.37
Zhong et al. (2020)	Log Mel-spectrogram	90.67
Hou et al. (2022)	Prosody features, MFCCs, MFSCs	92.51
Proposed	**Timbre features, spectral features**	**93.54**

mixing between IMF2 and IMF3, and between IMF4 and IMF5. The number of IMFs obtained by IMEMD, UPEMD, and ICEEMDAN is 14, 15, and 23, respectively, which proves that IMEMD can return a more compact representation than other methods. Noise residuals and mode mixing effects have bad effects on the frequency distribution of the IMFs, resulting in the spectrum becoming blurry (Sandoval and De Leon, 2017). So, the performance of IMEMD is better than that of UPEMD and ICEEMDAN.

3.4. Performance analysis of IMEMD-CRNN system

In this section, the proposed IMEMD-CRNN method is applied to the two publicly available Emo-DB and TESS datasets for speech emotion recognition experiments to show the significance and the robustness of the IMEMD-CRNN method. In the upcoming subsections, the experimental results will be described in detail.

3.4.1. Performance on the Emo-DB dataset

The utterances on the Emo-DB dataset are spoken by 10 actors intended to convey one of seven emotions. These seven emotion labels are anger, anxiety/fear, boredom, disgust, happiness, neutral, and sadness. We first preprocess each utterance (The preprocessing method is shown in Section 3.2). Secondly, the signal is decomposed by IMEMD to obtain IMFs. Then, we extract Hilbert spectrum distribution features, Hilbert contour features, SMFCC features, the first derivative of SMFCC, and the second derivative of SMFCC from IMFs (The feature extraction method is shown in Section 2). The dimension of features is 43. We use leave-one-speaker-out (LOSO) 10-fold cross-validation to provide an accurate assessment of the proposed IMEMD-CRNN model (Hou et al., 2022). In the LOSO 10-fold cross-validation method, utterances of 8 speakers are used as training set, one speaker is selected as the validation data, and utterances of the left-out speaker are used as the testing set. We repeat this procedure 10 times. The final classification accuracy is the average of the 10 folds. The initial values of hyperparameters of the CRNN model are referred to Adavanne et al. (2019) and Cao et al. (2019). We further utilize the validation set to debug the hyperparameters to obtain optimal hyperparameters.

Table 4 shows the recognition results of the proposed method with state-of-the-art (SOTA) methods. The unweighted accuracy of our method reaches 93.54%, greater than the SOTA method by 1.03%. To verify that the improvement in accuracy of the proposed method is statistically significant compared to the SOTA method (the method proposed by Hou et al. (2022)), a paired-sample *t*-test is used. The null hypothesis is that the pairwise difference between the UA of the two methods has a mean equal to zero. The significance level α of the hypothesis test is set to 0.05. The *value of p* of the paired-sample *t*-test is 0.01 ($p < 0.05$). Therefore, the improvement in the accuracy of IMEMD-CRNN compared with SOTA method is statistically significant. As shown in Table 4, combining hand-crafted features with deep learning is higher than the methods where the original signals are directly fed into the deep networks. The results demonstrate that effective hand-crafted features combined with deep-learning networks can build a more accurate and robust speech emotion recognition system. The accuracies obtained using our method for each emotion are anger (90.9%), anxiety/fear (96%), boredom (92.4%), disgust (97.6%), happiness (90%), neutral (92.8%), and sadness (95.1%). The results indicate that our proposed IMEMD-CRNN framework has the best performance for disgust and the worst performance for anger and happiness. Some angry samples are identified as happiness and anxiety. A part of happy samples is recognized as angry and anxious. This may be because all three emotions are relatively strong and, therefore, easily misclassified.

3.4.2. Performance on the TESS dataset

To compare with other SER methods, we use randomized 10-fold cross-validation to train and validate our method on the TESS dataset. The final performance is the averaged results of the 10 folds. The preprocessing and feature extraction steps are the same as the Emo-DB database. The initial values of hyperparameters of the CRNN model are referred to Adavanne et al. (2019) and Cao et al. (2019). We further utilize the validation set to debug the hyperparameters to obtain optimal hyperparameters. Table 5 shows the results of comparing the

TABLE 5 Comparison of different SER methods on the TESS dataset.

Methods	Input feature+Classifier	UA (%)
Krishnan et al. (2021)	Entropy features based on EMD + SVM	81.67
Krishnan et al. (2021)	Entropy features based on EMD + LDA	93.30
Chatterjee et al. (2021)	MFCCs +1D CNN	95.79
Proposed	**Timbre and spectral features + CRNN**	**100**

proposed method with the state-of-the-art method on the TESS dataset. From Table 5, it can be seen that the proposed method achieves a UA value of 100% in the TESS database; the UA value is improved by 4.21% compared to the best comparison method. We also use the paired-sample *t*-test to compare the results of IMEMD-CRNN and the method proposed by Chatterjee et al. (2021). The significance level α of the hypothesis test is set to 0.05. The value of p of the paired-sample *t*-test is 6.24×10^{-7} ($p < 0.05$). Therefore, the improvement in the accuracy of IMEMD-CRNN compared with the SOTA method is statistically significant.

4. Conclusion

This paper proposes a novel framework named IMEMD-CRNN to accurately extract emotional information from speech and effectively identify different emotions. The IMEMD-CRNN contains three parts. IMEMD is first used to extract physically meaningful IMFs from speech signals. Then, we extracted time-frequency features from the IMFs that can effectively express speech emotions. Finally, CRNN is employed to further model the speech emotion information in the time-frequency features to realize the recognition of emotion. Comprehensive experiments on the synthetic signals, the Emo-DB dataset, and TESS dataset verify the effectiveness of the proposed scheme. Simultaneously, simulations and emotional speech experiments indicate that our IMEMD mitigates mode mixing and improves decomposition accuracy under low computational cost. More importantly, we compare our proposed scheme with some state-of-the-art SER methods. The results show that our method can accurately extract speech emotion features and significantly improves the performance of SER. The proposed IMEMD-CRNN framework has potential applications in psychology, physiology, signal processing, and pattern recognition involving speech-based affective computing. In future work, to further reduce the mode mixing and improve the ability of IMEMD to decompose signals, the addition of optimization algorithms to the IMEMD will be investigated.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

CS was involved in experiment conduction, data analysis, and manuscript write-up. HL and LM were involved in the conception, supervision, and manuscript review. All authors contributed to the article and approved the submitted version.

Funding

This work was supported in part by the National Natural Science Foundation of China under Grant U20A20383, in part by the National Key R&D Program of China under Grant 2020YFC0833204, Provincial Key R&D Program of Heilongjiang under Grant GY2021ZB0206, Shenzhen Foundational Research Funding under Grant JCYJ20200109150814370, and Funds for National Scientific and Technological Development under Grant 2021SZVUP087 and Grant 2021SZVUP088.

Acknowledgments

The authors are grateful for the reviewers who made constructive comments.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Adavanne, S., Politis, A., Nikunen, J., and Virtanen, T. (2019). Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE J. Select Top. Signal Proc.* 13, 34–48. doi: 10.1109/JSTSP.2018.2885636

Anvarjon, T., Mustaqeem, , and Kwon, S. (2020). Deep-net: a lightweight CNN-based speech emotion recognition system using deep frequency features. *Sensors* 20:5212. doi: 10.3390/s20185212

Arano, K. A., Gloor, P., Orsenigo, C., and Vercellis, C. (2021). When old meets new: emotion recognition from speech signals. *Cogn. Comput.* 13, 771–783. doi: 10.1007/s12559-021-09865-2

Aydın, S., Demirtaş, S., Tunga, M. A., and Ateş, K. (2018). Comparison of hemispheric asymmetry measurements for emotional recordings from controls. *Neural Comput. Appl.* 30, 1341–1351. doi: 10.1007/s00521-017-3006-8

Basu, S., Chakraborty, J., Bag, A., and Aftabuddin, M. (2017). A review on emotion recognition using speech. In: 2017 international conference on inventive communication and computational technologies (ICICCT), Coimbatore, India, 109-114.

Bhattacharjee, M., Prasanna, S. R. M., and Guha, P. (2020). Speech/music classification using features from spectral peaks. *IEEE/ACM Transact. Audio Speech Lang. Proc.* 28, 1549–1559. doi: 10.1109/TASLP.2020.2993152

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B. (2005). A database of German emotional speech. In Proceedings of the 2005—Eurospeech, 9th European Conference on Speech Communication and Technology. Lisbon, Portugal. 1517–1520.

Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., et al. (2008). IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* 42, 335–359. doi: 10.1007/s10579-008-9076-6

Cao, Y., Kong, Q., Iqbal, T., An, F., Wang, W., and Plumbley, M. D. (2019). *Polyphonic sound event detection and localization using a two-stage strategy*. In DCASE.

Chatterjee, R., Mazumdar, S., Sherratt, R. S., Halder, R., Maitra, T., and Giri, D. (2021). Real-time speech emotion analysis for smart home assistants. *IEEE Trans. Consumer Electron* 67, 68–76. doi: 10.1109/TCE.2021.3056421

Chen, M., He, X., Jing, Y., and Zhang, H. (2018). 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. *Signal Process. Lett.* 25, 1440–1444. doi: 10.1109/LSP.2018.2860246

Chen, J., Li, H., Ma, L., and Soong, F. (2022). DEEMD-SPP: a novel framework for emotion recognition based on EEG signals. *Front. Psych.* 13:885120. doi: 10.3389/fpsyt.2022.885120

Colominas, M. A., Schlotthauer, G., and Torres, M. E. (2014). Improved complete ensemble EMD: a suitable tool for biomedical signal processing. *Biomed Signal Process Control* 14, 19–29. doi: 10.1016/j.bspc.2014.06.009

Deb, S., and Dandapat, S. (2019). Emotion classification using segmentation of vowel-like and non-vowel-like regions. *IEEE Trans. Affect. Comput.* 10, 360–373. doi: 10.1109/TAFFC.2017.2730187

Deering, R., and Kaiser, J. F. (2005). The use of a masking signal to improve empirical mode decomposition. In 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, PA, 485–488.

Deschamps-Berger, T., Lamel, L., and Devillers, L. (2021). End-to-end speech emotion recognition: challenges of real-life emergency call centers data recordings. In 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII), Nara, Japan, 1–8.

Dominguez-Jimenez, J. A., Campo-Landines, K. C., Santos, J. C., Delahoz, E. J., and Ortiz, S. H. (2020). A machine learning model for emotion recognition from physiological signals. *Biomed. Signal Process. Contr.* 55:101646. doi: 10.1016/j.bspc.2019.101646

Ekman, P., and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *J. Pers. Soc. Psychol.* 17, 124–129. doi: 10.1037/h0030377

Goshvarpour, A., Abbasi, A., and Goshvarpour, A. (2017). An accurate emotion recognition system using ECG and GSR signals and matching pursuit method. *Biom. J.* 40, 355–368. doi: 10.1016/j.bj.2017.11.001

Guidi, A., Gentili, C., Scilingo, E. P., and Vanello, N. (2019). Analysis of speech features and personality traits. *Biomed. Signal Process Contr.* 51, 1–7. doi: 10.1016/j.bspc.2019.01.027

Hou, M., Zhang, Z., Cao, Q., Zhang, D., and Lu, G. (2022). Multi-view speech emotion recognition via collective relation construction. *IEEE/ACM Transact. Audio Speech Lang. Process.* 30, 218–229. doi: 10.1109/TASLP.2021.3133196

Hsieh, I., and Liu, J. (2019). A novel signal processing approach to auditory phantom perception. *Psychon. Bull. Rev.* 26, 250–260. doi: 10.3758/s13423-018-1513-y

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., et al. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London. *Proc. R. Soc. Lond. A* 454, 903–995. doi: 10.1098/rspa.1998.0193

Karan, B., Sahu, S. S., Orozco-Arroyave, J. R., and Mahto, K. (2020). Hilbert spectrum analysis for automatic detection and evaluation of Parkinson's speech. *Biomed. Signal Process Contr.* 61:102050. doi: 10.1016/j.bspc.2020.102050

Kerkeni, L., Serrestou, Y., Raoof, K., Mbarki, M., Mahjoub, M. A., and Cleder, C. (2019). Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO. *Speech Commun.* 114, 22–35. doi: 10.1016/j.specom.2019.09.002

Kılıç, B., and Aydın, S. (2022). Classification of contrasting discrete emotional states indicated by EEG based graph theoretical network measures. *Neuroinformatics* 20, 863–877. doi: 10.1007/s12021-022-09579-2

Kim, J., and Saurous, R. A. (2018). Emotion recognition from human speech using temporal information and deep learning. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech, 2018-September. 937–940

Kim, J. W., Saurous, R. A., and Int, S. C. A. (2018). Emotion recognition from human speech using temporal information and deep learning. In 19th Annual Conference of the International Speech Communication Association, 937–940.

Krishnan, P. T., Joseph Raj, A. N., and Rajangam, V. (2021). Emotion classification from speech signal based on empirical mode decomposition and non-linear features. *Complex Intell. Systems* 7, 1919–1934. doi: 10.1007/s40747-021-00295-z

Li, H., Chen, J., Ma, L., Bo, H., Xu, C., and Li, H. (2020a). Review of speech dimensional emotion recognition. *J. Softw.* 31, 2465–2491. doi: 10.13328/j.cnki.jos.006078

Li, H., Sun, C., Ma, L., Bo, H., and Xu, Z. (2020b). Timbre feature extraction of musical instrument based on TVF-EMD and its application. *J. Signal Process.* 36, 934–941. doi: 10.16798/j.issn.1003-0530.2020.06.015

Liu, Z., Jin, Y., Zuo, M. J., and Feng, Z. (2017). Time-frequency representation based on robust local mean decomposition for multicomponent AM-FM signal analysis. *Mech. Syst. Signal Process.* 95, 468–487. doi: 10.1016/j.ymssp.2017.03.035

Liu, Z., Peng, D., Zuo, M. J., Xia, J., and Qin, Y. (2022). Improved Hilbert–Huang transform with soft sifting stopping criterion and its application to fault diagnosis of wheelset bearings. *ISA Trans.* 125, 426–444. doi: 10.1016/j.isatra.2021.07.011

Liu, P., Zhang, Y., Xiong, Z., Wang, Y., and Qing, L. (2022). Judging the emotional states of customer service staff in the workplace: a multimodal dataset analysis. *Front. Psychol.* 13:1001885. doi: 10.3389/fpsyg.2022.1001885

Lu, B., Zhang, Y., and Zheng, W. (2021). A survey of affective brain-computer interface. *Chin. J. Intellig. Sci. Technol.* 3, 36–48. doi: 10.11959/j.issn.2096-6652.202104

Muppidi, A., and Radfar, M. (2021). Speech emotion recognition using quaternion convolutional neural networks. Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6309–6313.

Mustaqeem, , and Kwon, S. (2021). MLT-DNet: speech emotion recognition using 1D dilated CNN based on multi-learning trick approach. *Expert Syst. Appl.* 167:114177. doi: 10.1016/j.eswa.2020.114177

Nitsche, M., Koschack, J., Pohlers, H., Hullemann, S., Paulus, W., and Happe, S. (2012). Effects of frontal transcranial direct current stimulation on emotional state and processing in healthy humans. *Front. Psych.* 3:58. doi: 10.3389/fpsyt.2012.00058

Picard, R. W., Vyzas, E., and Healey, J. (2001). Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Trans. Pattern Anal. Mach. Intell.* 23, 1175–1191. doi: 10.1109/34.954607

Pichora-Fuller, M. K., and Dupuis, K. (2020). Toronto emotional speech set (TESS). Vienna Borealis.

Quan, X., Zeng, Z., Jiang, J., Zhang, Y., Lv, B., and Wu, D. (2021). Physiological signals based affective computing: a systematic review. *Acta Automat. Sin.* 47, 1769–1784. doi: 10.16383/j.aas.c200783

Rilling, G., and Flandrin, P. (2008). One or two frequencies? The empirical mode decomposition answers. *IEEE Trans. Acoust. Speech Signal Process.* 56, 85–95. doi: 10.1109/TSP.2007.906771

Sandoval, S., and De Leon, P. L. (2017). Advances in empirical mode decomposition for computing instantaneous amplitudes and instantaneous frequencies. Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) March 2017, 4311–4315.

Sarma, M., Ghahremani, P., Povey, D., Goel, N. K., Sarma, K. K., and Dehak, N. (2018). Emotion identification from raw speech signals using DNNs. In Proceedings of the Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India. 3097–3101

Senroy, N., Suryanarayanan, S., and Ribeiro, P. F. (2007). An improved Hilbert–Huang method for analysis of time-varying waveforms in power quality. *IEEE Transact. Power Syst.* 22, 1843–1850. doi: 10.1109/TPWRS.2007.907542

Sharma, R., Bhukya, R. K., and Prasanna, S. R. M. (2018). Analysis of the Hilbert spectrum for text-dependent speaker verification. *Speech Commun.* 96, 207–224. doi: 10.1016/j.specom.2017.12.001

Suganya, S., and Charles, E. Y. A. (2019). Speech emotion recognition using deep learning on audio recordings. Proceedings of the 19th International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka 1–6.

Vieira, V., Coelho, R., and de Assis, F. M. (2020). Hilbert-Huang-Hurst-based non-linear acoustic feature vector for emotion classification with stochastic models and learning systems. *IET Signal Process.* 14, 522–532. doi: 10.1049/iet-spr.2019.0383

Wang, Y., Hu, K., and Lo, M. (2018). Uniform phase empirical mode decomposition: an optimal hybridization of masking signal and ensemble approaches. *IEEE Access.* 6, 34819–34833. doi: 10.1109/ACCESS.2018.2847634

Wang, C., Ren, Y., Zhang, N., Cui, F., and Luo, S. (2022). Speech emotion recognition based on multi-feature and multi-lingual fusion. *Multimed. Tools Appl.* 81, 4897–4907. doi: 10.1007/s11042-021-10553-4

Wang, X., Wang, M., Qi, W., Su, W., Wang, X., and Zhou, H. (2021). A novel end-to-end speech emotion recognition network with stacked transformer layers. Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada 6289–6293.

Wu, Z., and Huang, N. E. (2009). Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Adv. Adapt. Data Anal.* 01, 1–41. doi: 10.1142/S1793536909000047

Wu, J., Wang, X., Sun, X., and Liu, Y. (2014). Pure harmonics extracting from time-varying power signal based on improved empirical mode decomposition. *Measurement* 49, 216–225. doi: 10.1016/j.measurement.2013.11.041

Xu, G., Wang, X., and Xu, X. (2009). Time-varying frequency-shifting signal-assisted empirical mode decomposition method for AM–FM signals. *Mech. Syst. Signal Process.* 23, 2458–2469. doi: 10.1016/J.YMSSP.2009.06.006

Yang, Z., Zhang, Q., Zhou, F., and Yang, L. (2018). Hilbert spectrum analysis of piecewise stationary signals and its application to texture classification. *Digit. Signal Process.* 82, 1–10. doi: 10.1016/j.dsp.2018.07.020

Yao, X., Bai, W., Ren, Y., Liu, X., and Hui, Z. (2020). Exploration of glottal characteristics and the vocal folds behavior for the speech under emotion. *Neurocomputing* 410, 328–341. doi: 10.1016/j.neucom.2020.06.010

Zhong, Y., Hu, Y., Huang, H., and Silamu, W. (2020). A lightweight model based on separable convolution for speech emotion recognition. Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China 2020. 3331–3335.

TYPE Original Research
PUBLISHED 17 January 2023
DOI 10.3389/fnbot.2022.1057983

OPEN ACCESS

EDITED BY
Wenming Zheng,
Southeast University, China

REVIEWED BY
Jia-Bao Liu,
Anhui Jianzhu University, China
Xin Luo,
Chongqing Institute of Green and Intelligent Technology (CAS), China

*CORRESPONDENCE
Ruoyuan Zhang
✉ hf_zry@163.com

RECEIVED 30 September 2022
ACCEPTED 30 December 2022
PUBLISHED 17 January 2023

CITATION
Ma R and Zhang R (2023) Facial expression recognition method based on PSA—YOLO network. *Front. Neurorobot.* 16:1057983.
doi: 10.3389/fnbot.2022.1057983

Facial expression recognition method based on PSA—YOLO network

Ruoling Ma[1] and Ruoyuan Zhang[2]*

[1]Guangdong Finance and Trade of Vocational College, Guangzhou, China, [2]Anhui Water Conservancy Technical College, Hefei, China

In order to improve the recognition speed and accuracy of face expression recognition, we propose a face expression recognition method based on PSA—YOLO (Pyramids Squeeze Attention—You Only Look Once). Based on CSPDarknet53, the Focus structure and pyramid compression channel attention mechanism are integrated, and the network depth reduction strategy is adopted to build a PSA-CSPDarknet-1 lightweight backbone network with small parameters and high accuracy, which improves the speed of face expression recognition. Secondly, in the neck of the network, a spatial pyramid convolutional pooling module is built, which enhances the spatial information extraction ability of deep feature maps with a very small computational cost, and uses the α—CIoU loss function as the bounding box loss function to improve the recognition accuracy of the network for targets under high IoU threshold and improve the accuracy of face expression recognition. The proposed method is validated on the JAFFE, CK+, and Cohn-Kanade datasets. The experimental results show that the running time of the proposed method and the comparison method is reduced from 1,800 to 200 ms, and the recognition accuracy is increased by 3.11, 2.58, and 3.91%, respectively, so the method proposed in this paper has good applicability.

KEYWORDS

YOLOv4 network, PSA—YOLO network, facial expression recognition, channel attention mechanism, operation efficiency

1. Introduction

Nowadays, with the rapid development of computer technology, automatic facial expression recognition technology has been widely applied in networked learning, medical treatment, transportation, and social security fields (Yao et al., 2018; Zhang and He, 2021). Most methods perform expression recognition when the user's head is in the front or near the front state, and the face is basically unaffected by occlusion (Zhang and Xu, 2020). However, this restriction significantly reduces the robustness of the expression recognition algorithm. In addition, there are also some methods to learn user-related facial features by directly constraining users. This feature is particularly sensitive to the identity information of users, so the identification robustness of unknown users' needs to be improved (Lin et al., 2020).

At present, facial expression recognition is mainly divided into two methods: one is a single frame image, and the other is a video image. The former mainly extracts feature images from the input, while the latter can extract the temporal information of the image sequence and the features of each static image (Chen et al., 2018; Tan et al., 2019; Lin et al., 2020). Some facial expression recognition systems may have good performance in some image datasets but poor performance in others, and there is still room to improve the robustness of facial expression recognition (Li et al., 2020; Liu and Xin, 2020). Based on the above analysis, a facial expression recognition method on PSA—YOLO network is proposed to solve the problems of facial expression recogniti ıracy and data set universality.

Facial expressions correspond to a person's internal emotional state, intention, or social information. Literature Jan et al. (2018) defines six basic terms of "anger," "disgust," "fear," "happiness," "sadness," and "surprise," followed by the expression of "contempt." Facial expression recognition is a traditional problem in computer vision and an essential part of artificial intelligence technology. It has gradually attracted more and more attention, and scholars have proposed a large number of new methods (Islam and Hossain, 2019).

For example, reference Li et al. (2018) proposed a facial expression recognition algorithm combining HOG features and improving KC—FDDL (*K*-means Cluster and Fisher Discrimination Dictionary Learning) Dictionary Learning sparse representation. The HOG features of the normalized expression images were extracted to form the training set, the Fisher discriminant dictionary learning of the improved *K*-means clustering was carried out, and the expression classification was carried out with the sparse representation weighted by the residuals, which overcame the influence of illumination and occlusion in the process of facial expression recognition. However, this method cannot recover sufficient expression information for occluded regions. Literature Tamfous et al. (2020) used sparse coding and dictionary learning methods to study the time-varying shapes in Kendall shape space of 2D and 3D landmarks and studied intrinsic and non-intrinsic solutions to overcome the non-linearity of shape space on facial expression recognition, including action trajectory recognition. However, this method is highly dependent on data sets, and different data sets greatly impact the recognition results (Liu et al., 2020).

In recent years, CNN (Convolutional Neural Networks) has made great contributions to the image classification neighborhood. Many expression recognition methods based on CNN have emerged, which make up for the poor robustness of traditional methods (Wang et al., 2020). For example, a FER (Facial Expression Recognition) method based on a variant feature reduction model and iterative optimization classification strategy was proposed in the literature Du and Hu (2019). WPLBP (Weighted patch-based Local Binary Patterns) is used for feature extraction and expression classification, improving expression recognition accuracy. However, the accuracy of the feature extraction process should be further enhanced. Reference Keyu et al. (2018) proposes a UDADL (Unsupervised Domain Adaptive Dictionary Learning) model, which Bridges the source Domain and target Domain by Learning a shared Dictionary. The analytical dictionary finds approximate solutions as latent variables to simplify the identification process. Literature Liang et al. (2020) proposes a framework for co-learning FER's spatial characteristics and temporal dynamics. The deep network is used to extract spatial features from each frame, the convolution network is used to model the temporal dynamics, and BiLSTM (directional Long Short-Term Memory) network is used to collect clues from the fused functions to complete facial expression recognition. However, the user identity in practice is difficult to define. Literature Chen et al. (2020) proposes a method of facial expression recognition using GAN (Generative Adversarial Network), which focuses on the recognition of facial expressions with a large intra—class gap in the process of facial expression recognition in the real environment so as to better adapt to the tasks with significant intra—class differences.

At present, deep learning-based facial expression target facial expression recognition algorithms are mainly single-stage algorith with YOLO (You Only Look Once) series as the core and two-stage algorithms with RCNN (Region CNN) as the core (Muhammad et al., 2018). Studies in literature Jin et al. (2019) mainly replace or improve the backbone network in YOLO network to improve the facial expression recognition performance of the algorithm. However, the improved network still has shortcomings, such as insufficient attention to the details of expression images and insufficient utilization of semantic information contained in deep features, affecting the performance of facial expression recognition. Therefore, these factors should be fully considered and utilized to improve the performance of the YOLO network in facial expression and facial expression recognition.

To solve the above problems, Ours takes the YOLOv4 target facial expression recognition network as the basis, aiming at the task of facial expression, facial expression recognition, and aiming at improving the accuracy and speed of facial expression, facial expression recognition by the network, builds PSA—YOLO target facial expression recognition network with the characteristics of high facial expression recognition accuracy, fast facial expression recognition speed, and high facial expression recognition rate of small targets.

2. PSA—YOLO recognition algorithm

2.1. PSA—YOLO network structure

Ours proposes a PSA—YOLO network based on the YOLOv4 target facial expression recognition network, in which CBM represents "convolution—batch normalization—Mish activation function module" and BN (Batch Normalization), as shown in Figure 1. First, the Focus structure (Glenn, 2021) and PSA mechanism (Zhang et al., 2021) were added to the CSPDarknet53 backbone network, and residual blocks were stacked in the pattern of "1-1-4-4-2" to simplify the number of network layers. Second, SPC (Squeeze and Concat) module and SPP (Spatial Pyramid Pooling) module (He et al., 2014) are fused into SPCSP (Spatial Pyramid Convolution and Pooling) replaces the original SPP module. Finally, the *k*-means clustering method and α-CIOU loss function are used to perform dimension analysis and bounding box regression on the training image, and the facial expression recognition head part remains unchanged. These parts together constitute the basic structure of the PSA—YOLO network.

2.2. PSA—CSPDarknet feature extraction network

To pay more attention to the channels important to target facial expression recognition information in the initial stage of network forward propagation and fully extract the underlying features of facial expression edge texture to improve the accuracy of facial expression recognition, PSA—CSPDarknet network only retains residual blocks in CSP1-1 layer of CSP1-53 network and adds Focus structure and PSA module in front of residual blocks. The PSA—CSPDarkNet network structure is shown in Figure 2. The Focus structure has been used in the YOLOv5 (Hu et al., 2018) target facial

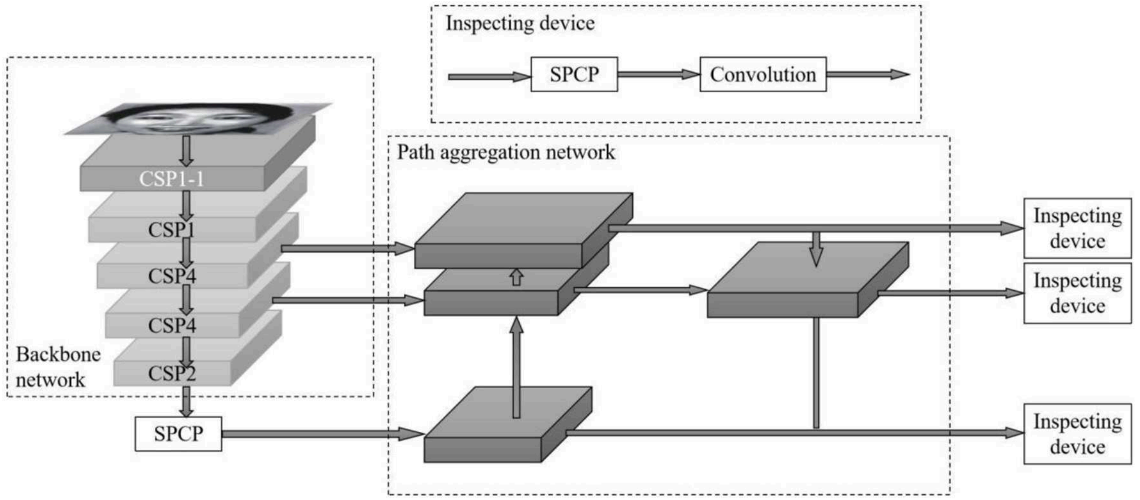


FIGURE 1
Structure of PSA YOLO.

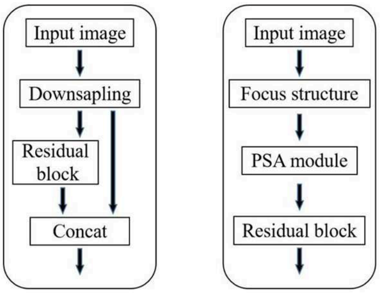


FIGURE 2
Structure of original CSP 1 1 and improved structure.

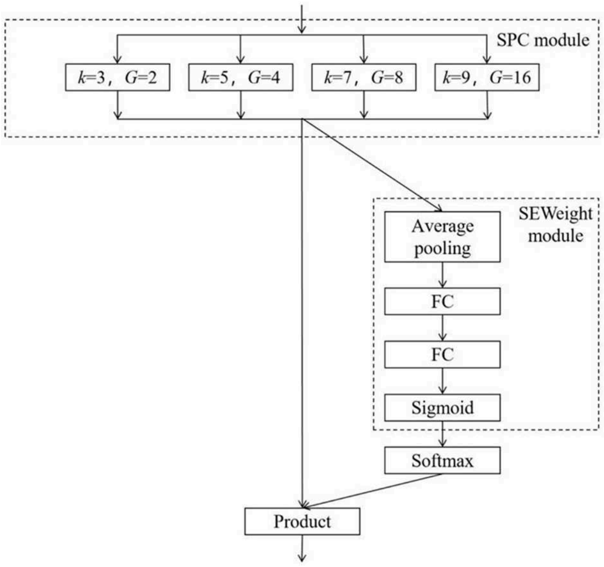


FIGURE 3
Structure of the PSA mechanism.

expression recognition network to replace the backbone network for the first downsampling, showing good facial expression recognition performance in the COCO dataset. The input image is cut into four similar feature maps by tensor slicing operation, and then the four feature maps are fused in the channel dimension to transform the spatial features into channel features without information loss to replace the first down-sampling in the original network.

The PSA module is divided into four parts, as shown in Figure 3 (*K* represents the convolution kernel size, *G* represents the convolution kernel grouping size, and FC represents the fully connected layer). Firstly, the SPC module effectively extracts and integrates the spatial information of different scales of the input feature map. For the spatial dimension of the input feature map, the SPC module uses convolution kernels of four sizes (3, 5, 7, and 9) to perform grouped convolution. The sizes of the grouped kernels of each size are 2, 4, 8, and 16, respectively, to realize grouped convolution and channel compression of the feature map. Then, the SEWeight module (He et al., 2016) is used to learn the weight of the feature map processed by the SPC module, coordinate the local and global attention, and assign different weights according to the importance of the feature channel to the classification task. Softn normalizes the weight of the included channel. The interact between attention weight and the channel is realized by multiplying the normalized weight with the feature map processed by the SPC module so that the channel, which is more important for expression facial expression recognition in the feature map, is assigned with higher weight.

In order to balance the speed and accuracy of the backbone network, based on CSPDarknet53 after the fusion of Focus structure and PSA module, the number of residual blocks is readjusted to simplify the number of network layers and reduce the network parameters and computation burden. Three models were constructed, PSA-CSPDarknet-1, PSA-CSPDarknet-2, and PSA-CSPDarknet-3. Among them, PSA-CSPDarknet-1 halved the number of residual blocks in CSP layer of CSPDarknet53 network and set it as "1-1-4-4-2." Inspired by the network structures of Resnet-18 and Resnet-34 in literature Liu et al. (2018), PSA-CSPDarknet-2 and PSA-CSPDarknet-3 residual arrangements were set as "1-2-2-2-2" and "1-3-4-6-3," respectively.

2.3. SPCP module

To further extract the multi-scale semantic information of spatial dimensions in the deep backbone network, a spatial pyramid convolution pooling module is constructed to replace the original spatial pyramid pooling module. In YOLOv4, the neck mainly consists of two parts: spatial pyramid pooling and PANet (Path Aggregation Network; Rezatofigh et al., 2019). Spatial pyramid pooling is a particular pooling method, which adopts the maximum pooling with a step size of 1 and convolution kernel size of 5 × 5, 9 × 9, and 13 × 13, which is closely integrated with the feature map of the deepest layer of the backbone network to expand the receptive field and integrate multi-scale spatial information. In PSA—YOLO target recognition network, the backbone network extracts local texture and pattern information to construct the semantic information required by the subsequent layer. However, with the increase in complexity, the width of the network will become larger, especially after the SPP module, the number of convolution kernels reaches 2,048, which increases the number of network parameters and computations. The SPC module is inserted before the SPP module, and the number of convolution kernels entering the SPP module is reduced by half by the method of grouping multi-scale convolution followed by recompression. The network computation is further balanced while the extraction of multi-scale spatial information is strengthened. As shown in Figure 4, before the SPC module is added to the SPP module, the number of channels entering the SPP module is compressed to 1,024 to build the SPCP module. On the premise of not affecting the speed of data propagation in the network, the efficiency of using local feature information and global feature information is improved. The bottom-up path enhancement is used in the path aggregation network to shorten the high-low fusion path of the multi-scale feature pyramid. The feature map information of the CSP4 layer, CSP2 layer, and three scales output by the SPCP module is fused in PSA—CSPDarknet. The feature information of shallow networks (CSP4 layer and CSP2 layer) can be used effectively.

2.4. Bounding box loss function

The commonly used bounding box loss functions are evolved based on the IoU loss, such as GIoU (Generalized IoU; Zheng et al., 2020), DIoU (Distance IoU), and CIoU (Complete IoU; He et al., 2021). The α-IoU series loss (Liliana et al., 2019) applies power transformation to summarize the above IOU-based loss. When the noise box with low IoU value appears, the α-IoU loss can adaptively increase the bounding box regression loss value so that the reduction of bounding box loss can be suppressed and the overfitting phenomenon can be avoided when the prediction box with controversy is trained. On the contrary, when the prediction box with high IoU value appears, the α-IoU loss will get lower bounding box loss than the noise box so that the network c predict more objects with high IoU value, and the average accuracy facial expression recognition at high IoU threshold can be improved. Under the action of the above two factors, the facial expression recognition performance of the network with high IoU threshold will be enhanced.

3. Experimental results and analysis

This experiment is based on python1.2 simulation platform, and the hardware environment is: Microsoft Windows 10 operating system, the CPU model is E5-1620 V4, the clock frequency is 3.5 GHz, the graphics card is NVIDIA TITAN V, the video memory size is 12 GB. In this experiment, PSA—YOLO network model was trained for 250 cycles, the minimum batch was 64, and its initial learning rate and learning rate change factor were 0.01 and 0.96, respectively. After each step, the learning rate was reduced. The maximum number of iterations, momentum, and weight decay are 2,000, 0.9, and 0.0002, respectively. After 1,600 iterations, the connections between PSA—YOLO networks have been formed, and the subsequent iterations are trained to enhance correlation and eliminate noise.

3.1. JAFFE dataset experiment

JAFFE is a database of facial expressions with just 213 still images. JAFFE dataset is used to test the effect of a small number of images on system training by different training methods. From the JAFFE dataset, 202 images were selected that were processed using image preprocessing techniques (the JAFFE dataset contains some mislabeled facial expressions that were later removed). This dataset has seven different facial expressions: angry, happy, neutral, surprised, sad, afraid and disgusted. A partial image example of the JAFFE dataset is shown in Figure 5.

In each test, 70% of the images were randomly selected as training images, and the remaining images were used as test images. The recognition effect of the proposed method is experimentally demonstrated on the JAFFE dataset. The confusion matrix of seven expressions is shown in Figure 6.

It can be seen from Figure 6 that the recognition accuracy of the proposed method in seven types of facial expressions is all higher than 60%, among which the recognition accuracy of happy, sad and surprised expressions is all higher than 85%, and the happy expression is the easiest to recognize with an accuracy of 89%. Confusion is often caused by the fact that angry and disgusted expressions are similar to each other in some cases, causing them to be indistinguishable in pixel space. In addition, the JAFFE dataset has a small number of images and is suitable for the PSA—YOLO network, so the overall recognition effect is satisfactory. In addition, in the JAFFE data set, the recognition accuracy of each emotion and the overall recognition accuracy obtained by the proposed method and other comparison methods (methods in literature Du and Hu, 2019; Chen et al., 2020; Liang et al., 2020) are shown in Table 1.

As can be seen from Table 1, both the recognition accuracy of each expression and the overall recognition accuracy, the results obtained by the proposed method are higher than other comparison methods, and the overall recognition accuracy is 83.84%. In literature Du and Hu (2019), WPLBP is used to extract expression features

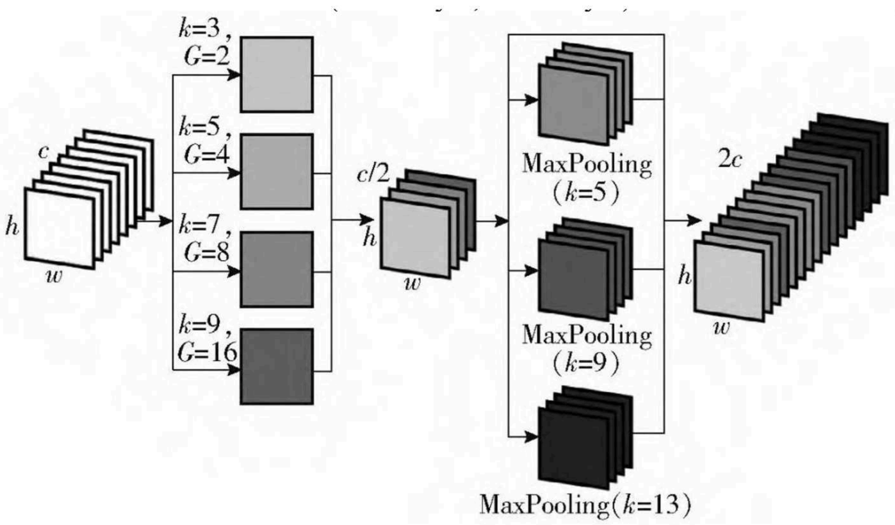


FIGURE 4
Structure of the SPCP module.

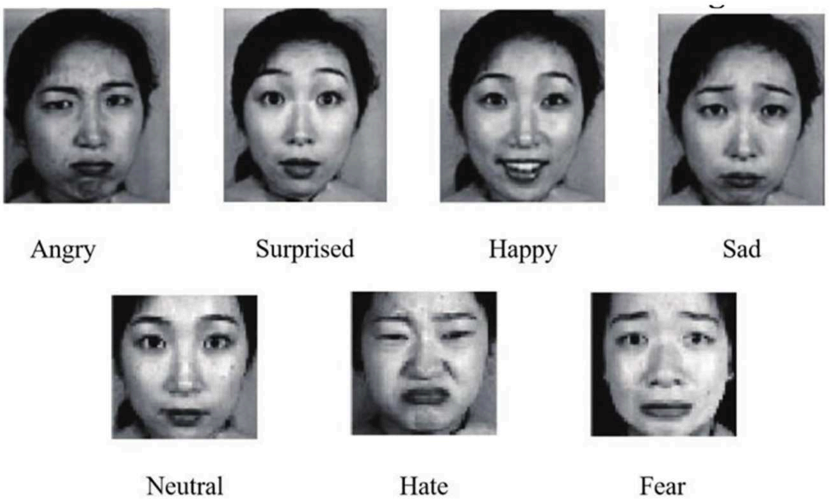


FIGURE 5
Some image examples of the JAFFE dataset.

and iterative optimization classification strategy is used to realize expression recognition. However, this method is greatly affected by the extraction accuracy, so the classification accuracy is low, and the overall recognition accuracy is 74.19%. In literature Liang et al. (2020), deep network is used to extract spatial features from each video frame, and facial expression recognition is completed through the BiLSTM network. Face recognition is completed from two perspectives of time and space, with many constraints, and the recognition accuracy is limited to a certain extent. The overall recognition accuracy is 78.56%. In literature Chen et al. (2020), GAN is used to realize facial expression recognition. This method is used primarily to recognize expressions with large intra-class gaps. Therefore, for expressions with small intra-class gaps, the recognition effect is insignificant, such as neutral and fearful expressions.

3.2. CK+ dataset experiment

The CK+ dataset contains 593 facial expression sequences, each of which can be viewed as several consecutive video frames, with ∼10,000 facial expression images from 123 models. Since these image sequences are continuous, there are many similar images. In the experiment, 693 images were selected and processed by image preprocessing technology after removing the similar images. Images with seven expressions were selected from the dataset: angry, happy, neutral, surprised, sad, afraid, and disgusted. A partial image example of the CK+ dataset is shown in Figure 7.

In each test, 70% of the images were randomly selected as training images, and the remaining images were used as test images. The recognition effect of the proposed method is experimentally

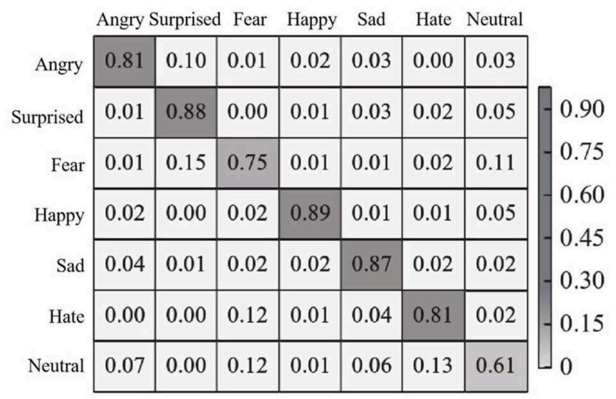


FIGURE 6
Expression recognition confusion matrixbased on JAFFE dataset.

TABLE 1 Expression recognition accuracy obtained by different methods in JAFFE dataset.

Expression	Reference Du and Hu (2019)	Reference Keyu et al. (2018)	Reference Liang et al. (2020)	Ours
Angry	75.02%	77.38%	79.94%	81.02%
Hate	76.39%	78.26%	80.07%	81.95%
Fear	71.88%	73.59%	74.63%	75.78%
Happy	84.39%	86.47%	88.16%	89.01%
Neutral	57.84%	58.91%	60.03%	61.56%
Sad	82.63%	84.19%	86.25%	87.34%
Surprised	83.68%	85.97%	87.35%	88.29%
Average	74.19%	78.56%	80.73%	83.84%

demonstrated on the CK+ dataset. The confusion matrix of seven expressions is shown in Figure 8.

As can be seen from Figure 8, the recognition accuracy of the proposed method is higher than 60% in all seven types of facial expressions, among which the recognition accuracy of happy and sad expressions is 94 and 93%, respectively, and the recognition accuracy of surprised, afraid and disgusted expressions is over 80%. Because angry and disgusted expressions are similar to each other in some cases, they are indistinguishable in pixel space, thus confusing. In addition, the large number of images in the CK+ dataset is conducive to model training, so the overall recognition accuracy is high. In addition, in the CK+ data set, the recognition accuracy of each emotion and the overall recognition accuracy obtained by the proposed method and other comparison methods (methods in literature Du and Hu, 2019; Chen et al., 2020; Liang et al., 2020) are shown in Table 2.

As can be seen from Table 2, the results obtained by the proposed method are higher than other methods in terms of both the recognition accuracy of each emotion and the overall recognition accuracy, with an overall recognition accuracy of 85.09%. Literature Du and Hu (2019) used WPLBP to extract expression features and iteratively optimized the classification strategy to realize express recognition. Literature Liang et al. (2020) used BiLSTM netw combined with deep network to extract spatial and temporal features to complete face recognition. In literature Chen et al. (2020), GAN was used to realize facial expression recognition. Compared with the other three methods, the overall recognition accuracy of the proposed method is improved by 7.32, 4.87, and 3.12%, respectively, which proves the superiority of the facial expression recognition performance.

3.3. Cohn-Kanade dataset experiment

The Cohn-Kanade Facial Expression Database was created in 2000 by the Robotics Institute and the Department of Psychology at CMU. The dataset consists of about 500 sequences of multiple expressions from 100 female adults, including African Americans, Latinos, Asians and others. In the experiment, images need to be normalized to obtain images with sizes of 64 × 64. Some images are shown in Figure 9.

In Cohn-Kanade data set on the experiment, the effect of the method inOurs to identify randomly selected 20 research objects, each object contains six different images of the expression, randomly selected 10 object used in the training, the remaining 10 object is used to test, 30 times to experiment on average, six kinds of expression of the confusion matrix is shown in Figure 10.

As can be seen from Figure 10, the recognition accuracy of each expression of the proposed method is higher than 75%. Since there is no neutral expression in this data set, expressions such as fear and disgust will not be confused with neutral expressions, so the accuracy has been improved to a certain extent. Similarly, happy and sad expressions were easy to recognize, with a recognition accuracy of 92 and 91%, respectively, both higher than 90%. In addition, in the Cohn-Kanade dataset, the recognition accuracy of each emotion and the overall recognition accuracy obtained by the proposed method and other comparison methods (methods in references Du and Hu, 2019; Chen et al., 2020; Liang et al., 2020) are shown in Table 3.

As can be seen from Table 3, consistent with the recognition structure of JAFFE and CK+ datasets, the proposed method has higher recognition accuracy than other comparison methods in each expression and overall recognition, with an overall recognition accuracy of 84.87%. The recognition accuracy of literature Du and Hu (2019) is greatly affected by the feature extraction accuracy of WPLBP method, so the classification accuracy is not high, and the overall recognition accuracy is 78.85%. Literature Liang et al. (2020) combines the spatiotemporal features of facial expressions and uses convolutional network to model the temporal dynamics, which makes it difficult to extract features. In reference Chen et al. (2020), GAN is used to realize facial expression recognition for expressions with large intra-class gap in the process of facial expression recognition. The application scenario is relatively single, and the recognition effect needs to be improved.

3.4. Identify error rates

In order to demonstrate the facial expression recognition performance of the proposed method in the JAFFE data set, CK+ data set and Cohn-Kanade data set, it is compared with the methods in literatures Du and Hu (2019), Chen et al. (2020), and Liang et al.

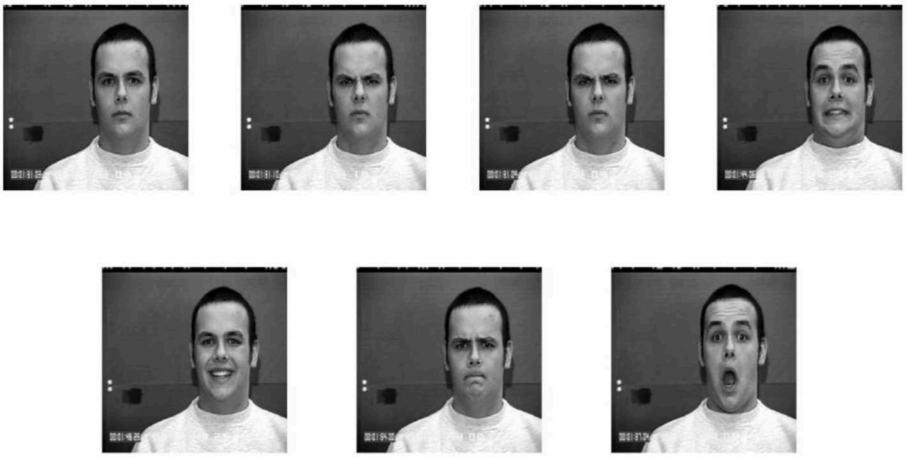
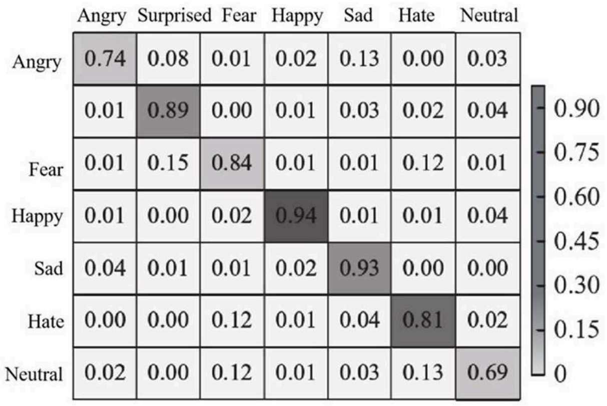


FIGURE 7
Some image examples of the CK+ dataset.



	Angry	Surprised	Fear	Happy	Sad	Hate	Neutral
Angry	0.74	0.08	0.01	0.02	0.13	0.00	0.03
	0.01	0.89	0.00	0.01	0.03	0.02	0.04
Fear	0.01	0.15	0.84	0.01	0.01	0.12	0.01
Happy	0.01	0.00	0.02	0.94	0.01	0.01	0.04
Sad	0.04	0.01	0.01	0.02	0.93	0.00	0.00
Hate	0.00	0.00	0.12	0.01	0.04	0.81	0.02
Neutral	0.02	0.00	0.12	0.01	0.03	0.13	0.69

FIGURE 8
Expression recognition confusion matrix of CK+ dataset.

TABLE 2 Expression recognition accuracy obtained by different methods in the CK+ dataset.

Expression	Reference Du and Hu (2019)	Reference Keyu et al. (2018)	Reference Liang et al. (2020)	Ours
Angry	71.24%	72.38%	73.41%	74.17%
Hate	77.45%	78.37%	79.94%	81.23%
Fear	81.86%	82.59%	83.63%	84.58%
Happy	86.91%	88.73%	90.67%	94.49%
Neutral	65.29%	67.67%	68.35%	69.81%
Sad	85.78%	87.56%	89.49%	93.04%
Surprised	85.68%	86.97%	87.35%	89.96%
Average	79.26%	81.14%	82.51%	85.09%

(2020), and the error rate of 5-fold cross-validation is shown in Table 4.

As can be seen from Table 4, the proposed method has the lowest error rate, which is 8.91%. Because the number of images in JAFFE database is very small, deep PSA—YOLO has not yet shown the best performance, so the performance of PSA—YOLO network is close to the recognition effect of GAN used in literature Chen et al. (2020). However, the proposed method adopts PSA—YOLO network model and spatial pyramid convolution pooling module to enhance the spatial information extraction ability of deep feature maps with minimal computational cost, so the expression recognition effect is better. The CK+ dataset, two images were selected for each expression for each subject, one of which was the frame at the beginning of the expression of the emotion, while the other was the frame in the image sequence when the emotion reached its expression peak. The combined classification of the two images can reduce the error rate, so the error rate of the proposed method is reduced compared with the JAFFE dataset. As can be seen from Table 4, proposed method achieves the lowest error rate of 6.92%. Due to limited number of images and limited network learning, the error rate of this dataset is higher than that of CK+ dataset, but lower than that of JAFFE dataset due to the lack of neutral expression, which avoids expression confusion.

3.5. Other factors affecting the average recognition rate

In order to further evaluate the performance of the proposed method, it is compared with the methods in literatures Du and Hu (2019), Chen et al. (2020), and Liang et al. (2020) in terms of the running time of the training network and the accuracy of facial expression recognition. The recognition accuracy and running time of different methods on the JAFFE, CK+, and Cohn-Kanade datasets are shown in Figure 11.

As can be seen from Figure 11, on JAFFE, CK+, and Cohn-Kanade datasets, compared with other methods, ours integrates Focus structure and PSA mechanism on the basis of CSPDarknet53, and adopts network depth reduction strategy. A lightweight PSA

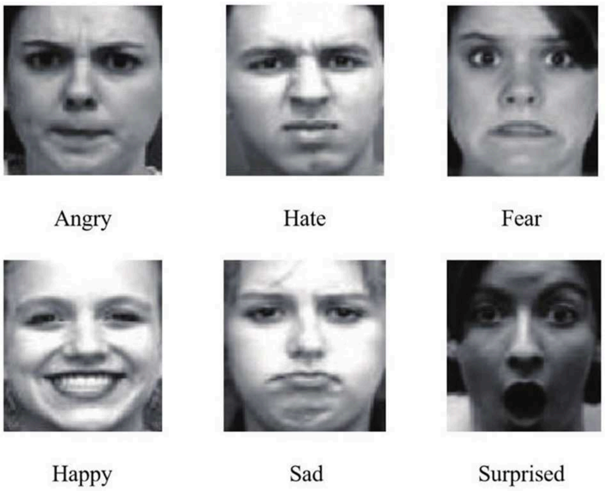
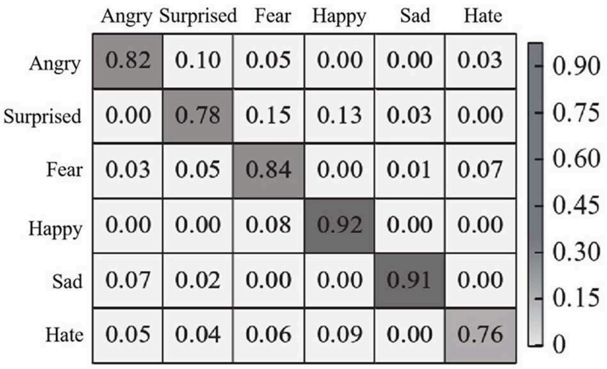


FIGURE 9
Some image examples of Cohn-Kanade dataset.



	Angry	Surprised	Fear	Happy	Sad	Hate
Angry	0.82	0.10	0.05	0.00	0.00	0.03
Surprised	0.00	0.78	0.15	0.13	0.03	0.00
Fear	0.03	0.05	0.84	0.00	0.01	0.07
Happy	0.00	0.00	0.08	0.92	0.00	0.00
Sad	0.07	0.02	0.00	0.00	0.91	0.00
Hate	0.05	0.04	0.06	0.09	0.00	0.76

FIGURE 10
Expression recognition confusion matrix of Cohn-Kanade dataset.

TABLE 3 Expression recognition accuracy obtained by different methods in Cohn-Kanade dataset.

Expression	Reference Du and Hu (2019)	Reference Keyu et al. (2018)	Reference Liang et al. (2020)	Ours
Angry	78.35%	79.81%	80.41%	82.02%
Hate	72.91%	74.62%	75.49%	76.13%
Fear	81.66%	82.57%	83.36%	84.47%
Happy	87.16%	89.98%	91.74%	92.61%
Sad	86.78%	87.65%	89.49%	91.08%
Surprised	75.02%	76.72%	77.35%	78.25%
Average	78.85%	80.04%	81.68%	84.87%

CSPDarknet 1 backbone network with a small number of parameters and high accuracy was constructed. Secondly, in the neck of the network, a spatial pyramid convolution pooling module is b to enhance the spatial information extraction ability of the d

TABLE 4 Error rates in different datasets and different methods.

Algorithm data set	Reference Du and Hu (2019)	Reference Keyu et al. (2018)	Reference Liang et al. (2020)	Ours
JAFFE (error rate)	26.72%	18.86%	11.08%	8.91%
CK+ (error rate)	21.19%	16.83%	10.95%	5.37%
Cohn-Kanade (error rate)	23.08%	18.15%	10.37%	6.92%

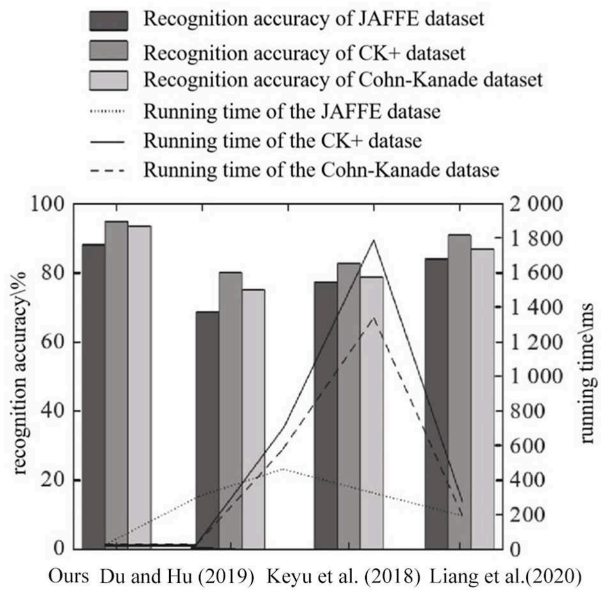


FIGURE 11
Comparison of running time and recognition accuracy.

feature map with minimal computational cost, and the α-CIO U loss function is used as the bounding box loss function to obtain high recognition accuracy. In literature Liang et al. (2020), BiLSTM network combined with spatial and temporal features extracted from deep network is used to complete face recognition and recognize the amount of system data. Therefore, the running time is the longest, which is close to 1,800 ms on CK+ dataset. The WPLBP method in reference Du and Hu (2019) and the GAN model system in reference Chen et al. (2020) are simple in composition, so the running time is reduced compared with that in reference Liang et al. (2020), but the recognition accuracy is lower than that of the proposed method. In addition, the ratio of training images to the images used in the test evaluation enables to evaluate the impact of the ratio of training images of different methods on the selected dataset. In the experiment, 70% of the images in the data set are used as the training set, and the rest are used as the test set. Taking JAFFE database as an example, different proportions of training images using different methods and the resulting recognition accuracies are shown in Figure 12.

As can be seen from Figure 12, when the ratio of training images increases, the recognition accuracy of all methods will improve, and the proposed method shows the best performance regardless of the

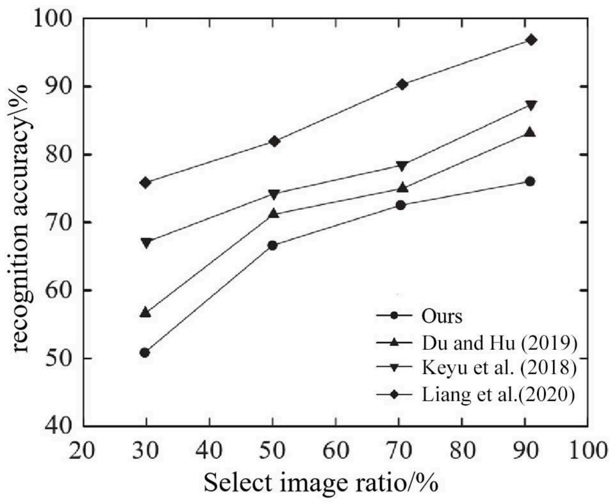


FIGURE 12
Influence of training image proportion on recognition accuracy of different methods.

TABLE 5 Ablation experimental results in CK+ dataset.

Ablation experiment entries	Average recognition rate
None feature extraction module	51.72%
None PSA module	59.07%
None CSPDdrknet53 module	75.12%
Complete network	83.84%

ratio of training images tested. In fact, when 90% of the images were randomly selected from the JAFFE database as training images and the remaining images were used as the test dataset, the recognition accuracy of the method reached 96.0%.

3.6. Ablation experiments

In order to clarify the influence of each network component on classification performance and operational efficiency, an ablation experiment was conducted using the CK+ dataset as an example to test the average accuracy of six expressions of happiness, sadness, anger, surprise, fear and disgust. The proposed method includes four main parts: feature extraction module, PAS module, CSPDdrknet53 module and classification module. Since the classification module is necessary for classification in the network in this paper, the classification module is retained, and the feature extraction module, PAS module, and CSPDdrknet53 module are deleted respectively, and then different experiments are performed, and the results are shown in Table 5. It can be seen that when one of the modules is deleted, the average recognition rate decreases to a certain extent compared to the complete network. Especially in the absence of the feature extraction module, the recognition rate decreased the most, only 51.72%. Generally, the initial features obtained are too coarse, and direct entry into subsequent processing will seriously affect the subsequent results. Therefore, the feature extraction module is required in the network. The PSA module is the core module of the proposed method, and the lack of this module also leads to a serious decrease in the recognition rate, which proves the importance of the PSA module. It can also be seen from Table 5 that without CSPDdrknet53 module, the average recognition rate is 75.12%. Therefore, each module has a certain boost in the final output.

4. Conclusion

In order to improve the recognition speed and accuracy of face expression recognition, ours propose a face expression recognition method based on PSA-YOLO. Based on the YOLOv4 network, comparative experiments were carried out on the backbone network, neck, and bounding box loss function. Based on CSPDarknet53, the Focus structure and pyramid compression attention mechanism are added, and the lightweight processing is carried out to build the PSA CSPDarknet backbone network. Secondly, the spatial pyramid convolution pooling module is used in the neck, and the α-CIoU loss is optimized as the bounding box loss function of the expression recognition network. Eventually, the PSA—YOLO network was built. Ablation validation of the proposed method was performed on the JAFFE, CK+, and Cohn-Kanade datasets. The experimental results show that the running time of the proposed method and the comparison method is reduced from 1,800 to 200 ms, and the recognition accuracy is increased by 3.11, 2.58, and 3.91%, respectively, which has obvious recognition advantages.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://tianchi.aliyun.com/dataset?spm=5176.27124976.J_3941670930.19.71de132aItNJg9.

Author contributions

Both authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Chen, J., Xu, R., and Liu, L. (2018). Deep peak-neutral difference feature for facial expression recognition. *Multimed. Tools Appl.* 77, 29871–29887. doi: 10.1007/s11042-018-5909-5

Chen, L., Wu, P., and Liu, Y. T. (2020). Depth learning recognition method for intra-class gap expression. *J. Image Graph.* 25, 679–687.

Du, L., and Hu, H. (2019). Weighted patch-based manifold regularization dictionary pair learning model for facial expression recognition using iterative optimization classification strategy. *Comput. Vis. Image Understand.* 18, 13–24. doi: 10.1016/j.cviu.2019.06.003

Glenn, J. (2021). *Yolov5*. Available online at: https:github.com/ultralytics/yolov5 (accessed November 22, 2022).

He, J., Erfani, S., Ma, X., Bailey, J., Chi, Y., and Hua, X. S. (2021). "Alpha Io U: A family of power intersection over union losses for bounding box regression," in *Conference and Workshop on Neural Information Processing Systems*, 13675.

He, K., Zhang, X., Ren, S., and Sun, J. (2014). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pat. Anal. Machine Intell.* 37, 1904–1916. doi: 10.1109/TPAMI.2015.2389824

He, K., Zhang, X. Y., Ren, S. Q., and Sun, J. (2016). "Deep residual learning for image recognition," in *Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 770–778. doi: 10.1109/CVPR.2016.90

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141. doi: 10.1109/CVPR.2018.00745

Islam, B., and Hossain, A. (2019). Fusion of features and extreme learning machine for facial expression recognition. *J. Comput. ENCES* 15, 1833–1841. doi: 10.3844/jcssp.2019.1833.1841

Jan, A., Ding, H., Meng, H., Chen, L., and Li, H. (2018). "Accurate facial parts localization and deep learning for 3D facial expression recognition," in *Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition*. (Xi'an), 466–472. doi: 10.1109/FG.2018.00075

Jin, X., Wu, L., Li, X., Zhang, X., Chi, J., Peng, S., et al. (2019). ILGNet: Inception modules with connected local and global features for efficient image aesthetic quality classification using domain adaptation. *IET Comput. Vis.* 13, 206–212. doi: 10.1049/iet-cvi.2018.5249

Keyu, Y., Wengming, Z., Zhen, C., Yuan, Z., Tong, Z., and Chuangao, T. (2018). Unsupervised facial expression recognition using domain adaptation based dictionary learning approach. *Neuro Comput.* 319, 84–91. doi: 10.1016/j.neucom.2018.07.003

Li, M., Peng, X. J., and Wang, Y. (2018). Facial expression recognition based on improved dictionary learning and sparse representation. *J. Syst. Simulat.* 30, 28–35. doi: 10.16182/j.issn1004731x.joss.201801004

Li, T. T., Hu, Y. L., and Wei, F. L. (2020). Improved facial expression recognition algorithm based on GAN and application. *J. Jilin Univ.* 58, 163–168. doi: 10.13413/j.cnki.jdxblxb.2019374

Liang, D., Liang, H., Yu, H., and Zhang, Y. (2020). Deep convolutional BiLSTM fusion network for facial expression recognition. *Vis. Comput.* 36, 499–508. doi: 10.1007/s00371-019-01636-3

Liliana, D. Y., Basaruddin, T., Widyanto, M. R., and Oriza, I. I. D. (2019). Fuzzy Emotion: A natural approach to automatic facial expression recognition from psychological perspective using fuzzy system. *Cogn. Process.* 20, 391–403. doi: 10.1007/s10339-019-00923-0

Lin, K. Z., Bai, J. X., Li, H. T., and Li, A. (2020). Facial expression recognition with small samples fused with different models under deep learning. *J. Front. Comput. Sci. Technol.* 14, 127–137. doi: 10.3778/j.issn.1673-9418.1904028

Liu, F., Li, M., Hu, J., Xiao, Y., and Qi, Z. (2020). Expression recognition based on low pixel face images. *Laser Optoelectron. Progr.* 57, 97–104. doi: 10.3788/LOP57.101008

Liu, Q. M., and Xin, Y. Y. (2020). Face expression recognition based on end-to-end low-quality face images. *J. Chin. Comput. Syst.* 41, 668–672.

Liu, S., Qi, L., Qin, H., Shi, J., and Jia, J. (2018). "Path aggregation network for instance segmentation," in *Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 8759–8768. doi: 10.1109/CVPR.2018.00913

Muhammad, N. A., Nasir, A. A., Ibrahim, Z., and Sabri, N. (2018). Evaluation of CNN, alexnet and GoogleNet for fruit recognition. *Indonesian J. Electr. Eng. Comput. Sci.* 12, 468–475. doi: 10.11591/ijeecs.v12.i2.pp468-475

Rezatofigh, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S. (2019). "Generalized intersection over union: A metric and a loss for bounding box regression," in *Conference on Computer Vision and Pattern Recognition* (Long Beach, CA), 658–666. doi: 10.1109/CVPR.2019.00075

Tamfous, A. B., Drira, H., and Amor, B. B. (2020). Sparse coding of shape trajectories for facial expression and action recognition. *IEEE Trans. Pat. Anal. Machine Intell.* 42, 2594–2607. doi: 10.1109/TPAMI.2019.2932979

Tan, L. Z., Ding, Y., and Xia, L. M. (2019). Facial expression recognition combined with orthogonal neighborhood preserving projection and CNN. *J. Chin. Comput. Syst.* 40, 2221–2226.

Wang, X. H., Liang, Y. C., and Ma, X. C. (2020). Facial expression classification algorithm research based on ideology of inception. *Opt. Technique* 46, 94–100.

Yao, Y., Huang, D., Yang, X., Wang, Y., and Chen, L. (2018). Texture and geometry are scattering representation-based facial expression recognition in 2D+3D videos. *ACM Trans. Multimed. Comput. Commun. Appl.* 14, 1–23. doi: 10.1145/3131345

Zhang, A. M., and Xu, Y. (2020). Attention hierarchical bilinear pooling residual network for expression recognition. *Comput. Eng. Appl.* 56, 161–166.

Zhang, H., Zu, K., Lu, J., Zou, Y., and Meng, D. (2021). EPSANet: An efficient pyramid squeeze attention block on convolutional neural network. *Ar Xiv E-prints.* doi: 10.48550/arXiv.2105.14447

Zhang, R., and He, N. (2021). A survey of micro-expression recognition methods. *Comput. Eng. Appl.* 57, 38–47.

Zheng, Z., Wang, P., Ren, D., Liu, W., Ye, R., Hu, Q., et al. (2020). "Distance Io U loss: Faster and better learning for bounding box regression," in *Conference on Association for the Advancement of Artificial Intelligence* (New York, NY), 12993–13000. doi: 10.1609/aaai.v34i07.6999

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 09 February 2023
DOI 10.3389/fnins.2023.1135986

OPEN ACCESS

EDITED BY
Xiaopeng Hong,
Harbin Institute of Technology, China

REVIEWED BY
Zhijun Yang,
Middlesex University London, United Kingdom
Yimin Wang,
Anhui Agricultural University, China

*CORRESPONDENCE
Liangliang Lou
✉ lou0908@126.com

SPECIALTY SECTION
This article was submitted to
Perception Science,
a section of the journal
Frontiers in Neuroscience

RECEIVED 02 January 2023
ACCEPTED 25 January 2023
PUBLISHED 09 February 2023

CITATION
Song M, Zhu R, Chen X, Zheng C and Lou L (2023) A lightweight attention deep learning method for human-vehicle recognition based on wireless sensing technology. *Front. Neurosci.* 17:1135986. doi: 10.3389/fnins.2023.1135986

A lightweight attention deep learning method for human-vehicle recognition based on wireless sensing technology

Mingxin Song[1,2], Rensheng Zhu[3], Xinquan Chen[1,2], Chunlei Zheng[4] and Liangliang Lou[1]*

[1]Institute of Intelligent Information Processing, Taizhou University, Taizhou, Zhejiang, China, [2]School of Science, Zhejiang University of Science and Technology, Hangzhou, Zhejiang, China, [3]China United Network Communications Co., Ltd., Taizhou Branch, Taizhou, Zhejiang, China, [4]Key Laboratory of Wireless Sensor Networks and Communications, Shanghai Institute of Microsystem and Information Technology, Shanghai, China

Wireless sensing-based human-vehicle recognition (WiHVR) methods have become a hot spot for research due to its non-invasiveness and cost-effective advantages. However, existing WiHVR methods shows limited performance and slow execution time on human-vehicle classification task. To address this issue, a lightweight wireless sensing attention-based deep learning model (LW-WADL) is proposed, which consists of a CBAM module and several depthwise separable convolution blocks in series. LW-WADL takes raw channel state information (CSI) as input, and extracts the advanced features of CSI by jointly using depthwise separable convolution and convolutional block attention mechanism (CBAM). Experimental results show that the proposed model achieves 96.26% accuracy on the constructed CSI-based dataset, and the model size is only 5.89% of the state of the art (SOTA) model. The results demonstrate that the proposed model achieves better performance on WiHVR tasks while reducing the model size compared to SOTA model.

KEYWORDS

human-vehicle recognition, channel status information, attention mechanism, depthwise separable convolution, wireless sensing

1. Introduction

INTELLIGENT Traffic Systems (ITS) is an important part of smart city (Santos et al., 2018; Jin and Ma, 2019; Zhao et al., 2019; Choy et al., 2020; de Oliveira et al., 2020), providing reliable, safe, and convenient services for road users (e.g., cars, motorcycle, pedestrians, etc.). As the number of road users continues to increase, a large number of existing ITSs are approaching their limits. In order to improve the performance of ITSs and relieve traffic pressure, the measurement of traffic parameters including road user behavior has become a research hotspot (Jiang et al., 2021; Park et al., 2021; Zhao and Huang, 2021). Generally, the behavior of road users includes human-vehicle recognition (HVR), traffic flow statistics, vehicle speed, and direction measurement, etc. As the foundation of road user behavior detection, the accuracy of human-vehicle recognition determines the performance of traffic parameter measurement (Won et al., 2017; Sliwa et al., 2020).

With the rapid development of artificial intelligence and deep learning techniques, image-based HVR methods (Huang et al., 2020; Du et al., 2021) have been widely used in ITSs. Such

HVR methods not only achieve excellent recognition performance, but also provide rich traffic image information for city managers. However, image-based HVR methods are susceptible to light so that their performance can be degraded rapidly in the case of low-light conditions such as night, cloudy, haze, etc. To alleviate the limitations of image-based HVR methods in low-light scenes, the microwave radar-based HVR method is proposed (Park et al., 2021; Singh et al., 2021; Tavanti et al., 2021). However, high-performance frequency-modulated continuous-wave (FMCW) radars come at a higher cost. In addition, the microwave radar has the problem of installation viewing angle, which leads to the high construction cost of the microwave radar-based HVR method.

Generally, the purpose of wireless sensing-based HVR (WiHVR) methods is to extract the energy information of surrounding wireless signals for target recognition. Since the propagation of wireless signals has no directionality, the WiHVR method does not have the problem with the above-mentioned viewing angle (Ma et al., 2019; Pan et al., 2019; Zhang et al., 2021). In recent years, the WiHVR methods based on extracted receiving signal strength (RSS) or channel state information (CSI) signatures from the wireless transceivers on 2.4 GHz band, such as Bluetooth (Sliwa et al., 2020; Wilby et al., 2020), ZigBee (Wang et al., 2017; Jiang et al., 2021), and WiFi (Chen et al., 2018; Wang F. et al., 2018), etc., have been widely employed to detect road users in ITSs.

1.1. CSI-based HVR methods

Wang W. et al. (2016) converted CSI signals to spectrograms, thereby describing human motion. Won et al. (2017) proposed a WiFi-based traffic monitoring system, in which the features of root mean square, median absolute deviation, mean, first quartile, and third quartile of the CSI signals were extracted, followed by a support vector machine for vehicle classification. Liu et al. (2017) showed a human motion detection method based on CSI phase difference. They discussed the situation of line of sight (LOS) and non-line of sight (NLOS). Arshad et al. (2018) proposed a WiFi-based device-free dangerous driving recognition system. This system extracted multi-domain features for both magnitude and phase of CSI signals. Wang J. et al. (2018) presented a new general device-free identification framework *via* empirical mode decomposition. They decomposed CSI signals into intrinsic mode functions (IMF) and extracted the time domain and frequency domain features from IMF components.

1.2. RSS-based HVR methods

Jiang et al. (2021) calculated the amplitude and mean information of RSS signals. They designed a HVR algorithm for WiHVR based on the calculated RSS features. Sliwa et al. (2020) provided a vehicle detection and classification method on the basis of the extracted RSS from transceivers on 2.4 GHz band. They used mean, minimum, standard deviation, and other characteristics of RSS signals to address the challenges of accuracy, robustness, and privacy. Abdelnasser et al. (2018) exploited a gesture recognition system in which the edge, frequency, and magnitude features of RSS signals were extracted for gesture recognition. Bhat et al. (2020) extracted the RSS power levels for human locomotion walking pattern recognition.

However, the above-mentioned WiHVR methods based on extracted RSS or CSI signatures from 2.4 GHz wireless transceivers like Bluetooth, ZigBee, and WiFi have the following drawbacks:

1) RSS is a coarse-grained signal, which leads to limited accuracy of HVR tasks based on RSS signals.
2) The effects of CSI or RSS on the performance of WiHVR in different application scenarios are not explored.

Recently, deep learning techniques (LeCun et al., 2015) consisting of a multi-layer network architecture have attracted much interest. One of the representative deep learning techniques is convolutional neural network (CNN) (Krizhevsky et al., 2012). Up to now, due to the powerful feature learning ability, CNNs have exhibited promising performance on various tasks such computer vision (Szegedy et al., 2016), speech signal processing (Zhang et al., 2017), natural language processing (Otter et al., 2020), and so on. However, few works have attempted to exploit the application of CNNs on WiHVR tasks.

To address the above-mentioned issues, this paper presents a novel WiHVR method based on the designed lightweight wireless sensing attention-based deep learning model (LW-WADL). Inspired by the recent-emerged convolutional block attention mechanism (Woo et al., 2018) (CBAM) and depthwise separable convolutions (Chollet, 2017), we propose a new deep model, which consists of a CBAM module and three depthwise separable convolution blocks in series to learn high-level features from preprocessing CSI signals for WiHVR. Compared with ordinary convolutions, depthwise separable convolutions have relatively low parameters and operations. Besides, we propose a novel CSI data enhancement method and a new subcarrier selection method. In particular, a new CSI-based dataset relates to road user behavior is constructed. In order to explore the effects of CSI on the performance of WiHVR in different application scenarios, the CSI dataset is divided into three taxonomies according to the number of categories, namely, two-category dataset, three-category dataset, and four-category dataset. Experimental results show that the accuracy of CSI-based methods decreases as the number of classification categories increases. For four-classification experiments, the proposed model achieves 96.26% accuracy and the model size is only 5.89% of the state of the art model.

To summarize, the main contributions of this paper are as follows:

1) This paper proposes a CSI data enhancement method, which preprocess the change trend of CSI data to one direction, thereby enhancing CSI data.
2) This paper provides a subcarrier selection method, which selects several subcarriers with large signal-to-noise ratios (SNR) as benchmarks and integrates them into a new CSI data.
3) This paper has proposed a lightweight wireless sensing attention-based deep learning model, and attempts to explore the effects of CSI on the performance of WiHVR in different application scenarios.

The remainder of this paper is organized as follows. Section "2. Preliminaries" introduces the CSI extraction and the theoretical analysis of WiHVR. Section "3. Proposed method" elaborates the proposed LW-WADL for WiHVR. Section "4. Experiment study" shows experimental results and analysis. Section "5. Conclusion and future work" gives the conclusions and future work.

2. Preliminaries

This paper aims to establish a lightweight and efficient WiHVR method to explore the effects of CSI on the performance of WiHVR in different application scenarios. The system architecture of the proposed WiHVR method is shown in Figure 1.

From Figure 1, it can be found that wireless transceiver prototype (WTP) is built and placed on both sides of the road. The WTP is mastered by the ESP32 chip for generating and receiving wireless signals. Once a road user appears in the WTP sensing area, the CSI signal collected by the WTP will be attenuated due to the road user. Therefore, WTP can extract CSI signals related to the road user information.

2.1. CSI extraction

This paper uses the designed WTP to extract CSI data related to road users. CSI represents the fine-grained channel features of wireless communication links between transmitters and receivers based on orthogonal frequency division multiplexing (OFDM) technology. Besides, CSI describes the changes of phase and amplitude caused by multipath effect and transmission loss in wireless signal transmission. The CSI channel gain matrix is expressed as:

$$Mcsi = \begin{pmatrix} h_{11} & \dots & h_{1n} \\ \vdots & \ddots & \vdots \\ h_{m1} & \cdots & h_{mn} \end{pmatrix} \tag{1}$$

where h_{mn} represents the different subcarriers. m and $n$represent the transmitting and receiving antennas, respectively. Each sub-element h_{mn} represents:

$$h_{mn} = ||h_{mn}||e^{j\eta_{mn}} \tag{2}$$

where $||h_{mn}||$ is the amplitude of the sub-carrier h_{mn}, and $e^{j\eta_{mn}}$ represent the phase of h_{mn} . From Eqs 1, 2, it can be known that CSI is not a supersession of all subcarrier signals, it describes a multipath signal with more characteristics. In this case, the CSI extracted by WTP contains multiple subcarrier information. These subcarriers have different sensitivities to road users, so it is necessary to filter out the subcarriers with lower sensitivity. The specific method will be elaborated in Section "3. Proposed method."

The specific process of CSI signal extraction is shown in Figure 2. The acquisition of CSI signal needs to be operated by inverse OFDM. In order to eliminate inter-symbol interference and inter-channel interference, OFDM will use cyclic prefix (C/P), but this part is not real data, so this part needs to be removed in inverse OFDM. After that, it is necessary to convert the series signal to the parallel signal (S/P), and perform discrete Fourier transform (DFT) or fast Fourier transform (FFT) to obtain the required CSI signal.

2.2. Theoretical analysis of WiHVR

The idea of WiHVR is based on the fact that road users of existence and movement affect the wireless propagation paths. To understand the relation of road users movement with received CSI, the wireless propagation model should be first studied. In a typical wireless environment, there is one main path line-of-sight (LOS) and several reflected paths by the surroundings. As shown in Figure 1, if a road user is present in the WTP sensing area, it will cause multipath propagation of the wireless signal. In this case, according to the free space model, the received power by a receiver antenna which is separated from a radiating transmitter antenna by a distance d, is given by the Friis free space equation,

$$P_r = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 d^2} \tag{3}$$

where P_r and P_t are the receiving and transmitting power, respectively. G_r and G_t are the receiving and transmitting antenna gains, respectively. λ is the wavelength in meters. d is the distance between the transmitter and receiver in meters, that is, the propagation path length. When a road user exists in the wireless environment, several scattered paths are produced by road user. Those scattered power should also be added in the final received power.

$$P_r = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 (d^2 + \delta^2)} \tag{4}$$

where δ is a brief representation of path length caused by road user. If a road user is static in the environment, P_r is almost stable. However,

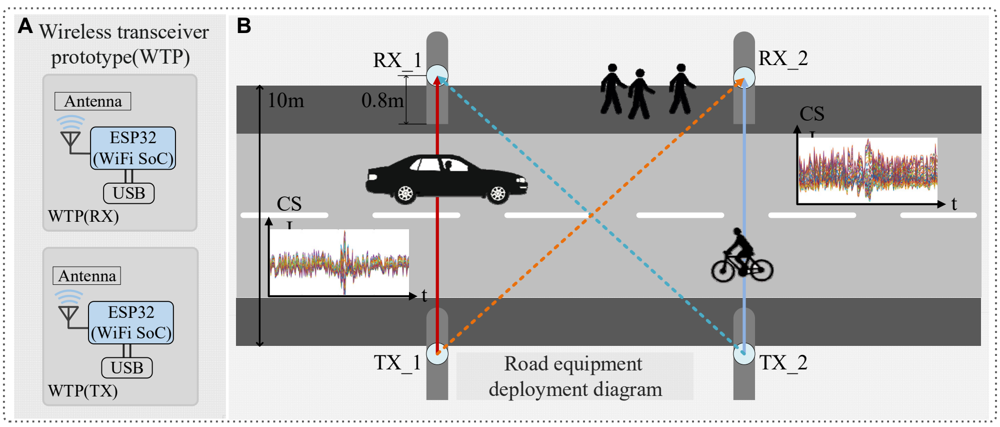


FIGURE 1
System architecture of the proposed WiHVR method. **(A)** Wireless transceiver prototype (WTP). **(B)** Road equipment deployment diagram.

along with the move of a road user, the scattered paths change in a fast speed, resulting in the variance in received signal power.

According to Eq. 4, the differences in size and speed of road users lead to different attenuation of wireless signals. Hence, the CSI readings measured by the WTP prototype are various.

3. Proposed method

According to the above analysis, since the differences in the size and speed of humans or vehicles moving on the road, the attributes of energy attenuation caused by two targets are various. In this case, it is feasible to design a WiHVR method. To this end, a deep learning-based WiHVR method is proposed in purpose of analyzing the effects of CSI on the performance of WiHVR in different application scenarios.

3.1. System overview

The overview of the proposed WiHVR based on a lightweight wireless sensing attention-based deep learning model (LW-WADL) is shown in Figure 3. The proposed WiHVR contains three key modules: Data collection, CSI preprocessing, and Deep feature extraction and classification. The data collection module consists of a pair of WTPs, both of which are made up of an ESP32 module, so as to collect CSI data of different road users in WTPs sensing area. The CSI preprocessing module includes CSI filtering, CSI augmentation, CSI subcarriers selection, and CSI segmentation. The core deep feature extraction and classification module, i.e., the proposed LW-WADL method consisting of a CBAM module and three depthwise separable convolution blocks, followed by a global average-pooling (GAP) layer for reducing computational complexity. In addition, GAP essentially is an average pooling operation which is intended to replace fully connected layers in classical CNNs. Thus, GAP is a special kind of average pooling where the sliding window of the average operation expands to the entire feature maps. Besides, after completing the final feature representations of the GAP layer, a C-class vector (C is the number of categories) is output through the Softmax function.

3.2. Data collection

As shown in Figure 3, this paper captures the CSI data in space through the developed WTP. To extract CSI data, a threshold-based road user detection algorithm is exploited in this paper. The purpose of road user detection is to find out whether there are dynamic targets in the sensing area. According to the analysis in Section "2. Preliminaries," it can be found that when there are no road users in the wireless environment, the CSI patterns stabilize around a reference value. Once a road user passes through the wireless

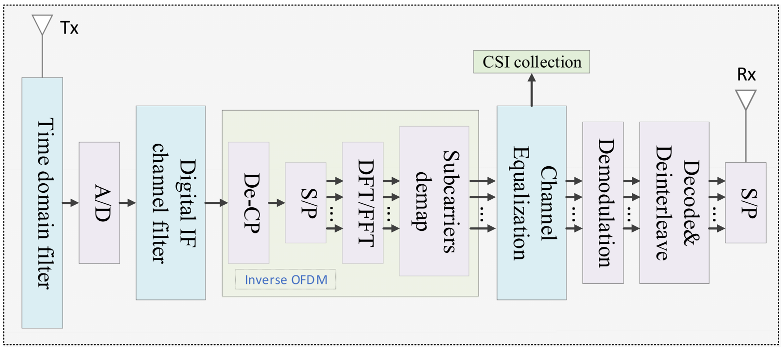


FIGURE 2
The specific process of CSI extraction.

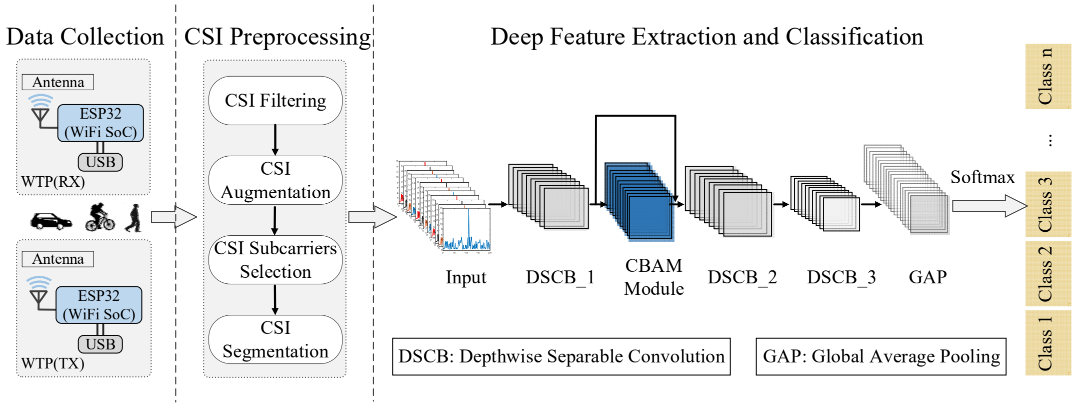


FIGURE 3
Overview of the proposed WiHVR based on a lightweight wireless sensing attention-based deep learning model (LW-WADL).

environment, the amplitude of CSI patterns will drop sharply. Therefore, the presence of road users in the region of interest can be detected by the following threshold-based algorithm:

$$X_{\det ection}[k+1] = \tag{5}$$

$$\begin{cases} Static,\ X_{\det ection}[k] = Detected \text{ and} \\ \quad \prod_{n=k-W+1}^{k} sign(|x[n] - x_{static}[n]| < T_{object}) > 0 \\ Detected,\ X_{\det ection}[k] = Static \text{ and} \\ \quad \prod_{n=k-W+1}^{k} sign(|x[n] - x_{static}[n]| \geq T_{object}) > 0 \end{cases}$$

where $x[n]$ $(n > 0)$ represents the n-th CSI reading. $x_{static}[n]$ is the average of CSI readings when there are no road users in the wireless environment, namely, CSI baseline. T_{object} is the decision threshold to determine whether there is a road user (dBm). Here, T_{object} is set to 4 dBm in this work. W is the size of the judgment window, and is set to 50 when the sampling rate is 50 Hz. $X_{\det ection}[]$ is the object detection result. "Detected" indicates that there are road users in the range of interest, and otherwise "Static" denotes no road users.

Moreover, the environmental factors such as rain, fog, temperature, etc., can affect the CSI baseline $x_{static}[]$. Thus, to improve the performance of the above-mentioned fluctuation detection algorithm, an adaptive baseline adjustment method is proposed, which can be calculated by:

$$x_{static}[n+1] = \begin{cases} \beta \cdot x_{static}[n] + (1-\beta) \cdot x[n+1], \\ \qquad X_{\det ection}[n] = X_{\det ection}[n+1] = Static \\ x_{static}[n],\ others \end{cases} \tag{6}$$

where β is a correction factor with a value of 0.96 in this paper. It can be seen from Eq. 6 that the CSI baseline will be updated as long as there are no road users in the wireless environment, otherwise it will not be updated. Hence, the problem of CSI baseline drift caused by environmental factors can be solved efficiently, as well as the robustness of the fluctuation detection algorithm can be improved.

Finally, the CSI data extracted by WTP contains 52 subcarriers, and each subcarrier contains amplitude and phase information. In order to improve execution efficiency of LW-WADL, this paper converts the raw two-dimensional CSI data containing amplitude information into one-dimensional data. Then, one-dimensional CSI data containing road user behavior information will be sent to the second stage for data preprocessing.

3.3. CSI preprocessing

The CSI preprocessing module includes the following four steps: CSI filtering, CSI augmentation, CSI subcarriers selection, and CSI segmentation.

3.3.1. CSI filtering

To guarantee the robustness of road users recognition, smoothing filtering is used to remove noise from the raw CSI data, as defined by:

$$X_{filter}(n) = \frac{1}{N} \sum_{j=1}^{N-1} X_{raw}(n-j) \tag{7}$$

where X_{raw} represent the raw CSI data, and $X_{filter}(n)$ is the average processed data and then the filter shift window size used is N, where is set to five. The raw CSI waveform vs. filtered waveform is shown in Figure 4. As can be seen from Figure 4, by applying moving average filter, the high-frequency noise has been removed from the CSI waveform without changing the trends of the waveform. The waveform changes of the filtered data (Figures 4C, D) are more pronounced than before filtering, thereby improving the efficiency and accuracy of road user detection.

3.3.2. CSI augmentation

Channel state information augmentation aims to find a way to enhance the CSI features without changing the raw CSI features. According to the characteristics of the raw CSI signal waveform, this paper proposes a novel CSI data enhancement method. This method first calculates the average value of a set of CSI amplitude, and then takes the absolute value of the CSI amplitude which is smaller than the average value. In this way, the decay of the CSI amplitude is amplified, thereby enhancing CSI features. First, the baseline X_{base} of a set of CSI data needs to be calculated, which can be expressed as:

$$X_{base} = \frac{1}{I} \cdot \frac{1}{T} \sum_{i=1}^{I} \sum_{n=1}^{T} X_{csi}(i, n) \tag{8}$$

where i represents the i-th CSI subcarrier, n represents the n-th sampling point of the i-th subcarrier. I represents the number of CSI subcarriers, which is 52 in this paper. T is the number of sampling points of a group of CSI data. The enhanced CSI data$X_{csi_aug}(i, n)$ can be obtained according to the CSI baseline X_{csi_base} :

$$X_{csi\_aug}(i, n) = \left| X_{csi}(i, n) - X_{csi\_base} \right| \tag{9}$$

where | | denotes the absolute value operation. According to the Eqs 8, 9, the enhanced CSI data can be obtained.

3.3.3. CSI subcarriers selection

Although CSI augmentation have enhanced CSI features related to road users. In practical applications, different subcarriers of CSI have different sensitivities to road users, e.g., some subcarriers fluctuate greatly when encountering road users, while other subcarriers fluctuate less. Therefore, to further enhance CSI data, we design a raw CSI subcarrier selection method to remove subcarriers with low sensitivity in CSI data. In order to evaluate the sensitivity of CSI subcarriers, this paper calculates the SNR of the CSI data amplitude, as expressed by:

$$SNR = 10 \lg \left| \frac{x_{peak} - x_{static}}{n_{noise} - x_{static}} \right| \tag{10}$$

where x_{peak} is the peak value of CSI with respect to a road user. x_{static} is the average of CSI readings when there are no road users within a wireless environment. n_{noise} is the peak value of noise. According to Eq. 10, the SNR $X_{csi\_SNR}(n)$ of all subcarriers in a set of CSI data is obtained:

$$X_{csi\_SNR}(n) = \{x_{SNR}(1), x_{SNR}(2), \ldots, x_{SNR}(m), \ldots, x_{SNR}(n)\} \tag{11}$$

where $x_{SNR}(m)$ represents the SNR value of the m-th subcarrier. For the convenience of calculation, it is assumed that $x_{SNR}(n)$has been arranged in descending order of SNR, that is, $\{x_{SNR}(1) > x_{SNR}(2) > \ldots > x_{SNR}(m) > \ldots > x_{SNR}(n)\}$. According to Eq. 11, "m" subcarriers with larger SNR are selected, where "m" is defined as the CSI factor. The selection of the CSI factor "m" is discussed in detail in Section "4. Experiment study." The mean of "m" subcarriers is calculated, which can be expressed as:

$$\bar{X}_{csi\_aug}(n) = \frac{1}{m} \sum_{n=1}^{m} X_{csi\_aug}(n) \tag{12}$$

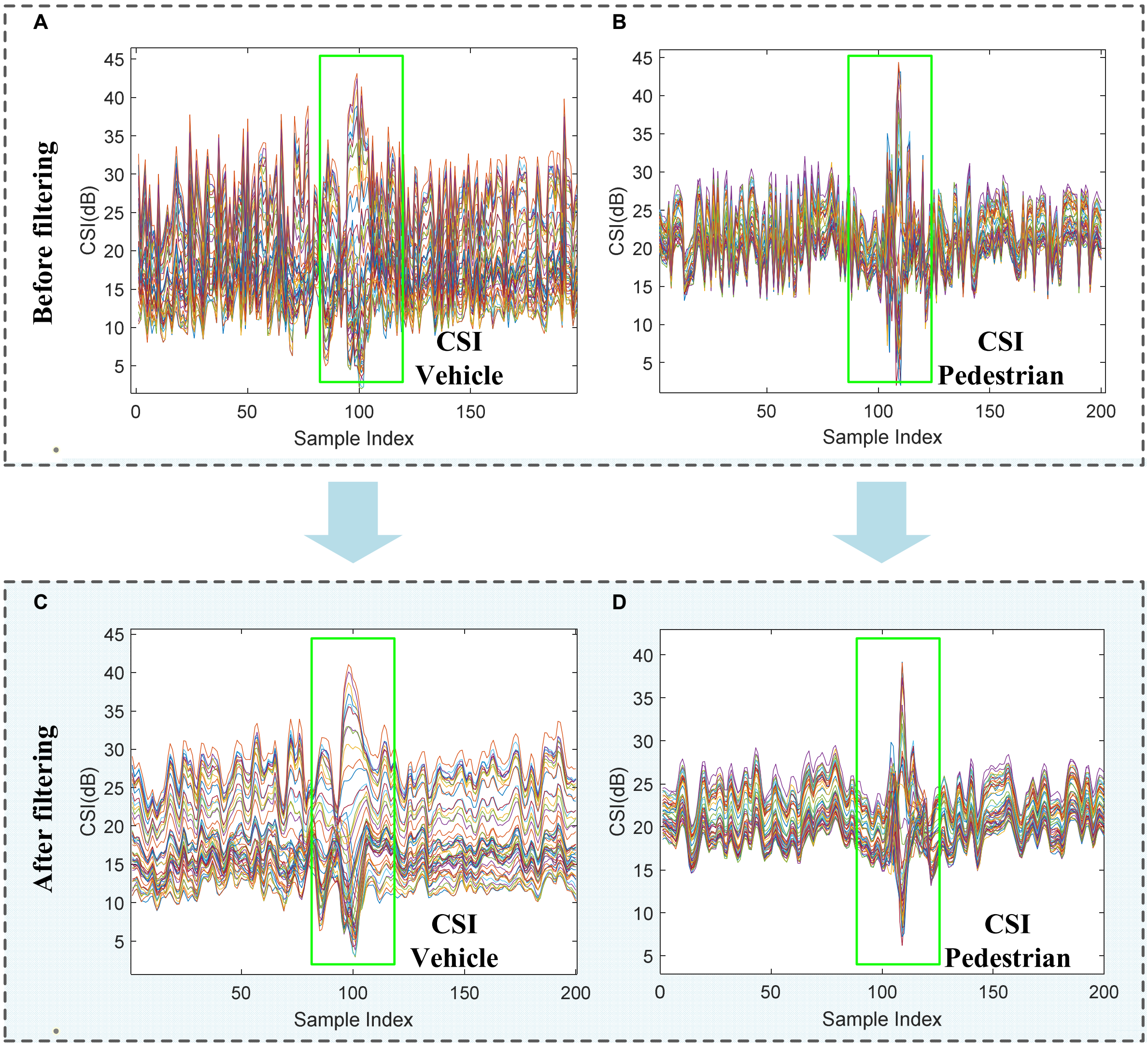


FIGURE 4
Raw CSI waveform vs. filtered waveform. **(A)** The raw CSI waveform of the vehicle. **(B)** The raw CSI waveform of the pedestrian. **(C)** The filtered CSI waveform of the vehicle. **(D)** The filtered CSI waveform of the pedestrian.

To demonstrate the validity of Eq. 12, we compare our proposed method with k-subcarriers weight fusion (Kong et al., 2019) and average-subcarriers (Wang Y. et al., 2016), as shown in **Figure 5**. It can be seen that on a pedestrian and a vehicle CSI sample, our CSI subcarrier selection method performs best, the SNR of the CSI amplitude is 3.5 and 6.1 dB, respectively.

3.3.4. CSI segmentation

The selected CSI data $\bar{X}_{csi\_aug}(n)$ containing multiple CSI features is split into certain segment-level sub-samples, each of which consists of one complete CSI feature of a road user. Specifically, the single CSI feature can be divided in terms of the local minimum in $\bar{X}_{csi\_aug}(n)$. These local minimums are defined as decision points. Specially, $x_d[i]$ represents the i-th decision point, $x_d[i]$ can be calculated as:

$$x_d[i] = \min_{s\cdot \mathrm{i}-\mathrm{W}+1\leq n\leq s\cdot i} \bar{X}_{csi\_aug}(n),\;\; i = 1, 2, 3, ..., L \qquad (13)$$

where $x_d[i]$ is the minimize value within the value range of $\bar{X}_{csi\_aug}(n)$. L is the number of decision points. w is the size of sliding window, and is set to 50, s is the step size of the window, and is set to 200. Additionally, the index of decision points in $\bar{X}_{csi\_aug}(n)$ is represented by P_i . According to Eq. 13 and P_i, $\bar{X}_{csi\_aug}(n)$ can be divided into L segments. $\tilde{x}_i[n]$ is the $_i$-th segment, which can be defined as:

$$\tilde{x}_i[n] = \left\{\bar{X}_{csi\_aug}[P_i - c_0], ..., \bar{X}_{csi\_aug}[P_i], ..., \bar{X}_{csi\_aug}[P_i + c_0]\right\} \qquad (14)$$

where c_0 is the slicing factor, and is set to 100. In this case, a new CSI dataset is developed. About 500 samples of four categories are included in the dataset: pedestrian, bicycle, motorcycle, and car.

3.4. Deep feature extraction and classification

According to the features of CSI signals containing time series features, as shown in **Figure 3**, a lightweight wireless sensing attention-based recognition algorithm, namely LW-WADL is proposed for deep feature learning from CSI features on HVR tasks. The proposed LW-WADL contains a CBAM module and three

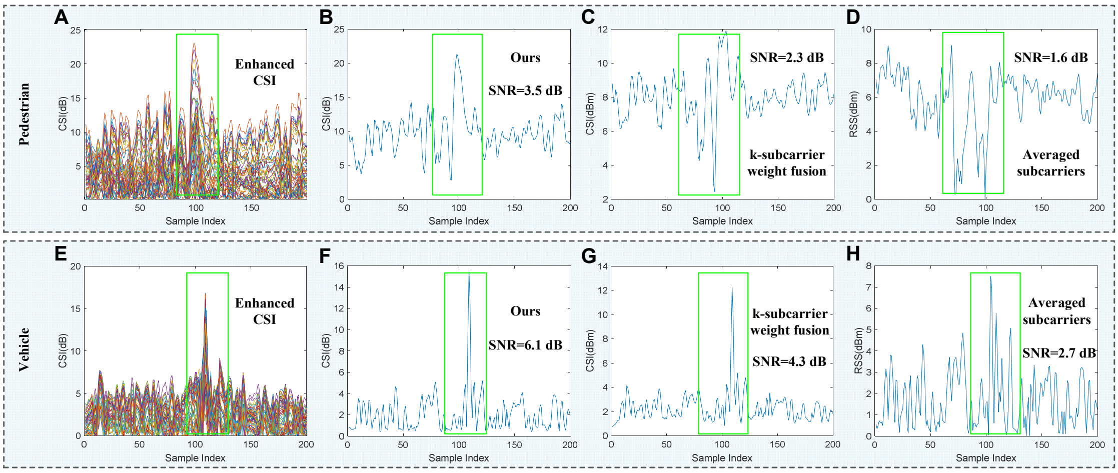


FIGURE 5
Different CSI subcarrier selection method. **(A,E)** Represent the enhanced raw CSI waveforms of the vehicle and pedestrian, respectively. **(B–D,F–H)** Represent the CSI waveforms of the vehicle and pedestrian generated by the three subcarrier selection methods, respectively.

depthwise separable convolution modules, followed by a GAP layer, as described below.

3.4.1. LW-WADL network structure

The overall network structure of the proposed LW-WADL is presented in Figure 3. LW-WADL involves of Three depthwise separable convolution blocks (DSCB_1, DSCB_2 and DSCB_3) in series. Then, in order to focus on learning the relevant information of feature maps while suppressing the irrelevant information, a CBAM module is concatenated after DSCB_1. The CBAM module can further improve the discriminating power of feature representations learned by DSCB_1. Finally, output features in DSCB_3 are achieved through a GAP layer.

3.4.2. CBAM attention module

The attention mechanism makes the model tend to pay attention to some information about the auxiliary classification in the feature map, while suppressing other useless information, thereby improving the classification ability of the model. The CBAM module consists of a channel attention module and a spatial attention module. The detailed structure is shown in Figure 6.

The channel attention module first performs maxpooling and average pooling based on the height and width of the DSCB_1 feature map to obtain two one-dimensional vectors. Then, it is input into the shared multi-layer perceptron (Shared MLP), and the corresponding elements of the output features of the MLP are summed point by point. The result is input into the Sigmoid activation function, and then the inner product operation is performed with the initial input feature map. The final output feature map is used as the input of the spatial attention module.

The spatial attention module performs maxpooling and average pooling based on the channel, and then uses the convolution (abbreviated as Conv) operation to merge the output features on the channel dimension. The merged features are input into a sigmoid activation function, then an inner product operation is performed on the obtained output features and the input of the spatial attention module. Finally, the output of the inner product operation is combined with the output of the DSCB_1 module to form the input features of the DSCB_2 module.

3.4.3. Softmax classifier output

WiHVR is fundamentally a multi-classification task, so we choose the Softmax function to produce final classification results. Through the Softmax function, the output values of classifier can be converted into a probability distribution in the range [0, 1].

The cross-entropy loss function is implemented as the training objective function for LW-WADL:

$$L_{loss} = -\sum_{i} \widehat{y}_i \log(y_i) \tag{15}$$

where $\hat{y}_i = 1$ if the class is i, otherwise $\hat{y}_i = 0$. y_i represents the output of the LW-WADL model, the probability that the class is i. L_{loss} is a loss measure of the difference between two probability distributions.

4. Experiment study

4.1. Experiment setup

As can be seen from Figure 7, the proposed WTP prototype contains two main components: antenna, ESP32. ESP32 is a WiFi SoC working at a frequency of 2.4 GHz. In the experiment, the two WTP prototypes were installed on both sides of a road with a width of 10 m, and antenna heights is set to 1 m. In addition, for training LW-WADL models, the Adam optimizer with a learning rate of 0.0001 is used. The batch-size is 16 and the maximum of epochs is 200. Besides, to explore the effects of CSI on the performance of WiHVR in different application scenarios, the developed CSI dataset is divided into three taxonomies according to the number of categories, namely, two-category dataset, three-category dataset and four-category dataset. Finally, 80% of the data in the dataset is used as the training set, while the rest is used for testing.

4.2. Evaluation indicators

The performance of the designed LW-WADL is evaluated by three typical metrics such as “Accuracy,” “Recall,” and “Precision.”

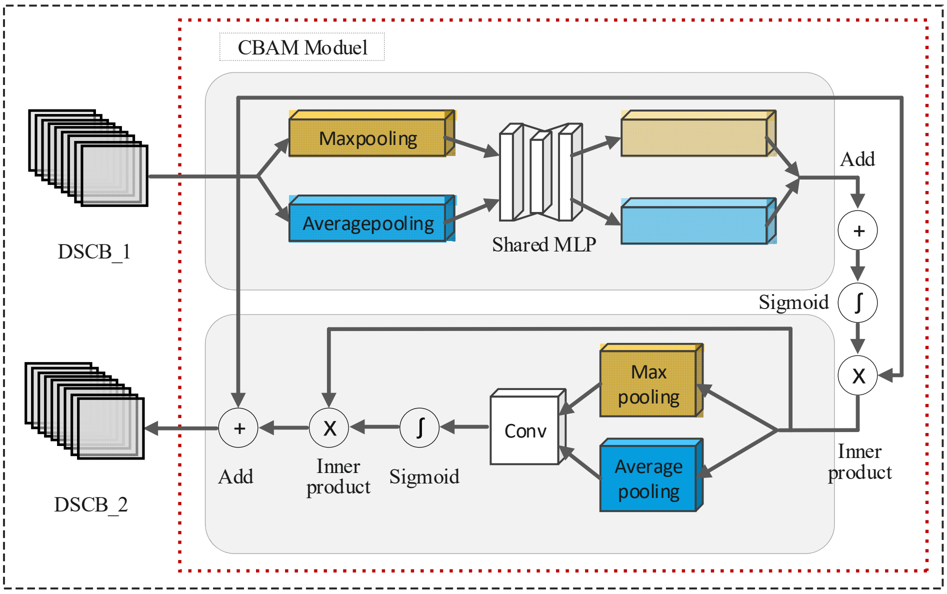


FIGURE 6
CBAM structure diagram.

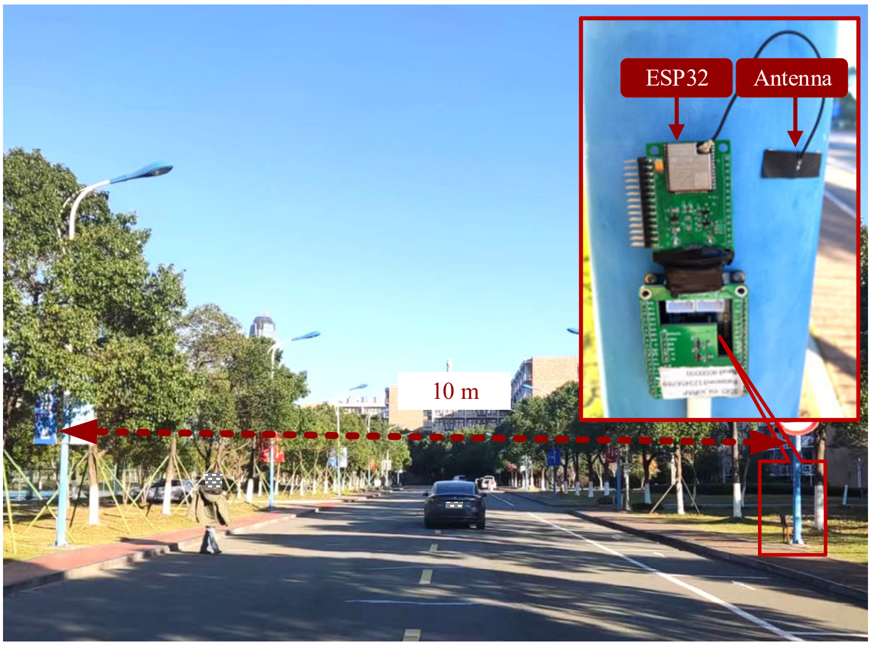


FIGURE 7
Experimental scenarios and WTP installation details.

For the computational complexity analysis of deep learning methods, two well-known computational indicators, the network parameters (abbreviated as param.) and floating-point operations (FLOPs) are employed. Specifically, "Accuracy" is the ratio of all correct predictions to the whole number of predictions. "Precision" is the ratio of correct predictions with positive values to total predictions with positive values. "Recall" is the ratio of predicted positives to the total number of actual positives. They are defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + TN} \tag{16}$$

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

where TP denotes the number of true positive samples classified as positive. FP denotes the number of true negative samples classified as positive. FN denotes the number of true positive samples classified as negative. TN denotes the number of true negative samples classified as negative.

4.3. Comparison of different methods

To verify the effectiveness of our LW-WADL for WiHVR, we adopt the developed CSI four-category dataset, so as to compare the performance of various deep learning models on WiHVRs tasks. Table 1 present a performance comparison of four methods, including full convolutional network (FCN) (Long et al., 2015), DeepWiTraffic (Won et al., 2019), and deep residual network (ResNet) (He et al., 2016). Among them, FCN and ResNet are used as baseline methods to provide benchmarking performance, whereas DeepWiTraffic is used as comparing work. Besides, FCN is composed of three convolutional layers. ResNet contains three residual blocks. DeepWiTraffic contains two convolutional layers and two max pooling layers. Experimental results are shown in Table 1.

From Table 1, it can be found that the designed deep learning model has the highest classification performance with an accuracy of 96.26%, a precision of 96.23%, and a recall of 96.16%. Compared with DeepWitraffic, our model not only makes an improvement of 1.67%, but also exhibits much lower computational complexity in which 93.25% of the parameters (Param, FLOPs) can be reduced. Moreover, the test time is just 0.0575 s, which is much less than DeepWitraffic. This shows that our model is a lightweight model. Additionally, compared with FCN and ResNet, our method yields an accuracy improvement of 2.8 and 1.87%.

4.4. Selection of CSI factor m

A set of experiments are designed to investigate the effect of CSI factor "m" on the accuracy of WiHVR. Figure 8 shows the accuracy of the WiHVR for different "m," where "all" represents the maximum "m." Experiments are performed on the CSI four-category dataset. Each CSI factor "m" corresponds to a CSI four-category dataset, and these datasets are identical except for the CSI factor "m." To make the results more reliable, the ResNet, DeepWiTraffic, FCN, and our model are used. The experimental results are shown in Figure 8.

As shown in Figure 8, with the increase of "m," both the ResNet, DeepWiTraffic, FCN, and ours model show a trend of increasing first and then decreasing, reaching the highest accuracy of 94.39, 94.59, 93.46, and 96.26% when "m" is 4, respectively. It can be found that only one subcarrier with the highest SNR or the average of all subcarriers cannot obtain the best HVR performance. This is because the sensitivity of different CSI subcarriers varies greatly. Some subcarriers are less sensitive, while some subcarriers with higher sensitivity are too sensitive to environmental changes, resulting in reduced recognition ability.

4.5. Comparison of CSI subcarrier selection methods

To verify the performance of the proposed CSI subcarrier selection method, we conduct comparative experiments on the CSI four-category dataset. Recently-merged CSI subcarrier selection methods such as k-subcarrier weight fusion (Kong et al., 2019) and averaged subcarriers (Wang Y. et al., 2016) are used for comparative experiments. The recognition accuracy of FCN, DeepWiTraffic, ResNet, and our LW-WADL under different CSI subcarrier selection methods are shown in Table 2.

As shown in Table 2, the four used models perform best under our CSI subcarrier selection method, and the accuracy of HVR is 93.46, 94.59, 94.39, and 96.26%, respectively. The results show the superiority of our CSI subcarrier selection method, which is consistent with the conclusion drawn in Figure 5. In addition, the

TABLE 1 Performance comparison of different methods on four-category dataset.

Method	Accuracy (%)	Precision (%)	Recall (%)	Param (M)	FLOPs (M)	Testing time(s)
FCN	93.46%	93.34%	93.46%	1.6451	3.2888	0.2019
DeepWiTraffic	94.59%	94.45%	94.59%	0.2090	0.4176	0.1398
ResNet	94.39%	95.28%	94.39%	0.7427	1.5047	0.2982
Ours	**96.26%**	**96.23%**	**96.16%**	**0.0123**	**0.0282**	**0.0575**

The bold values represent the indicator with the best performance, that is, the highest Accuracy, Precision, and Recall as well as the lowest Param, FLOPs, and Testing time.

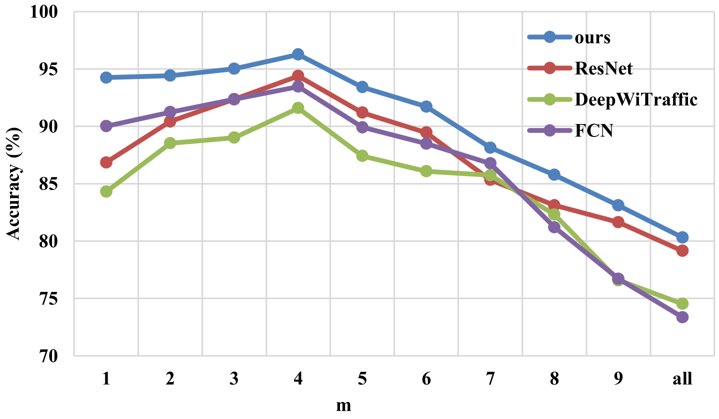


FIGURE 8
The classification accuracy of different CSI factors "m" on the CSI four-category dataset.

TABLE 2 Accuracy of different subcarrier selection methods.

CSI subcarrier selection methods	FCN	DeepWiTraffic	ResNet	LW-WADL
k-subcarrier weight fusion	89.17%	90.86%	90.77%	93.01%
Averaged subcarriers	82.09%	87.89%	83.36%	90.53%
Ours	**93.46%**	**94.59%**	**94.39%**	**96.26%**

The bold values represent that our subcarrier selection method achieves the highest accuracy among the four compared deep learning models.

TABLE 3 Performance evaluation of different classification tasks.

Tasks	FCN	DeepWiTraffic	ResNet	LW-WADL
2	100%	100%	100%	**100%**
3	95.12%	96.86%	96.21%	**97.96%**
4	93.46%	94.59%	94.39%	**96.26%**

The bold values indicate that the designed deep learning model (LW-WADL) outperforms on different classification tasks.

k-subcarrier weight fusion and averaged subcarriers methods will not remove those CSI subcarriers with too low or too high sensitivity, which may have a negative impact on CSI waveform. In this case, the accuracy of the above two methods in Table 2 is lower than that of our proposed method.

4.6. Performance evaluation of different classification tasks

To explore the performance of CSI signals on different classification tasks, three groups of experiments are set up, namely two-classification tasks, three-classification tasks, and four-classification tasks. Each group of experiments selects four methods to test, FCN, DeepWiTraffic, ResNet, and our LW-WADL, respectively, so that the results are more credible. The experimental results are shown in Table 3.

The results of Table 3 shows that the compared methods perform best and the same on the two-classification task, and the accuracy of HVR reaches 100%. However, with the increase of road user categories, the classification accuracy of the used four methods decrease. For three-classification task, the classification accuracy of four methods are 95.12, 96.86, 96.21, and 97.96%, respectively. For four-classification task, the classification accuracy of four methods is 1.66, 2.27, 1.82, and 1.73% lower than three-classification task. Overall, our method achieves more than 96% accuracy in different classification tasks.

4.7. CSI confusion matrices

To further display the recognition accuracy for each class of road users, Figure 9 shows confusion matrices of the classification

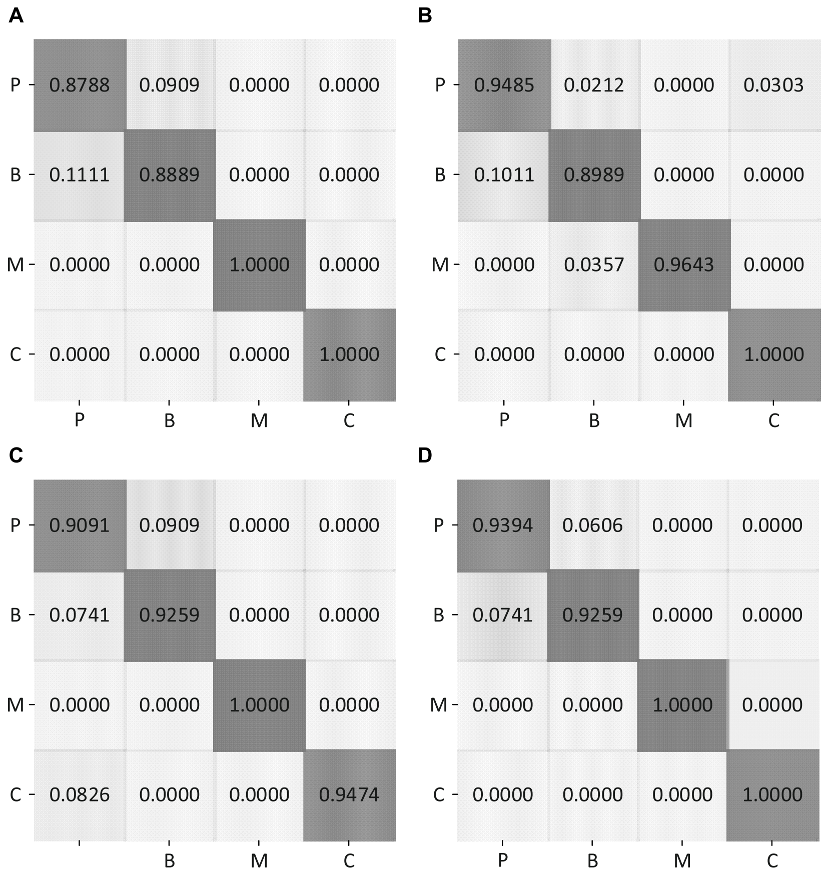


FIGURE 9
Confusion matrices of four methods on CSI four-classification dataset. **(A)** CSI confusion matrix of FCN. **(B)** CSI confusion matrix of DeepWiTraffic. **(C)** CSI confusion matrix of ResNet. **(D)** CSI confusion matrix of ours. P, pedestrian; B, bicycle; M, motorcycle; C: car.

results, when FCN, DeepWiTraffic, ResNet, and our LW-WADL methods obtain 93.46, 94.59, 94.39, and 96.26% accuracy. As shown in Figure 9, the "car" category is well-recognized for most of used methods. Among all the models, only ResNet incorrectly classify the "car" as the "pedestrian." This may be attributed to the fact that a car usually has a much larger volume than other road users. As a result, the attenuation of CSI readings caused by cars is quite different from those cases caused by other road users. For all used models, a major part of the error arises from misclassifying "pedestrian" as "bicycle." This reveals that the group, i.e., "pedestrian" vs. "bicycle" is easily confused with each other. This phenomenon is hinted by the overlap among some real-world road user shapes.

5. Conclusion and future work

This paper has proposed a lightweight wireless sensing attention-based deep learning model (LW-WADL). In order to evaluate the classification ability of LW-WADL, three CSI-based datasets are established, namely two-category dataset, three-category dataset. and four-category dataset. The experimental results on the developed dataset show that the classification accuracy of LW-WADL decreases with the increase of road user categories, but it is higher than 96%. In addition, this paper provides a novel CSI subcarrier selection method, which calculates the SNR of all subcarriers and selects the first four subcarriers with larger SNR for fusion. Besides, a new CSI data enhancement method is exploited to preprocess the change trend of CSI data to one direction, thereby enhancing CSI data.

In future, the performance of other advanced deep learning-based WiHVR methods will be investigated. It is also significant to explore the human-vehicle recognition task based on multiple sets of WTPs. Additionally, it is meaningful to explore the applications of the proposed methods in real scenarios such as multi-lane roads.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

MS was responsible for the writing of the manuscript and some experiments. RZ was responsible for some experiments of the manuscript. XC was responsible for the data collection and processing of the manuscript. CZ was responsible for the revision of the manuscript. LL provided the fund support for this project. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by Zhejiang Provincial National Science Foundation of China under Grant No. LGG22F030009 and partially supported by Taizhou Science and Technology Plan Project under Grant No. 21gya29.

Conflict of interest

RZ was employed by China United Network Communications Co., Ltd., Taizhou Branch.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abdelnasser, H., Harras, K., and Youssef, M. (2018). A ubiquitous WiFi-based fine-grained gesture recognition system. *IEEE Trans. Mobile Comput.* 18, 2474–2487. doi: 10.1109/TMC.2018.2879075

Arshad, S., Feng, C., Elujide, I., Zhou, S., and Liu, Y. (2018). "SafeDrive-Fi: A multimodal and device free dangerous driving recognition system using WiFi," in *Proceedings of the 2018 IEEE international conference on communications (ICC)* (Kansas City, MO: IEEE), 1–6. doi: 10.1109/ICC.2018.8422431

Bhat, S. A., Mehbodniya, A., Alwakeel, A. E., Webber, J., and Al-Begain, K. (2020). "Human motion patterns recognition based on rss and support vector machines," in *Proceedings of the 2020 IEEE wireless communications and networking conference (WCNC)* (Seoul: IEEE), 1–6. doi: 10.1109/WCNC45663.2020.9120797

Chen, Z., Zhang, L., Jiang, C., Cao, Z., and Cui, W. (2018). WiFi CSI based passive human activity recognition using attention based BLSTM. *IEEE Trans. Mobile Comput.* 18, 2714–2724. doi: 10.1109/TMC.2018.2878233

Chollet, F. (2017). "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Honolulu, HI, 1251–1258. doi: 10.1109/CVPR.2017.195

Choy, J. L. C., Wu, J., Long, C., and Lin, Y.-B. (2020). Ubiquitous and low power vehicles speed monitoring for intelligent transport systems. *IEEE Sens. J.* 11, 5656–5665.

de Oliveira, L. F. P., Manera, L. T., and Da Luz, P. D. G. (2020). Development of a smart traffic light control system with real-time monitoring. *IEEE Internet Things J.* 8, 3384–3393. doi: 10.1109/JIOT.2020.3022392

Du, Y., Qin, B., Zhao, C., Zhu, Y., Cao, J., and Ji, Y. (2021). A novel spatio-temporal synchronization method of roadside asynchronous MMW radar-camera for sensor fusion. *IEEE Trans. Intell. Transp. Syst.* 23, 22278–22289. doi: 10.1109/TITS.2021.3119079

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, 770–778. doi: 10.1109/CVPR.2016.90

Huang, Q., Cai, Z., and Lan, T. (2020). A new approach for character recognition of multi-style vehicle license plates. *IEEE Trans. Multimed.* 23, 3768–3777. doi: 10.1109/TMM.2020.3031074

Jiang, Y., Shuai, Y., He, X., Wen, X., and Lou, L. (2021). An energy-efficient street lighting approach based on traffic parameters measured by wireless sensing technology. *IEEE Sens. J.* 21, 19134–19143. doi: 10.1109/JSEN.2021.3089208

Jin, J., and Ma, X. (2019). A multi-objective agent-based control approach with application in intelligent traffic signal system. *IEEE Trans. Intell. Transp. Syst.* 20, 3900–3912. doi: 10.1109/TITS.2019.2906260

Kong, H., Lu, L., Yu, J., Chen, Y., Kong, L., and Li, M. (2019). "Fingerpass: Finger gesture-based continuous user authentication for smart homes using commodity wifi," in *Proceedings of the twentieth ACM international symposium on mobile ad hoc networking and computing*, Catania, 201–210. doi: 10.1145/3323679.3326518

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* 25, 1097–1105.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Liu, J., Wang, L., Guo, L., Fang, J., Lu, B., and Zhou, W. (2017). "A research on CSI-based human motion detection in complex scenarios," in *Proceedings of the 2017 IEEE 19th international conference on e-health networking, applications and services (Healthcom)* (Dalian: IEEE), 1–6. doi: 10.1109/HealthCom.2017.8210800

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Boston, MA, 3431–3440. doi: 10.1109/CVPR.2015.7298965

Ma, X., Zhao, Y., Zhang, L., Gao, Q., Pan, M., and Wang, J. (2019). Practical device-free gesture recognition using WiFi signals based on metalearning. *IEEE Trans. Industr. Inform.* 16, 228–237. doi: 10.1109/TII.2019.2909877

Otter, D. W., Medina, J. R., and Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. *IEEE Trans. Neural Netw. Learn. Syst.* 32, 604–624. doi: 10.1109/TNNLS.2020.2979670

Pan, X., Jiang, T., Li, X., Ding, X., Wang, Y., and Li, Y. (2019). "Dynamic hand gesture detection and recognition with WiFi signal based on 1d-CNN," in *Proceedings of the 2019 IEEE international conference on communications workshops (ICC Workshops)* (Shanghai: IEEE), 1–6. doi: 10.1109/ICCW.2019.8756690

Park, J.-K., Choi, I.-O., and Kim, K.-T. (2021). Length prediction of moving vehicles using a commercial FMCW radar. *IEEE Trans. Intell. Transp. Syst.* 23, 14833–14845. doi: 10.1109/TITS.2021.3134408

Santos, P. M., Rodrigues, J. G. P., Cruz, S. B., Lourenço, T., d'Orey, P. M., Luis, Y., et al. (2018). PortoLivingLab: An IoT-based sensing platform for smart cities. *IEEE Internet Things J.* 5, 523–532. doi: 10.1109/JIOT.2018.2791522

Singh, R., Saluja, D., and Kumar, S. (2021). R-comm: A traffic based approach for joint vehicular radar-communication. *IEEE Trans. Intell. Vehicles* 7, 83–92. doi: 10.1109/TIV.2021.3074389

Sliwa, B., Piatkowski, N., and Wietfeld, C. (2020). The channel as a traffic sensor: Vehicle detection and classification based on radio fingerprinting. *IEEE Internet Things J.* 7, 7392–7406. doi: 10.1109/JIOT.2020.2983207

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, 2818–2826. doi: 10.1109/CVPR.2016.308

Tavanti, E., Rizik, A., Fedeli, A., Caviglia, D. D., and Randazzo, A. (2021). A short-range FMCW radar-based approach for multi-target human-vehicle detection. *IEEE Trans. Geosci. Remote Sens.* 60, 1–16. doi: 10.1109/TGRS.2021.3138687

Wang, F., Gong, W., and Liu, J. (2018). On spatial diversity in WiFi-based human activity recognition: A deep learning-based approach. *IEEE Internet Things J.* 6, 2035–2047. doi: 10.1109/JIOT.2018.2871445

Wang, J., Zhao, Y., Fan, X., Gao, Q., Ma, X., and Wang, H. (2018). Device-free identification using intrinsic CSI features. *IEEE Trans. Vehicular Technol.* 67, 8571–8581. doi: 10.1109/TVT.2018.2853185

Wang, Q., Zheng, J., Xu, H., Xu, B., and Chen, R. (2017). Roadside magnetic sensor system for vehicle detection in urban environments. *IEEE Trans. Intell. Transp. Syst.* 19, 1365–1374. doi: 10.1109/TITS.2017.2723908

Wang, W., Liu, A. X., and Shahzad, M. (2016). "Gait recognition using wifi signals," in *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*, Heidelberg, 363–373. doi: 10.1145/2971648.2971670

Wang, Y., Wu, K., and Ni, L. M. (2016). Wifall: Device-free fall detection by wireless networks. *IEEE Trans. Mobile Comput.* 16, 581–594. doi: 10.1109/TMC.2016.2557792

Wilby, M. R., González, A. B. R., Pozo, R. F., and Díaz, J. J. V. (2020). Short-term prediction of level of service in highways based on bluetooth identification. *IEEE Trans. Intell. Transp. Syst.* 23, 142–151. doi: 10.1109/TITS.2020.3008408

Won, M., Sahu, S., and Park, K.-J. (2019). "DeepWiTraffic: Low cost WiFi-based traffic monitoring system using deep learning," in *Proceedings of the 2019 IEEE 16th international conference on mobile ad hoc and sensor systems (MASS)* (Monterey, CA: IEEE), 476–484. doi: 10.1109/MASS.2019.00062

Won, M., Zhang, S., and Son, S. H. (2017). "WiTraffic: Low-cost and non-intrusive traffic monitoring system using WiFi," in *Proceedings of the 2017 26th international conference on computer communication and networks (ICCCN)* (Vancouver, BC: IEEE), 1–9. doi: 10.1109/ICCCN.2017.8038380

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). "Cbam: Convolutional block attention module," in *Proceedings of the european conference on computer vision (ECCV)*, Munich, 3–19. doi: 10.1007/978-3-030-01234-2_1

Zhang, S., Zhang, S., Huang, T., and Gao, W. (2017). Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Trans. Multimed.* 20, 1576–1590. doi: 10.1109/TMM.2017.2766843

Zhang, X., Tang, C., Yin, K., and Ni, Q. (2021). WiFi-based cross-domain gesture recognition via modified prototypical networks. *IEEE Internet Things J.* 9, 8584–8596. doi: 10.1109/JIOT.2021.3114309

Zhao, L., and Huang, Z. (2021). A moving object detection method using deep learning-based wireless sensor networks. *Complexity* 2021:5518196. doi: 10.1155/2021/5518196

Zhao, L., Wang, J., Liu, J., and Kato, N. (2019). Routing for crowd management in smart cities: A deep reinforcement learning perspective. *IEEE Commun. Mag.* 57, 88–93. doi: 10.1109/MCOM.2019.1800603

TYPE Review
PUBLISHED 27 February 2023
DOI 10.3389/fpsyg.2023.1126994

Check for updates

OPEN ACCESS

EDITED BY
Wenming Zheng,
Southeast University,
China

REVIEWED BY
Guijun Chen,
Taiyuan University of Technology,
China
Shou-Yu Chen,
Tongji University,
China

*CORRESPONDENCE
Yinzhen Huang
✉ hyz@hnit.edu.cn

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Perception Science, a section of the journal Frontiers in Psychology

RECEIVED 19 December 2022
ACCEPTED 11 January 2023
PUBLISHED 27 February 2023

CITATION
Wang X, Ren Y, Luo Z, He W, Hong J and Huang Y (2023) Deep learning-based EEG emotion recognition: Current trends and future perspectives.
Front. Psychol. 14:1126994.
doi: 10.3389/fpsyg.2023.1126994

Deep learning-based EEG emotion recognition: Current trends and future perspectives

Xiaohu Wang[1†], Yongmei Ren[2†], Ze Luo[1], Wei He[2], Jun Hong[1] and Yinzhen Huang[3]*

[1]School of Intelligent Manufacturing and Mechanical Engineering, Hunan Institute of Technology, Hengyang, China, [2]School of Electrical and Information Engineering, Hunan Institute of Technology, Hengyang, China, [3]School of Computer and Information Engineering, Hunan Institute of Technology, Hengyang, China

Automatic electroencephalogram (EEG) emotion recognition is a challenging component of human–computer interaction (HCI). Inspired by the powerful feature learning ability of recently-emerged deep learning techniques, various advanced deep learning models have been employed increasingly to learn high-level feature representations for EEG emotion recognition. This paper aims to provide an up-to-date and comprehensive survey of EEG emotion recognition, especially for various deep learning techniques in this area. We provide the preliminaries and basic knowledge in the literature. We review EEG emotion recognition benchmark data sets briefly. We review deep learning techniques in details, including deep belief networks, convolutional neural networks, and recurrent neural networks. We describe the state-of-the-art applications of deep learning techniques for EEG emotion recognition in detail. We analyze the challenges and opportunities in this field and point out its future directions.

KEYWORDS

human–computer interaction, electroencephalogram, emotion recognition, deep learning, survey

1. Introduction

Emotion recognition (or detection) is a major scientific problem in affective computing, which mainly solves the problem of computer systems accurately processing, recognizing, and understanding the emotional information expressed by human beings. Affective computing requires interdisciplinary knowledge, including psychology, biology, and computer science. As emotion plays a key role in the field of human–computer interaction (HCI) and artificial intelligence, it has recently received extensive attention in the field of engineering research. Research of emotion recognition technology can further promote the development of various disciplines, including computer science, psychology, neuroscience, human factors engineering, medicine, and criminal investigation.

As a complex psychological state, emotion is related to physical behavior and physiological activities (Cannon, 1927). Researchers have conducted numerous studies to enable computers to correctly distinguish and understand human emotions. These studies aim to enable computers to generate various emotional features similar to human beings, so as to achieve the purpose of natural, sincere, and vivid interaction with human beings. Some of these methods mainly use non physiological signals, such as speech (Zhang et al., 2017; Khalil et al., 2019; Zhang S. et al., 2021), facial expression (Alreshidi and Ullah, 2020), and body posture (Piana et al., 2016). However, their accuracy depends on people's age and cultural characteristics, which are subjective, so accurately judging the true feelings of others is difficult. Other methods use physiological activities (or physiological clues), such as heart rate (Quintana et al., 2012), skin impedance (Miranda et al., 2018), respiration (Valderas et al., 2019) or brain signals, functional magnetic resonance imaging (Chen et al., 2019a), magnetoencephalography (Kajal et al., 2020), and electroencephalography, to identify emotional states. Some studies have shown that physiological activities and emotional expression are correlated, although the sequence of the two processes is still debated (Cannon, 1927). Therefore, the method based on calculating physiological signals is considered an effective supplement to the recognition method based on nonphysiological signals. The subject cannot control the automatically generated electroencephalogram (EEG) signal. For those who cannot speak clearly and express their feelings through natural speech or have physical disabilities and cannot express their feelings through facial expressions or body postures, emotion recognition of voice, expression, and posture becomes impossible. Therefore, EEG is an appropriate means to extract human emotions, and studying emotional cognitive mechanisms and recognizing emotional states by directly using brain activity information, such as EEG, are particularly important.

From the perspective of application prospects, EEG-based emotion recognition technology has penetrated into various fields, including medical, education, entertainment, shopping, military, social, and safe driving (Suhaimi et al., 2020). In the medical field, timely acquisition of patients' EEG signals and rapid analysis of their emotional state can help doctors and nurses to accurately understand the patients' psychological state and then make reasonable medical decisions, which has an important effect on the rehabilitation of some people with mental disorders, such as autism (Mehdizadehfar et al., 2020; Mayor-Torres et al., 2021; Ji et al., 2022), depression (Cai et al., 2020; Chen X. et al., 2021), Alzheimer's disease (Güntekin et al., 2019; Seo et al., 2020), and physical disabilities (Chakladar and Chakraborty, 2018). In terms of education, the emotion recognition technology based on EEG signals can enable teaching staff to adjust teaching methods and teaching attitudes in a timely manner in accordance with the emotional performance of different trainees in class, such as increasing or reducing the workload (Menezes et al., 2017). In terms of entertainment, such as computer games, researchers try to detect the emotional state of players to adapt to the difficulty, punishment, and encouragement of the game (Stavroulia et al., 2019). In the military aspect, the emotional status of noncommissioned officers and soldiers can be captured timely and accurately through EEG signals, so that the strategic layout can be adjusted in time to improve the winning rate of war (Guo et al., 2018). In terms of social networks, we can enhance barrier-free communication in the HCI system, increase the mutual understanding and interaction in the human–machine–human interaction channel, and avoid some unnecessary misunderstandings and frictions through the acquisition of emotional information (Wu et al., 2017). In terms of safe driving, timely detection of EEG emotional conditions can enable a vehicle to perform intelligent locking during startup to block driving or actively open the automatic driving mode to intervene in the vehicle's motion trajectory until parking at a safe position, thereby greatly reducing the occurrence of accidents (Fan et al., 2017).

Recently, automatic recognition of emotional information from EEG has become a challenging problem, and has attracted extensive attention in the fields of artificial intelligence and computer vision. The flow of emotion recognition research is shown in Figure 1. Essentially, human emotion recognition using EEG signals belongs to one type of pattern recognition research.

In the early EEG-based automated emotion recognition literature, a variety of machine learning-based studies, such as support vector machine (SVM; Lin et al., 2009; Nie et al., 2011; Jie et al., 2014; Candra et al., 2015), k-nearest neighbor (KNN; Murugappan et al., 2010; Murugappan, 2011; Kaundanya et al., 2015), linear regression (Bos, 2006; Liu et al., 2011), support vector regression (Chang et al., 2010; Soleymani et al., 2014), random forest (Lehmann et al., 2007; Donos et al., 2015; Lee et al., 2015), and decision tree (Kuncheva et al., 2011; Chen et al., 2015), have been developed.

Although the abovementioned hand-crafted EEG signal features associated with machine learning approaches can produce good domain-invariant features for EEG emotion recognition, they are still low-level and not highly discriminative. Thus, obtaining high-level domain-invariant feature representations for EEG emotion recognition is desirable.

The recently-emerged deep learning methods may present a possible solution to achieve high-level domain-invariant feature representations and high-precision classification results of EEG emotion recognition. The representative deep leaning techniques contain recurrent neural networks (RNNs; Elman, 1990), long short-term memory (LSTM; Hochreiter and Schmidhuber, 1997; Zhang et al., 2019), deep belief networks (DBNs; Hinton et al., 2006), and convolutional neural networks (CNNs; Krizhevsky et al., 2012). To date, deep learning techniques have shown outstanding performance on object detection and classification (Wu et al., 2020), natural language processing (Otter et al., 2020), speech signal processing (Purwins et al., 2019), and multimodal emotion recognition (Zhou et al., 2021) due to its strong feature learning ability. Figure 2 shows the evolution of EEG emotion recognition with deep learning algorithms, emotion categories and databases.

Inspired by the lack of summarizing the recent advances in various deep learning techniques for EEG-based emotion recognition, this paper aims to present an up-to-date and comprehensive survey of EEG emotion recognition, especially for various deep learning techniques in this area. This paper highlights the different challenges and opportunities on EEG emotion recognition tasks and points out its future trends. In this survey, we have searched the published literature between January 2012, and December 2022 through Scholar. google, ScienceDirect, IEEEXplore, ACM, Springer, PubMed, and Web of Science, on the basis of the following keywords: "EEG emotion recognition," "emotion computing," "deep learning," "RNNs," "LSTM," "DBNs," and "CNNs." There is no any language restriction for the searching process.

In this work, our contributions can be summarized as follows:

1. We provide an up-to-date literature survey on EEG emotion recognition from a perspective of deep learning. To the best of our knowledge, this is the first attempt to present a comprehensive review covering EEG emotion recognition and deep learning-based feature extraction algorithms in this field.

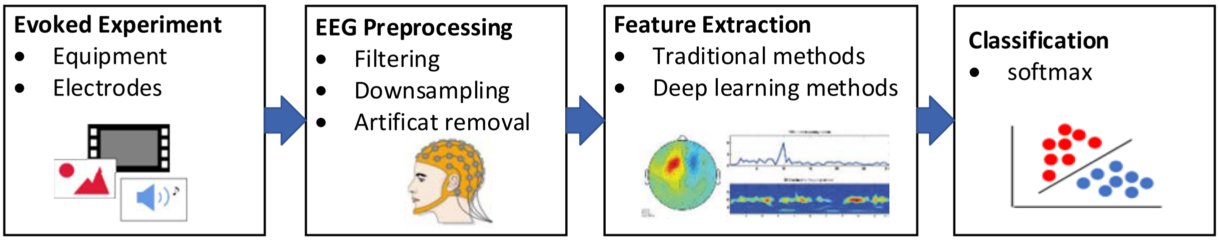


FIGURE 1
Flowchart of emotion recognition using EEG signals.

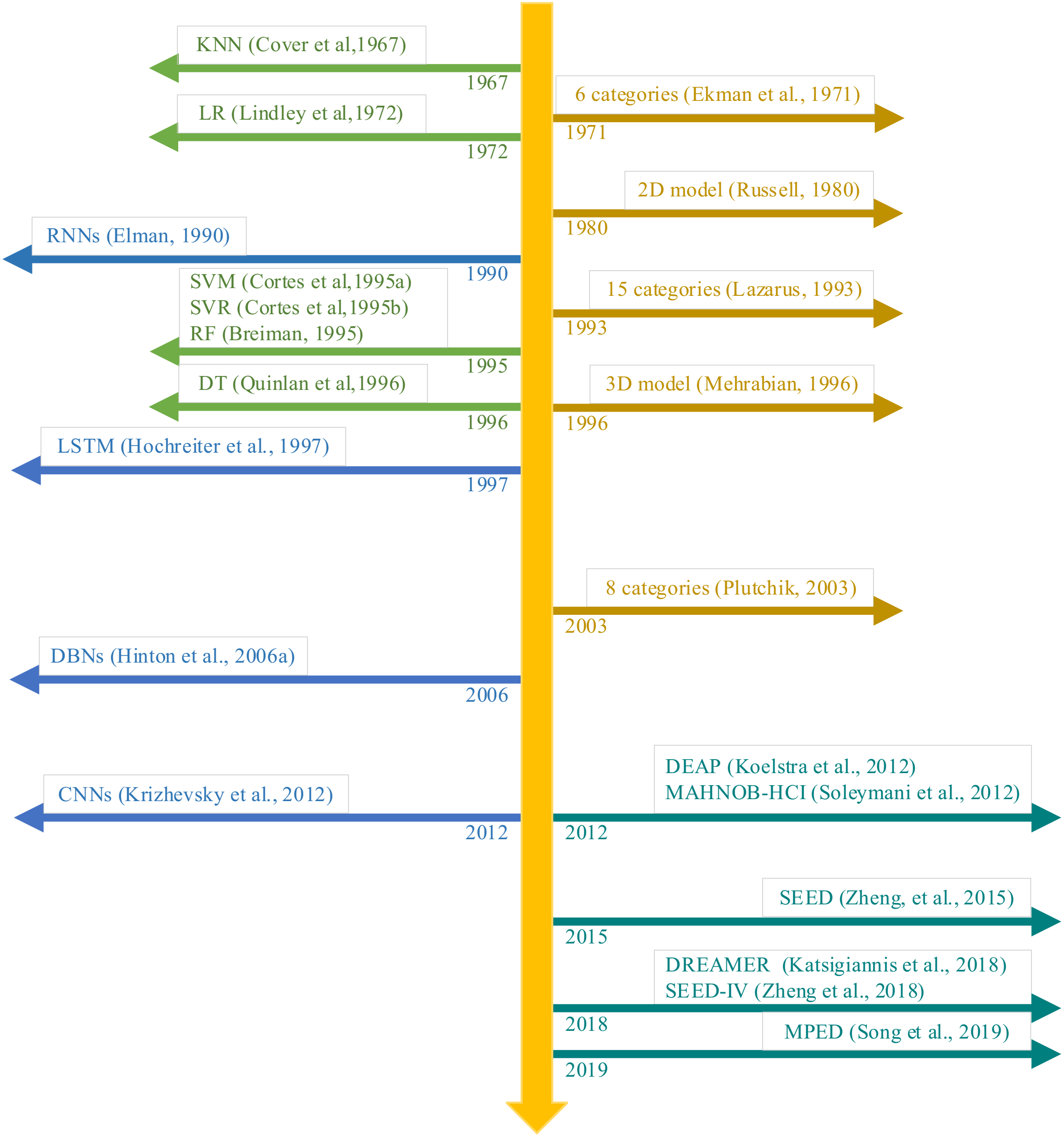

KNN (Cover et al,1967)
1967
LR (Lindley et al,1972)
1972
RNNs (Elman, 1990)
1990
SVM (Cortes et al,1995a)
SVR (Cortes et al,1995b)
RF (Breiman, 1995)
1995
DT (Quinlan et al,1996)
1996
LSTM (Hochreiter et al., 1997)
1997
DBNs (Hinton et al., 2006a)
2006
CNNs (Krizhevsky et al., 2012)
2012
6 categories (Ekman et al., 1971)
1971
2D model (Russell, 1980)
1980
15 categories (Lazarus, 1993)
1993
3D model (Mehrabian, 1996)
1996
8 categories (Plutchik, 2003)
2003
DEAP (Koelstra et al., 2012)
MAHNOB-HCI (Soleymani et al., 2012)
2012
SEED (Zheng, et al., 2015)
2015
DREAMER (Katsigiannis et al., 2018)
SEED-IV (Zheng et al., 2018)
2018
MPED (Song et al., 2019)
2019

FIGURE 2
The evolution of EEG emotion recognition with deep learning algorithms, emotion categories and databases.

2. We analyze and discuss the challenges and opportunities faced to EEG emotion recognition and point out future directions in this field.

The organization of this paper is as follows. We first present the preliminaries and basic knowledge of EEG emotion recognition. We review benchmark datasets and deep learning techniques in detail. We show the recent advances of the applications of deep learning techniques for EEG emotion recognition. We give a summary of open challenge and future directions. We provide the concluding remarks.

2. Preliminaries and basic knowledge

2.1. Definition of affective computing

Professor Picard (1997) of the MIT and his team clearly defined affective computing, that is, the calculation of factors triggered by emotion, related to emotion, or able to affect and determine emotional change. In accordance with the research results in the field of emotion, emotion is a mechanism gradually formed in the process of human adaptation to social environment. When different individuals face the same environmental stimulus, they may have the same or similar emotional changes or they may have different emotional changes due to the difference in individual living environment. This psychological mechanism can play a role in seeking advantages and avoiding disadvantages. Although computers have strong logic computing ability, human beings cannot communicate more deeply when interacting with computers due to the lack of psychological mechanisms similar to human beings. Emotion theory is an effective means to solve this problem. Therefore, an effective method to realize computer intelligence is to combine logical computing with emotional computing, which is a research topic that many researchers focus on at present (Zhang S. et al., 2018).

2.2. Classification of emotional models

Many researchers cannot reach a unified emotional classification standard when conducting emotional computing research due to the high complexity and abstractness of emotion. At present, researchers usually divide emotion models into discrete model and dimensional space model.

In the discrete model, each emotion is distributed discretely, and these discrete emotions combine to form the human emotional world. In the discrete model, designers have different definitions of emotions, and they are divided into different emotional categories. American psychologist Ekman and Friesen (1971) divided human emotions into six basic emotions, namely, anger, disgust, fear, happiness, sadness, and surprise, by analyzing human facial expressions. Lazarus (Lazarus, 1993), one of the modern representatives of American stress theory, divided emotions into 15 categories, such as anger, anxiety and happiness, and each emotional state has a corresponding core related theme. Psychologist Plutchik (2003) divided emotions into eight basic categories: anger, fear, sadness, disgust, expectation, surprise, approval, and happiness. These discrete emotion classification methods are relatively simple and easy to understand, and have been widely used in many emotion recognition studies.

The dimensional space model of emotion can be divided into 2D and 3D. The 2D expression model of emotion was first proposed by psychologist Russell (1980). It uses 2D coordinate axis to describe the polarity and intensity of emotion. Polar axis is used to describe the positive and negative types of emotion, and intensity coordinate axis refers to the intensity of emotion. The 2D emotion model is consistent with people's cognition of emotion. Currently, the VA model that divides human emotions into two dimensions is widely used, which are the valency dimension and arousal dimension, as shown in Figure 3.

Considering that the 2D space representation of emotions cannot effectively distinguish some basic emotions, such as fear and anger, Mehrabian (1996) proposed a 3D space representation of emotions, and its three dimensions are pleasure, activation, and dominance, as shown in Figure 4. Centered on the origin, pleasure (P) represents the difference between positive and negative emotions; arousal (A) indicates the activation degree of human emotions; dominance (D) indicates the degree of human control over current things. At the same time, the coordinate values of the three dimensions can be used to describe specific human emotions.

2.3. Deep learning techniques

2.3.1. DBNs

DBNs proposed by Hinton et al. (2006) are a generative model aim to train the weights among its neurons and make the entire neural network generate training data in accordance with the maximum probability.

At present, DBN has been applied to many areas of life, such as voice, graphics, and other visual data classification tasks, and achieved good recognition results. Tong et al. used a DBN model to classify hyperspectral remote sensing images, improved the training process of DBN, and used the hyperspectral remote sensing image dataset Salinas to verify the proposed method (Tong et al., 2017). Compared with traditional model classification methods, the classification accuracy of DBN model can reach more than 90%. In the text classification event, Payton L et al. proposed an ME learning algorithm for DBN (Payton et al., 2016). This algorithm is specially designed to deal with limited training data. Compared with the maximum likelihood learning method, the method of maximizing the entropy of parameters in DBN has more effective generalization ability, less data distribution deviation, and robustness to over fitting. It achieves good classification effect on Newsgroup, WebKB, and other datasets. The DBN model also achieves good classification results in speech classification events. Wen et al. tried to recognize human emotions from speech signals (Wen et al., 2017) using a random DBN (RDBN) integrating method. The experimental results on the benchmark speech emotion database show that the accuracy of RDBN is higher than that of KNN and other speech emotion recognition methods. Kamada S and Ichimura T extended the learning algorithm of adaptive RBM and DBN to time series analysis by using the idea of short-term memory and used the adaptive structure learning method to search the optimal network structure of DBN in the training process. This method was applied to MovingMNIST, a benchmark dataset for video recognition, and its prediction accuracy exceeded 90% (Kamada and Ichimura, 2019).

2.3.2. CNNs

The concept of neural networks originated from the neural mathematical model first proposed in 1943 (Mcculloch and Pitts, 1943). However, the artificial neural networks confined to the shallow network architecture fell into a low tide in the late 1960s due to the constraints of early computing power, data and other practical conditions. The real rise of neural network method began with AlexNet (Krizhevsky et al., 2012) proposed by Hinton et al. (2012). This CNN model won the Image Net Large Scale Visual Recognition Challenge (Deng et al., 2009) with a huge

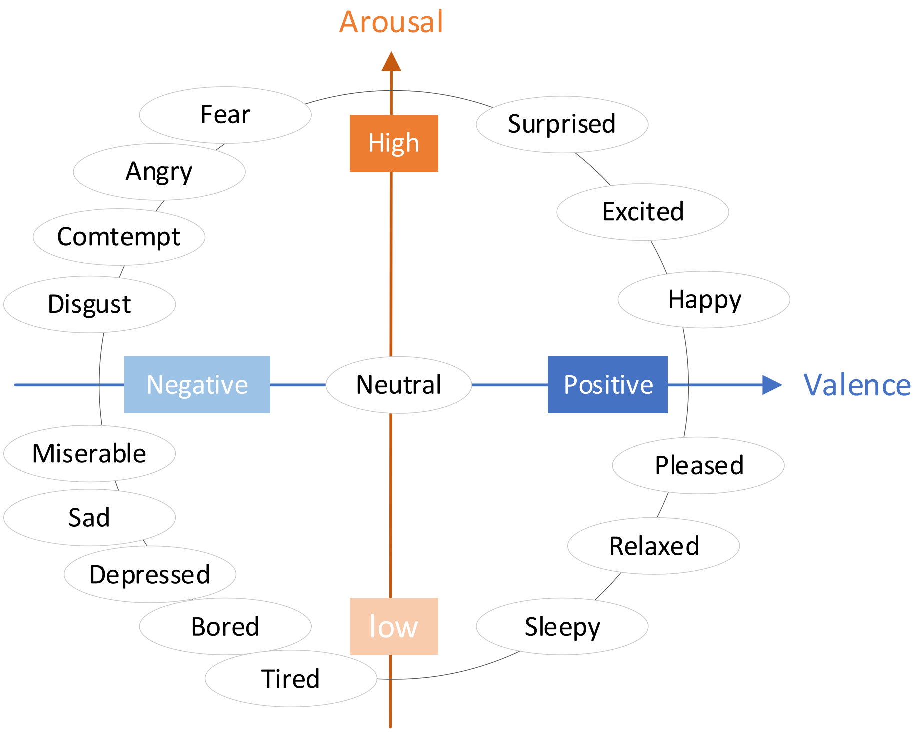


FIGURE 3
Two-dimensional model for valence–arousal.

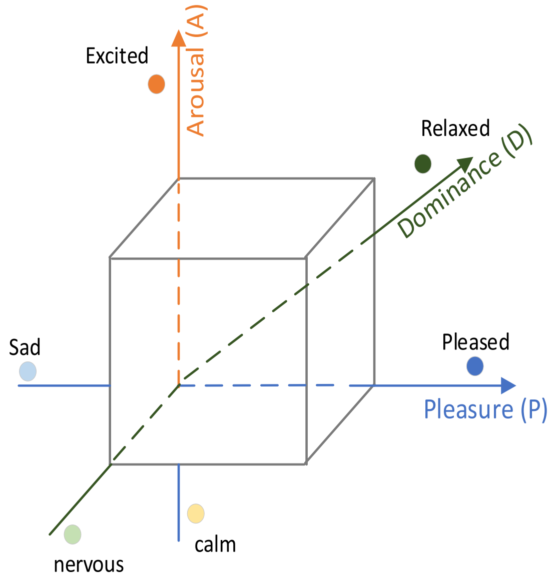


FIGURE 4
PAD 3D emotional model.

advantage of 10.9%. Therefore, deep neural networks (DNNs) have gradually attracted extensive attention from the industry and academia. In a broad sense, the DNNs can be divided into feedforward neural networks (Rumelhart et al., 1986; Lecun and Bottou, 1998) and RNNs (Williams and Zipser, 1989; Yi, 2010; Yi, 2013) in accordance with the difference in connection modes between neurons. In accordance with the differences in use scenarios and HCI methods, the DNNs can be divided into single input networks and multi-input networks, which are continuously extended to a variety of HCI scenarios and have achieved breakthrough results (Wang et al., 2018), allowing AI products to be actually implemented in practical applications.

2.3.3. RNNs

Compared with CNNs, RNNs are better in processing data with sequence characteristics and can obtain time-related information in data (Lecun et al., 2015). Researchers have widely used RNNs in natural language processing, including machine translation, text classification, and named entity recognition (Schuster and Paliwal, 1997). RNNs have achieved outstanding performance in audio-related fields and made great breakthroughs. They have been widely used in speech recognition, speech synthesis, and other fields. Considering that RNN only considers the preorder information and ignores the postorder information, a bidirectional RNN (BRNN) was proposed (Schuster and Paliwal, 1997). To solve the problems of gradient disappearance and gradient explosion in the training process of the long sequence of RNN, researchers improved the structure of RNN and built a LSTM (Hochreiter and Schmidhuber, 1997). LSTM has modified the internal structure of RNN in the current time step, making the hidden layer architecture more complex, which can have a better effect in longer sequences and is a more widely used RNN in the general sense.

The bidirectional LSTM (BILSTM) can be obtained by combining LSTM and BRNN (Graves and Schmidhuber, 2005). It replaces the original RNN neuron structure in BRNN with the neuron structure of LSTM and combines the forward LSTM and backward LSTM to form a network. BILSTM retains the advantages of BRNN and LSTM at the same time. It can retain the context information of the current time node and record the relationship between the front and back features. Therefore, BILSTM improves the generalization ability of the network model and the ability to handle long sequences, and avoids the problem of gradient explosion and gradient disappearance according to the difference of connection modes between neurons.

In recent years, gated recurrent unit (GRU) networks (Cho et al., 2014) were proposed. They discard the three LSTM gated (force gate, input gate, and output gate) networks, selects reset gate and update gate, and combines the current state of neurons and the hidden layer state, which are uniformly expressed as h_t. Compared with LSTM, the GRU model is simpler, with fewer parameters and easier convergence.

3. Benchmark datasets

The proposed EEG emotion recognition algorithms should be verified on EEG data with emotion ratings or labels. However, some researchers are limited by conditions and cannot build a special experimental environment. Most researchers are interested in verifying their algorithms and comparing with relevant studies on the recognized benchmark datasets. Hence, a variety of open-source EEG emotion databases have been developed for EEG emotion recognition. Table 1 presents a brief summary of existing speech emotion databases. In this section, we describe briefly these existing EEG emotion databases as follows.

3.1. Database for emotion analysis using physical signals

DEAP is a large multimodal physiological and emotional database jointly collected and processed by Koelstra and other research institutions of four famous universities (Queen Mary University in London, Twente University in the Netherlands, Geneva University in Switzerland, and Swiss Federal Institute of Technology; Koelstra et al., 2012). The collection scene for the DEAP database is shown in Figure 5, and it is an open-source data set for analyzing human emotional states. The DEAP database collected 32 participants for the experiment, where 16 of them were male and 16 were female. In the experiment, the EEG and peripheral physiological signals of the participants were collected, and the frontal facial expression videos of the first 22 participants were recorded. The participants read the instructions of the experiment process and wore the detection equipment before starting the data acquisition experiment. Each participant watched 40 music video clips with a duration of 1 min in the experiment. The subjective emotional experience in induction experiments was self-evaluated and rated on assessment scales that cover four emotional dimensions, namely, arousal, valence, dominance, and like. During self-assessment, the participants saw the content displayed on the screen and clicked to select the option that matched their situation at that time. The EEG and peripheral physiological signals were recorded by using a Biosemi ActiveTwo system. The EEG information was collected by using electrode caps with 32 AgCl electrodes. The EEG sampling rate was 512 Hz. The data set recorded 40 channels in total, the first 32 channels were EEG signal channels, and the last 8 channels were peripheral signal channels.

3.2. Multimodal database for affect recognition and implicit tagging

MAHNOB-HCI is a multimodal physiological emotion database collected by Soleymani et al. (2014) through a reasonable and normal experimental paradigm. The MAHNOB-HCI dataset collected EEG signals and peripheral physiological signals from 30 volunteers with different cultural and educational backgrounds using emotional stimulation videos. Among the 30 young healthy adult participants, 17 were women and 13 were men, and the age ranged from 19 to 40. Thirty participants watched 20 different emotional video clips selected from movies and video websites. These video clips can stimulate the subjects to have five emotions: disgust, amusement, fear, sadness, and joy. The duration of watching videos was 35 to 117 s. The participants evaluated the arousal and potency dimensions rated on assessment scales after watching each video clip. In the data collection experiment, six cameras were used to record the facial expressions of the subjects at a frame rate of 60 frames per second. The collection scene for the MAHNOB-HCI database is shown in Figure 6.

3.3. SJTU emotion EEG dataset

The SEED is an EEG emotion dataset released by the BCMI Research Center in Shanghai Jiaotong University (SJTU; Zheng and Lu, 2015), and the protocol used in the emotion experiment is shown in Figure 7. The SEED dataset selected 15 people (7 men and 8 women) as the subjects of the experiment and collected data of 62 EEG electrode channels from the participants. In the experiment, 15 clips of Chinese movies were selected for the subjects to watch. These videos contained three types of emotions: positive, neutral, and negative. Each genre had five clips, and each clip was about 4 min. Clips containing different emotions appeared alternately. In the experiment, the subjects

TABLE 1 Description of public datasets.

Name	Participants	Documented Signals	Stimulus	Task models/ Emotions
DEAP	32	EEG, EMG, EOG, GSR, Temperature, and Face Video	40 Video clips	VAD model
MAHNOB-HCI	27	EEG, ECG, GSR, ERG, Respiration Amplitude, Skin Temperature, Face Video, Audio Signals, and Eye Gaze	20 Video clips and Pictures	VAD model
SEED	15	EEG, Face Video, and Eye tracking	15 Video clips	Positive, Neutral, and Negative
DREAMER	23	EEG, ECG	18 Video clips	VAD model
SEED-IV	15	EEG, and EM	168 Video clips	Happiness, Sadness, Fear and Neutrality
MPED	23	EEG、ECG、RSP、and GSR	28 Video clips	Joy, Funny, Anger, Fear, Disgust, Sadness, and Neutrality

had a 5-s prompt before watching each video. The subjects conducted a 45 s self-assessment, followed by a 15 s rest. During the experiment, the subjects were asked to complete three experiments repeatedly in a week or even longer. Each subject watched the same 15 clips of video and recorded their self-evaluation emotions.

FIGURE 5
Collection scene for the Deap database (Koelstra et al., 2012). Reproduced with permission from IEEE. Licence ID: 1319273-1.

FIGURE 6
Collection scene for the MAHNOB-HCI database (Soleymani et al., 2014). Reproduced with permission from IEEE. Licence Number: 5493400200365.

3.4. Dreamer

The Dreamer database (Katsigiannis and Ramzan, 2018) is a multimodal physiological emotion database released by the University of Western Scotland in 2018. It contains 18 audio-visual clips during the emotion induction experiments and collects the EEG and electrocardiogram (ECG) signals simultaneously. The video duration is between 65 and 393 s, with an average duration of 199 s. Twenty-three subjects with an average age of 26.6 years were invited to participate in the experiment. The subjects were asked to conduct a self-assessment between 1 and 5 points in the emotional dimensions of valence, arousal, and dominance after each emotional induction experiment.

3.5. Seed-iv

SEED-IV is another version of the SEED dataset released by SJTU (Zheng et al., 2018), which has been widely used in recent related work. The protocol of SEED-IV for four emotions is shown in Figure 8. Forty-four participants (22 women, all college students) were recruited to self-evaluate their emotions during the induction experiment, and 168 film clips were selected as the material library of four emotions (happiness, sadness, fear, and neutrality). It follows the experimental paradigm adopted in SEED, 62-channel EEG of 15 selected subjects were recorded in the three tests. They chose 72 film clips with four different emotional labels (neutral, sad, fear, and happy). Each subject watched six film clips in each session, resulting in 24 trials in total.

3.6. Multi-modal physiological emotion database

The MPED contains four physiological signals of 23 subjects (10 men and 13 women) and records seven types of discrete emotion (joy, funny, anger, fear, disgust, sadness, and neutrality) when they watch 28

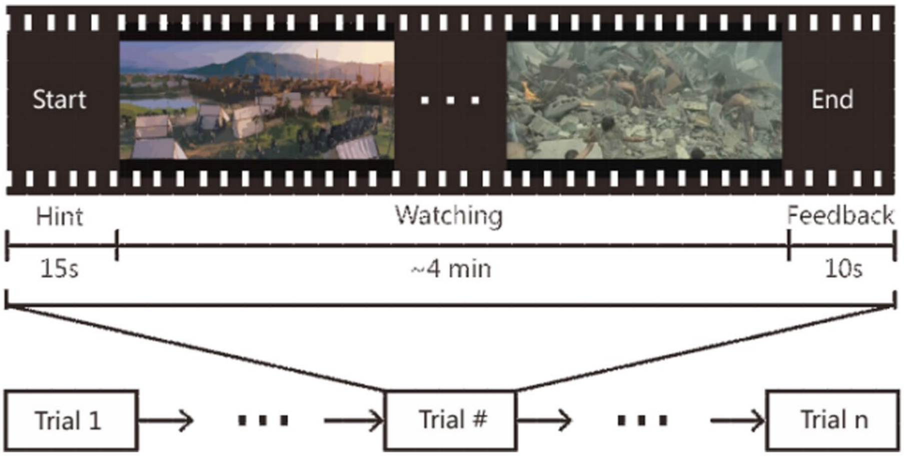


FIGURE 7
Protocol used in the emotion experiment (Zheng and Lu, 2015). Reproduced with permission from IEEE. Licence Number: 5493480479146.

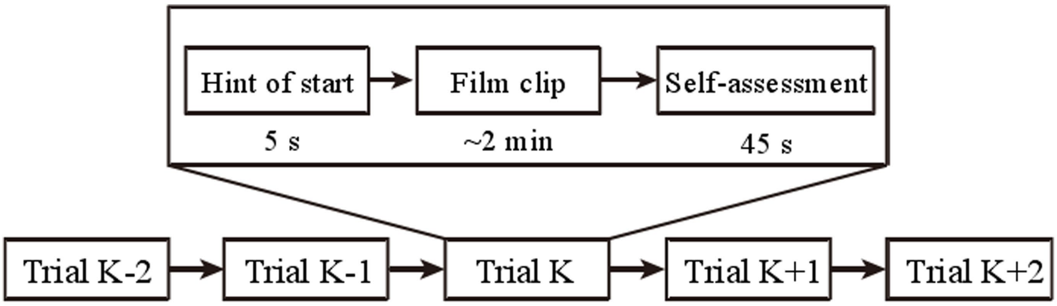


FIGURE 8
Protocol of SEED-IV for four emotions (Zheng et al., 2018). Reproduced with permission from IEEE. Licence Number: 5493600273300.

video clips (Song et al., 2019). The experiments are divided into two sessions with an interval of at least 24 h. Twenty-one EEG data are used as training data, and the remaining 7 EEG data are used as test data.

4. Review of EEG emotion recognition techniques

4.1. Shallow machine learning methods for EEG emotion recognition

The emotion recognition method of EEG signals based on machine learning is usually divided into two steps: manual feature extraction and classifier selection (Li et al., 2019). The feature extraction methods mainly include time domain analysis, frequency domain analysis, time frequency domain analysis, multivariate statistical analysis, and nonlinear dynamic analysis (Donos et al., 2018; Rasheed et al., 2020). Principal component analysis, linear discriminant analysis (LDA; Alotaiby et al., 2017), and independent component analysis (Ur et al., 2013; Acharya et al., 2018a; Maimaiti et al., 2021) are widely used unsupervised time-domain methods to summarize EEG data. Frequency domain features include spectral center, coefficient of variation, power spectral density, signal energy, spectral moment, and spectral skewness, which can provide key information about data changes (Yuan et al., 2018; Acharya et al., 2018a). The abovementioned time-domain or frequency-domain methods have limitations and cannot provide accurate frequency or time information at a specific time point. Wavelet transformation (WT) is usually used to decompose EEG signals into their frequency components to express the relationship between signal information and time. Time frequency signal processing algorithms, such as discrete wavelet transform analysis and continuous wavelet transform, are a necessary means to solve different EEG behavior, which can be described in the time and frequency domains (Martis et al., 2012; UR et al., 2013). Statistical parameters, such as mean, variance, skewness, and kurtosis, have been widely used to extract feature information from EEG signals. Variance represents the distribution of data, skewness represents the symmetry information of data, and kurtosis provides the peak information in data (Acharya et al., 2018a).

In classifier selection, previous work mainly used shallow machine learning methods, such as LDA, SVM, and KNN, to train emotion recognition models based on manual features. Although the method of "manual features+shallow classifier" has made some progress in previous emotion recognition systems, the design of manual features requires considerable professional knowledge, and the extraction of some features (such as linear features) is time consuming.

4.2. Deep learning for EEG emotion recognition

Traditional machine learning techniques extract EEG features manually, which not only have high redundancy in the extracted features, but also have poor universality. Therefore, manual feature extraction techniques can not achieve the ideal results in EEG emotion recognition. Obviously, with the increasing progress of deep learning technology (Chen B. et al., 2021), EEG emotion recognition research ground on various neural networks has gradually become a research hotspot. Different from shallow classifier, deep learning has the advantages of strong learning ability and good portability, which can automatically learn good feature representations instead of manually design. Recently, various deep learning models, such as DNN, CNN, LSTM, and RNN models, were tested on public datasets. Compared with CNN, RNN is more suitable for processing sequence-related tasks. LSTM has been proven to be capable of capturing time information in the field of emotion recognition (Bashivan et al., 2015; Ma et al., 2019). As a type of sequence data, most studies on EEG are based on RNN and LSTM models. Li et al. (2017) designed a hybrid deep learning model by combining CNN and RNN to mine inter-channel correlation. The results demonstrated the effectiveness of the proposed methods, with respect to the emotional dimensions of Valence and Arousal. Zhang T. et al. (2018) proposed a spatial–temporal recurrent neural network (STRNN) for emotion recognition, which integrate the feature learning from both spatial and temporal information of signal sources into a unified spatial–temporal dependency model, as shown in Figure 9. Experimental results on the benchmark emotion datasets of EEG and facial expression show that the proposed method is significantly better than those state-of-the-art methods. Nath et al. (2020) compared the emotion recognition effects of LSTM with KNN, SVM, DT, and RF. Among them, LSTM has the best robustness and accuracy.

EEG signals are essentially multichannel time series signals. Thus, a more effective method for emotional recognition of EEG signals is to obtain the long-term dependence of EEG signals based on RNN. Li et al. (2020) proposed a BILSTM network framework based on multimodal attention, which is used to learn the best time characteristics, and inputted the learned depth characteristics into the DNN to predict the emotional output probability of each channel. A

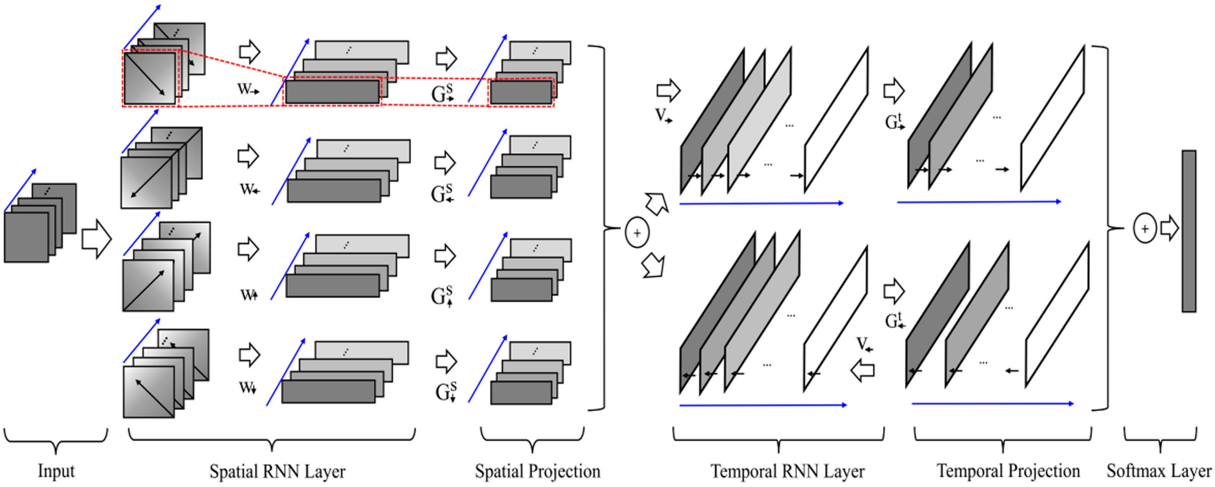


FIGURE 9
The used STRNN framework for EEG emotion recognition (Zhang T. et al., 2018). Reproduced with permission from IEEE. Licence Number: 5493591292973.

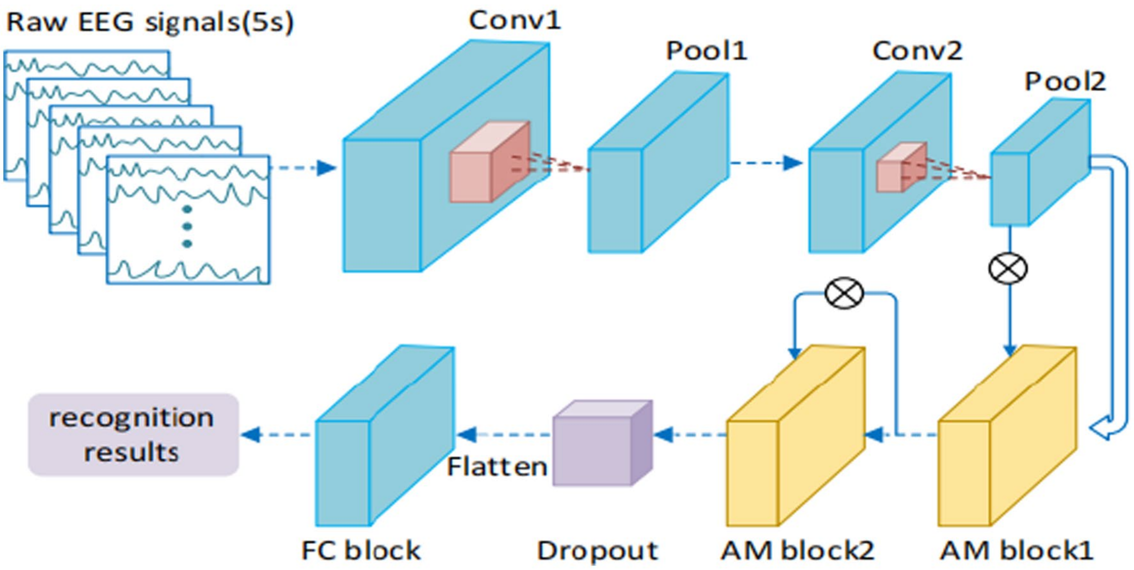


FIGURE 10
The flow of the 3DCANN algorithm (Liu et al.,2021). Reproduced with permission from IEEE. Licence Number: 5493600805113.

decision-level fusion strategy is used to predict the final emotion. The experimental results on AMIGOS dataset show that this method is superior to other advanced methods. Liu et al. (2017) proposed an emotion recognition algorithm model ground on multi-layer long short-term memory recurrent neural network (LSTM-RNN), which combines temporal attention (TA) and band attention (BA). Experiments on Mahnob-HCI database demonstrate the proposed method achieves higher accuracy and boosts the computational efficiency. Liu et al.(2021) studied an original algorithm named three-dimension convolution attention neural network (3DCANN) for EEG emotion recognition, which is composed of spatio-temporal feature extraction module and EEG channel attention weight learning module. Figure 10 presents the details of the used 3DCANN scheme. Alhagry et al. (2017) proposed an end-to-end deep learning neural network to identify emotions from original EEG signals. It uses LSTM-RNN to learn features from EEG signals and uses full connection layer for classification. Li et al. (2022) proposed a C-RNN model using CNN and RNN, and used multichannel EEG signals to identify emotions. Although the method based on RNN has great advantages in processing time series data and has made great achievements, it still has shortcomings in the face of multichannel EEG data. GRU or LSTM can connect the relationship between different channels through multichannel fusion, but this processing ignores the spatial distribution of EEG channels and cannot reflect the dynamics of the relationship between different channels.

Among various network algorithms, CNNs have a good ability to extract features of convolution kernels. They can extract information features by transferring each part of the image with multiple kernels. They have been widely used in image processing tasks. For EEG signal, they can process raw EEG data well and can be used for spectrum diagram. Considering that the use of CNN to train EEG data can reduce the effect of noise, most studies use CNN for the emotional recognition of EEG signals to reduce the complexity of training. Thodoroff et al. (2016) combined CNN and RNN to train robust features for automatic detection of seizures. Shamwell et al. (2016) explored a new CNN architecture with 4 convolutional layers and 3 fully connected layers to classify EEG signals. To reduce the over fitting of the model, Manor and Geva (2015) proposed a CNN model based on spatiotemporal regularization, which is used to classify single track EEG in RSVP (fast serial visual rendering). Sakhavi et al. (2015) proposed a parallel convolutional linear network, which is an architecture that can represent EEG data as dynamic energy input, and used CNN for image classification. Ren and Wu (2014) applied convolutional DBN to classify EEG signals. Hajinoroozi et al. (2017) used covariance learning to train EEG data for driver fatigue prediction. Jiao et al. (2018) proposed an improved CNN method for mental workload classification tasks. Gao et al. (2020) proposed a gradient particle swarm optimization (GPSO) model to achieve the automatic optimization of the CNN model. The experimental results show that the proposed method based on the GPSO-optimized CNN model achieve a prominent classification accuracy. Figure 11 presents the details of the used GPSO scheme.

CNN can use EEG to identify many human diseases. Antoniades et al. (2016) used deep learning to automatically generate features of EEG data in time domain to diagnose epilepsy. Page et al. (2016) conducted end-to-end learning through the maximum pool convolution neural network (MPCNN) and proved that transfer learning can be used to teach the generalized characteristics of MPCNN raw EEG data. Acharya et al. (2018b) proposed a five-layer deep CNN for detecting normal, pre seizure, and seizure categories.

The summary of recent state-of-the-art methods related to EEG-based emotion recognition system using machine learning and deep learning approaches is given in Table 2.

5. Open challenges

To date, although a number of literature related to EEG emotion recognition using deep learning technology is reported, showing its certain advance, a few challenges still exist in this area. In the following, we discuss these challenges and opportunities, and point out potential research directions in the future.

5.1. Research on the basic theory of affective computing

At present, the theoretical basis of emotion recognition mainly includes discrete model and continuous model, as shown in Figure 3. Although they are related to each other, they have not formed a unified theoretical framework. The relationship between explicit information (such as happy, sad, and other emotional categories) and implicit information (such as the signal characteristics of different frequency bands of EEG signals corresponding to happy, sad, and other emotional categories) in emotional computing is worthy of further study. Digging out the relationship between them is extremely important for understanding the different emotional states represented by EEG signals.

5.2. EEG emotion recognition data sets

For EEG-based emotion recognition, most publicly available datasets for affective computing use images, videos, audio, and other external methods to induce emotional changes. These emotional changes are passive, which are different from the emotional changes that individuals actively produce in real scenes and may lead to differences

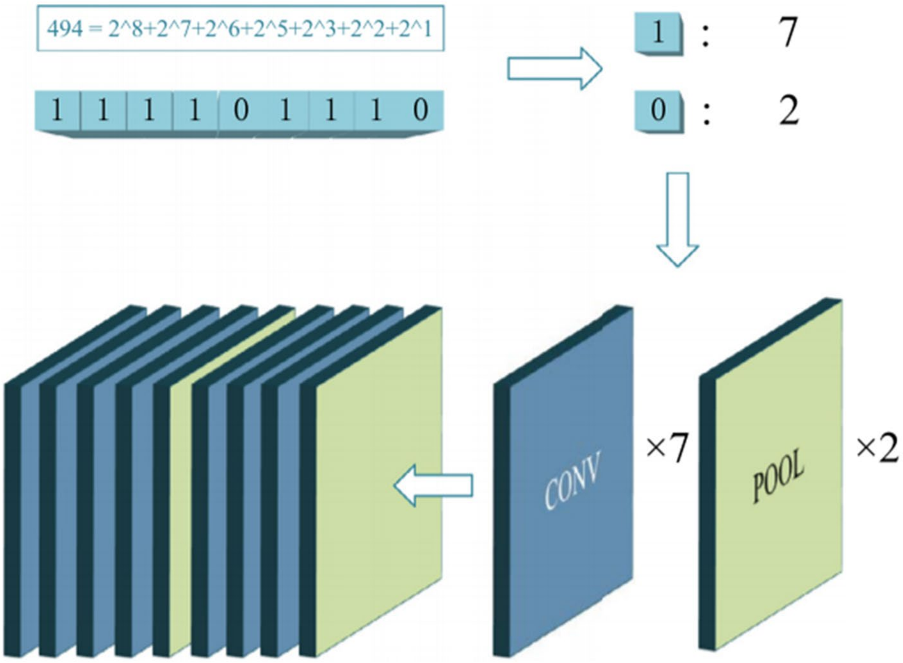


FIGURE 11
The schematic diagram of the GPSO algorithm (Gao et al., 2020). Reproduced with permission from Elsevier. Licence Number: 5493610089714.

TABLE 2 Summary of EEG emotion recognition papers using Deep learning methods from 2017 to 2022.

Year	References	Stimulus	Classification methods	Emotion	Acc.(%)
2017	Alhagry et al. (2017)	DEAP	LSTM-RNN	Valence, Arousal, and Liking	Arousal: 85.65, Valence: 85.45, Liking: 87.99
2017	Li et al. (2017)	DEAP	CNN + RNN	Arousal and Valence	Arousal: 72.06, Valence: 74.12
2017	Liu et al. (2017)	Mahnob-HCI	LSTM-RNN	Valence, Arousal, and F1-score	Arousal: 73.1, Valence: 74.5; F1-score: Arousal: 72.3, Valence: 73.0
2018	Zhang T. et al. (2018)	SEED and CK+	STRNN	Positive, Negative, and Neutral	SEED: Overall Accuracy: 89.5 CK+: Overall Accuracy: 95.4
2018	Jiao et al. (2018)	The Sternberg memory task	Deep CNN	Types of mental load	Fused CNNs 1: 91.32 Fused CNNs 2: 92.37
2018	Song et al. (2018)	SEED and DREAMER	DGCNN	Positive, neutral and negative; Arousal, Valence, and Dominance	SEED: 90.40, DREAMER: Arousal: 84.54 Valence: 86.23 Dominance: 85.02
2018	Salama et al. (2018)	DEAP	3D-CNN	Arousal and Valence	Arousal: 88.49 Valence: 87.44
2019	Chao et al. (2019)	DEAP	CapsNet	Valence, Arousal and Dominance	Valence: 66.73 Arousal: 68.28 Dominance: 67.25
2019	Gonzalez et al. (2019)	DEAP, IAPS and DREAMER	CNN	Valence and Arousal	Single subject: Valence: 70.26 Arousal: 72.42
2019	Garg et al. (2019)	DEAP	Merged LSTM	Arousal, Valence, Liking and Dominance	Arousal: 83.85, Valence: 84.89 Liking: 80.72 Dominance: 84.37
2019	Wang et al. (2019)	SEED	DNNs	Positive, Negative, and Neutral	Overall Accuracy: 93.28
2019	Chen et al. (2019b)	DEAP	Bagging Tree (BT), SVM, LDA, Bayesian LDA, Deep CNN	Valence and Arousal	Valence: 99.97 (using CVCNN), Arousal: 99.58 (using GSLTCNN)
2019	Ma et al. (2019)	DEAP	multimodal residual LSTM	Arousal and Valence	Valence: 92.30 Arousal: 92.87
2019	Pandey and Seeja (2019)	DEAP	DNN	Arousal and Valence	Arousal: 61.25 Valance: 62.50
2020	Nakisa et al. (2020)	Audio-video Clips	ConvNet long short-term memory (LSTM; early and late fusion)	Low Arousal Positive, High Arousal Positive, Low Arousal Negative, High Arousal Negative	Overall accuracy: Early fusion:71.61 Late fusion: 70.17
2020	Nath et al. (2020)	DEAP	LSTM	Arousal and Valence	Valence: 94.69 Arousal: 93.13
2020	Gao et al. (2020)	Film clips	GPSO-optimized CNN	Fear, happiness, and sadness	Overall accuracy: 92.44 ± 3.60
2020	Joshi and Ghongade (2020)	Own dataset	BiLSTM	Positive, neutral and negative	Overall Accuracy:72.83
2020	Wei et al. (2020)	SEED	SRU	Positive, neutral and negative	Overall Accuracy:80.02
2020	Sharma et al. (2020)	DEAP and SEED	LSTM	Arousal and Valence Positive, neutral and negative	DEAP: 4 classes: 82.01 Arousal: 85.21 Valance: 84.16 SEED: 90.81
2020	Alhalaseh and Alasasfeh (2020)	DEAP	CNN, k-NN, NB, DT	Valence and Arousal	Overall accuracy: 95.20 (using CNN)
2020	Cui et al. (2020)	DEAP and DREAMR	Regional- Asymmetric Convolutional Neural Network (RACNN)	Valence and Arousal	Overall accuracy: 96.65 (Valence), 97.11 (Arousal)

(Continued)

TABLE 2 (Continued)

Year	References	Stimulus	Classification methods	Emotion	Acc.(%)
2020	Hassouneh et al. (2020)	Own dataset	LSTM	Happy, fear, anger, sad, Surprise and disgust	Overall accuracy: 7.25
2020	Liu et al. (2020)	SEED	DECNN	Positive and negative	Overall accuracy: 97.56
2020	Li et al. (2020)	AMIGOS	Bidirectional LSTM-RNNs	Valence and Arousal	Arousal: F1-Score: 61.3, ACC: 73.5; Valence: F1-Score: 58.3, ACC: 67.8
2021	Topic and Russo (2021)	DEAP,DREAMER, SEED and AMIGOS.	CNN + SVM	Arousal and Valence; Positive and negative	DEAP: Arousal:77.7 andValence: 76.6 DREAMER: Arousal: 90.4 andValence: 88.2 AMIGOS: Arousal: 90.5 andValence: 78.4 SEED: 88.5
2021	Liu and Fu (2021)	DEAP	multi-channel feature fusion	SROCC and PLCC	SROCC: 78.9, PLCC: 84.3
2021	Sakalle et al. (2021)	Own dataset, DEAP and SEED	LSTM	Disgust, sadness, surprise, and anger Positive, negative, and neutral	DEAP: 91.38 SEED: 89.34 Own dataset: 4 class: 94.12 3 class: 92.66
2021	Huang et al. (2021)	DEAP	BiDCNN	Arousal and Valence	Subject-dependent Arousal:94.72 Valence: 94.38 Subject-independent Arousal: 63.94 Valence: 68.14
2022	Chowdary et al. (2022)	Own dataset	RNN, LSTM, and GRU	positive, negative, and neutral	average accuracy: RNN: 95, LSTM: 97, GRU:96
2022	Algarni et al. (2022)	DEAP	Bi-LSTM	arousal, valence and liking	average accuracy: valence: 99.45, arousal: 96.87, liking: 99.68
2022	Tuncer et al. (2022)	DREAMER	LEDPatNet19	arousal, dominance, and valance	valence: 94.58, arousal: 92.86, arousal: 94.44
2022	Li et al. (2022)	DEAP and SEED	ensemble learning	arousal and valence	DEAP average accuracy: Arousal: 65.70, valence: 64.22 SEED average accuracy: 84.44
2022	He et al. (2022)	DREAMER and DEAP	adversarial discriminative-temporal convolutional networks (AD-TCNs)	arousal and valence	DEAP average accuracy: Arousal: 64.33, valence: 63.25 DREAMER average accuracy: Arousal: 66.56, valence: 63.69
2022	Wang et al. (2022)	DEAP	2D CNN	arousal and valence	Average accuracy: Arousal: 99.99, valence: 99.98

in their EEG signals. Therefore, how to solve the difference between the external-induced emotional change and the internal active emotional change is a subject worthy of study.

Different individuals may not induce the same emotion for the same emotion-inducing video due to the differences in the physiology and psychology between different subjects. Although the same emotion is generated, the EEG signals will have some differences due to the physiological differences between individuals. To effectively solve the problem of individual differences, we can build a personalized emotional computing model from the perspective of individuals. However, building an emotion recognition model with better generalization ability is a relatively more economical solution because the collection and annotation of physiological signals will bring about a large cost. An effective method to improve the generalization ability of affective computing models is transfer learning (Pan et al., 2011). Therefore, how to combine the agent independent classifier model with the transfer learning technology may be a point worth being considered in the future.

The privacy protection of users' personal information is an important ethical and moral issue in the Internet era. The EEG and other physiological signals collected in emotional computing belong to users' private information, so privacy protection should be paid attention. At present, research in this area has only started (Cock et al., 2017; Agarwal et al., 2019).

5.3. EEG signal preprocessing and feature extraction

In the EEG signal acquisition experiment, many equipment are needed, and the noise acquisition should be minimized. However, EEG signal acquisition is more complex, and the acquisition results are often vulnerable to external factors. Therefore, collecting EEG signals with high efficiency and quality is an important part of affective computing. Effective preprocessing can remove the noise in the original EEG signal, improve the signal quality, and help feature extraction, which is another important part for affective computing.

The common features of EEG signal include power spectral density, differential entropy, asymmetric difference of differential entropy, asymmetric quotient of differential entropy, discrete wavelet analysis, empirical mode decomposition, empirical mode decomposition sample entropy (EMD_SampEn), and statistical features (mean, variance, etc.). How to extract appropriate features or fuse different features will have an important effect on affective computing models.

6. Conclusion

Multiple recent studies using deep learning have been conducted for EEG emotion recognition associated with promising performance due to the strong feature learning and classing ability of deep learning. This paper attempts to provide a comprehensive survey of existing EEG emotion recognition methods. The common open data sets of EEG-based affective computing are introduced. The deep learning techniques are summarized with specific focus on the common methods of emotional calculation of EEG signals, related algorithms. The challenges faced by emotional computing based on EEG signals and the problems to be solved in the future are analyzed and summarized.

Author contributions

XW and YR contributed to the writing and drafted this article. ZL and JH contributed to the collection and analysis of existing literature. YH contributed to the conception and design of this work and revised this article. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by Hunan Provincial Natural Science Foundation of China (under Grant Nos. 2022JJ50148 and 2021JJ50082), Hunan Provincial Education Department Science Research Fund of China (under Grant Nos. 22A0625 and 21B0800), and National Innovation and Entrepreneurship Training for University of PRC (under Grant Nos. 202211528036 and 202211528023).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Acharya, U. R., Hagiwara, Y., and Adeli, H. (2018a). Automated seizure prediction. *Epilepsy Behav.* 88, 251–261. doi: 10.1016/j.yebeh.2018.09.030

Acharya, U. R., Adeli, et al. (2018b). Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. *Comput. Biol. Med.* 100, 270–278. doi: 10.1016/j.compbiomed.2017.09.017

Alhagry, S., Aly, A., and Reda, A. (2017). Emotion recognition based on EEG using LSTM recurrent neural network. *Int. J. Adv. Comput. Sci. Appl.* 8, 355–358. doi: 10.14569/IJACSA.2017.081046

Alhalaseh, R., and Alasasfeh, S. (2020). Machine-learning-based emotion recognition system using EEG signals. *Computers* 9:95. doi: 10.3390/computers9040095

Algarni, M., Saeed, F., Al-Hadhrami, T., et al. (2022). Deep learning-based approach for emotion recognition using electroencephalography (EEG) signals using bi-directional long short-term memory (bi-LSTM). *Sensors* 22:2976. doi: 10.3390/s22082976

Alreshidi, A., and Ullah, M. (2020). Facial emotion recognition using hybrid features. *Informatics* 7:6. doi: 10.3390/informatics7010006

Alotaiby, T. N., Alshebeili, S. A., Alotaibi, F. M., and Alrshoud, S. R. (2017). Epileptic seizure prediction using CSP and LDA for scalp EEG signals. *Comput. Intell. Neurosci.* 2017, 1–11. doi: 10.1155/2017/1240323

Antoniades, A., Spyrou, L., Took, C. C., and Sanei, Saeid (2016). "Deep learning for epileptic intracranial EEG data". in *2016 IEEE 26th international workshop on machine learning for signal processing (MLSP). IEEE.*

Agarwal, A., Dowsley, R., McKinney, N. D., Wu, D., Lin, C. T., de Cock, M., et al. (2019). Protecting privacy of users in brain-computer interface applications. *IEEE Trans. Neural Syst. Rehabil. Eng.* 27, 1546–1555. doi: 10.1109/TNSRE.2019.2926965

Bashivan, P., Rish, I., Yeasin, M., and Codella, N. (2015). Learning representations from EEG with deep recurrent-convolutional neural networks. *Computer Science*. doi: 10.48550/arXiv.1511.06448

Bos, D. O. (2006). EEG-based emotion recognition. The influence of visual and auditory stimuli. *Computer Science* 56, 1–17.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* 37, 373–384. doi: 10.1080/00401706.1995.10484371

Cannon, W. B. (1927). The James-Lange theory of emotions: a critical examination and an alternative theory. *Am. J. Psychol.* 39, 106–124. doi: 10.2307/1415404

Candra, H., Yuwono, M., Chai, R., Handojoseno, Ardi, Elamvazuthi, Irraivan, Nguyen, Hung T., and Su, Steven (2015) "Investigation of window size in classification of EEG-emotion signal with wavelet entropy and support vector machine". in *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC).* pp. 7250–7253. IEEE.

Cai, H., Qu, Z., Li, Z., Zhang, Y., Hu, X., and Hu, B. (2020). Feature-level fusion approaches based on multimodal EEG data for depression recognition. *Inf. Fusion* 59, 127–138. doi: 10.1016/j.inffus.2020.01.008

Chakladar, D. D., and Chakraborty, S. (2018). EEG based emotion classification using "correlation based subset selection". *Biol. Inspired Cogn. Archit.* 24, 98–106. doi: 10.1016/j.bica.2018.04.012

Chang, C. Y., Zheng, J. Y., and Wang, C. J. (2010). "Based on support vector regression for emotion recognition using physiological signals". in *The 2010 international joint conference on neural networks (IJCNN).* pp. 1–7. IEEE.

Chen, X., Xu, L., Cao, M., Zhang, T., Shang, Z., and Zhang, L. (2021). Design and implementation of human-computer interaction systems based on transfer support vector machine and EEG signal for depression patients' emotion recognition. *J. Med. Imaging Health Infor.* 11, 948–954. doi: 10.1166/jmihi.2021.3340

Chen, J., Hu, B., Moore, P., Zhang, X., and Ma, X. (2015). Electroencephalogram-based emotion assessment system using ontology and data mining techniques. *Appl. Soft Comput.* 30, 663–674. doi: 10.1016/j.asoc.2015.01.007

Chen, X., Zeng, W., Shi, Y., Deng, J., and Ma, Y. (2019a). Intrinsic prior knowledge driven CICA FMRI data analysis for emotion recognition classification. *IEEE Access* 7, 59944–59950. doi: 10.1109/ACCESS.2019.2915291

Chen, J. X., Zhang, P. W., Mao, Z. J., Huang, Y. F., Jiang, D. M., and Zhang, Y. N. (2019b). Accurate EEG-based emotion recognition on combined features using deep convolutional neural networks. *IEEE Access* 7, 44317–44328. doi: 10.1109/ACCESS.2019.2908285

Chen, B., Liu, Y., Zhang, Z., Li, Y., Zhang, Z., Lu, G., et al. (2021). Deep active context estimation for automated COVID-19 diagnosis. *ACM Trans. Multimed. Comput. Appl.* 17, 1–22. doi: 10.1145/3457124

Chao, H., Dong, L., Liu, Y., and Lu, B. (2019). Emotion recognition from multiband eeg signals using capsnet. *Sensors* 19:2212. doi: 10.3390/s19092212

Cho, K., Merrienboer, B. V., Gulcehre, C., Bahdanau, Dzmitry, Bougares, Fethi, Schwenk, Holger, et al. (2014). "Learning phrase representations using RNN encoder-decoder for statistical machine translation". in *Conference on empirical methods in natural language processing (EMNLP 2014).*

Chowdary, M. K., Anitha, J., and Hemanth, D. J. (2022). Emotion recognition from EEG signals using recurrent neural networks. *Electronics* 11:2387. doi: 10.3390/electronics11152387

Cui, H., Liu, A., Zhang, X., Chen, X., Wang, K., and Chen, X. (2020). EEG-based emotion recognition using an end-to-end regional-asymmetric convolutional neural network. *Knowl.-Based Syst.* 205:106243. doi: 10.1016/j.knosys.2020.106243

Cock, M., Dowsley, R., McKinney, N., Nascimento, A. C. A., and Wu, D. D. (2017), "Privacy preserving machine learning with EEG data". In: Proceedings of the 34th international conference on machine learning. Sydney, Australia.

Cortes, C., and Vapnik, V. (1995). Support vector machine. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

Cover, T., and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* 13, 21–27. doi: 10.1109/TIT.1967.1053964

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., and Fei-Fei, L. (2009). "Imagenet: a large-scale hierarchical image database". in *2009 IEEE conference on computer vision and pattern recognition (pp. 248–255).* IEEE.

Donos, C., Maliia, M. D., Dümpelmann, M., and Schulze-Bonhage, A. (2018). Seizure onset predicts its type. *Epilepsia* 59, 650–660. doi: 10.1111/epi.13997

Donos, C., Dümpelmann, M., and Schulze-Bonhage, A. (2015). Early seizure detection algorithm based on intracranial EEG and random forest classification. *Int. J. Neural Syst.* 25:1550023. doi: 10.1142/S0129065715500239

Elman, J. L. (1990). Finding structure in time. *Cogn. Sci.* 14, 179–211. doi: 10.1016/0364-0213(90)90002-E

Ekman, P., and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *J. Pers. Soc. Psychol.* 17, 124–129. doi: 10.1037/h0030377

Fan, J., Wade, J. W., Key, A. P., Warren, Z. E., and Sarkar, N. (2017). EEG-based affect and workload recognition in a virtual driving environment for ASD intervention. *IEEE Trans. Biomed. Eng.* 65, 43–51. doi: 10.1109/TBME.2017.2693157

Garg, A., Kapoor, A., Bedi, A. K., and Sunkaria, Ramesh K.2019 International Conference on Data Science and Engineering (ICDSE) (2019). "Merged LSTM model for emotion classification using EEG signals". in *2019 international conference on data science and engineering (ICDSE).* pp. 139–143. IEEE.

Gao, Z., Li, Y., Yang, Y., Wang, X., Dong, N., and Chiang, H. D. (2020). A GPSO-optimized convolutional neural networks for EEG-based emotion recognition. *Neurocomputing* 380, 225–235. doi: 10.1016/j.neucom.2019.10.096

Guo, J., Fang, F., Wang, W., and Ren, Fuji (2018). "EEG emotion recognition based on granger causality and capsnet neural network". in *2018 5th IEEE international conference on cloud computing and intelligence systems (CCIS).* pp. 47–52. IEEE.

Güntekin, B., Hanoğlu, L., Aktürk, T., Fide, E., Emek-Savaş, D. D., Ruşen, E., et al. (2019). Impairment in recognition of emotional facial expressions in Alzheimer's disease is represented by EEG theta and alpha responses. *Psychophysiology* 56:e13434. doi: 10.1111/psyp.13434

Gonzalez, H. A., Yoo, J., and Elfadel, I. M. (2019). "EEG-based emotion detection using unsupervised transfer learning". in *2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC).* pp. 694–697. IEEE.

Graves, A., and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* 18, 602–610. doi: 10.1016/j.neunet.2005.06.042

Hajinoroozi, M., Zhang, J. M., and Huang, Y. (2017). "Driver's fatigue prediction by deep covariance learning from EEG". in *2017 IEEE international conference on systems, man, and cybernetics (SMC).* pp. 240–245. IEEE.

Hassouneh, A., Mutawa, A. M., and Murugappan, M. (2020). Development of a real-time emotion recognition system using facial expressions and EEG based on machine learning and deep neural network methods. *Inform. Med. Unlocked* 20:100372. doi: 10.1016/j.imu.2020.100372

He, Z., Zhong, Y., and Pan, J. (2022). An adversarial discriminative temporal convolutional network for EEG-based cross-domain emotion recognition. *Comput. Biol. Med.* 141:105048. doi: 10.1016/j.compbiomed.2021.105048

Hinton, G. E., Osindero, S., and Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput.* 18, 1527–1554. doi: 10.1162/NECO.2006.18.7.1527

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Huang, D., Chen, S., Liu, C., Zheng, L., Tian, Z., and Jiang, D. (2021). Differences first in asymmetric brain: a bi-hemisphere discrepancy convolutional neural network for EEG emotion recognition. *Neurocomputing* 448, 140–151. doi: 10.1016/j.neucom.2021.03.105

Ji, S., Niu, X., Sun, M., Shen, T., Xie, S., Zhang, H., and Liu, H. (2022). "Emotion recognition of autistic children based on EEG signals". in *International conference on computer engineering and networks*. pp. 698–706. Springer, Singapore.

Jiao, Z., Gao, X., Wang, Y., Li, J., and Xu, H. (2018). Deep convolutional neural networks for mental load classification based on EEG data. *Pattern Recogn.* 76, 582–595. doi: 10.1016/j.patcog.2017.12.002

Jie, X., Cao, R., and Li, L. (2014). Emotion recognition based on the sample entropy of EEG. *Biomed. Mater. Eng.* 24, 1185–1192. doi: 10.3233/BME-130919

Joshi, V. M., and Ghongade, R. B. (2020). IDEA: intellect database for emotion analysis using EEG signal. *J. King Saud Univ. - Comput. Inf. Sci.* 34, 4433–4447. doi: 10.1016/j.jksuci.2020.10.007

Katsigiannis, S., and Ramzan, N. (2018). DREAMER: a database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE J. Biomed. Health Inform.* 22, 98–107. doi: 10.1109/JBHI.2017.2688239

Kaundanya, V. L., Patil, A., and Panat, A. (2015). Performance of k-NN classifier for emotion detection using EEG signals. in *2015 international conference on communications and signal processing (ICCSP).* pp. 1160–1164. IEEE.

Kamada, S.，Ichimura, T. (2019). A Video Recognition Method by using Adaptive Structural Learning of Long Short Term Memory based Deep Belief Network. *In 2019 IEEE 11th International Workshop on Computational Intelligence and Applications (IWCIA). IEEE.* 2019, 21–26. doi:10.1109/IWCIA47330.2019.8955036.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Proc. Adv. Neural Inf. Process. Syst.* 60, 84–90. doi: 10.1145/3065386

Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., and Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: a review. *IEEE Access* 7, 117327–117345. doi: 10.1109/ACCESS.2019.2936124

Koelstra, S., Muhl, C., Soleymani, M., Jong-Seok Lee, , Yazdani, A., Ebrahimi, T., et al. (2012). DEAP: a database for emotion analysis; using physiological signals. *IEEE Trans. Affect. Comput.* 3, 18–31. doi: 10.1109/T-AFFC.2011.15

Kuncheva, L. I., Christy, T., Pierce, I., and Mansoor, S.P. (2011) Multi-modal biometric emotion recognition using classifier ensembles. in *International conference on industrial, engineering and other applications of applied intelligent systems*. pp. 317–326. Springer, Berlin, Heidelberg.

Lazarus, R. S. (1993). From psychological stress to the emotions: a history of changing outlooks. *Annu. Rev. Psychol.* 44, 1–22. doi: 10.1146/annurev.ps.44.020193.000245

Lee, J. K., Kang, H. W., and Kang, H. B. (2015). Smartphone addiction detection based emotion detection result using random Forest. *J. IKEEE* 19, 237–243. doi: 10.7471/ikeee.2015.19.2.237

Lehmann, C., Koenig, T., Jelic, V., Prichep, L., John, R. E., Wahlund, L. O., et al. (2007). Application and comparison of classification algorithms for recognition of Alzheimer's disease in electrical brain activity (EEG). *J. Neurosci. Methods* 161, 342–350. doi: 10.1016/j.jneumeth.2006.10.023

Lecun, Y., and Bottou, L. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Li, C., Bao, Z., Li, L., and Zhao, Z. (2020). Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition. *Inf. Process. Manag.* 57:102185. doi: 10.1016/j.ipm.2019.102185

Li, X., Song, D., Zhang, P., Yu, Guangliang, Hou, Yuexian, and Hu, Bin (2017). Emotion recognition from multi-channel EEG data through convolutional recurrent neural network. in *2016 IEEE international conference on bioinformatics and biomedicine (BIBM).* pp. 352–359. IEEE.

Li, Y., Zheng, W., Wang, L., Zong, Y., and Cui, Z. (2019). From regional to global brain: a novel hierarchical spatial-temporal neural network model for EEG emotion recognition. *IEEE Trans. Affect. Comput.* 13, 568–578. doi: 10.1109/TAFFC.2019.2922912

Li, R., Ren, C., Zhang, X., and Hu, B. (2022). A novel ensemble learning method using multiple objective particle swarm optimization for subject-independent EEG-based emotion recognition. *Comput. Biol. Med.* 140:105080. doi: 10.1016/j.compbiomed.2021.105080

Liu, S., Wang, X., Zhao, L., Zhao, J., Xin, Q., and Wang, S. H. (2020). Subject-independent emotion recognition of EEG signals based on dynamic empirical convolutional neural network. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 18, 1710–1721. doi: 10.1109/TNSRE.2020.3019063

Liu, Y., and Fu, G. (2021). Emotion recognition by deeply learned multi-channel textual and EEG features. *Futur. Gener. Comput. Syst.* 119, 1–6. doi: 10.1016/j.future.2021.01.010

Liu, S., Wang, X., Zhao, Z., Li, B., Hu, W., Yu, J., et al. (2021). 3DCANN: A Spatio-Temporal Convolution Attention Neural Network for EEG Emotion Recognition. *IEEE Journal of Biomedical and Health Informatics.* doi: 10.1109/JBHI.2021.3083525

Liu, Y., Sourina, O., and Nguyen, M. K. (2011). *Real-time EEG-based emotion recognition and its applications*. Springer Berlin Heidelberg.

Liu, J., Su, Y., and Liu, Y. (2017). *Multi-modal emotion recognition with temporal-band attention based on LSMT-RNN*. Springer, Cham.

Lin, Y. P., Wang, C. H., Wu, T. L., Jeng, Shyh-Kang, and Chen, Jyh-Horng2009 IEEE International Conference on Acoustics, Speech and Signal Processing (2009). "EEG-based emotion recognition in music listening: a comparison of schemes for multiclass support vector machine." in *2009 IEEE international conference on acoustics, speech and signal processing*. pp. 489–492. IEEE.

Lindley, D. V., and Smith, A. F. M. (1972). Bayes estimates for the linear model. *J. R. Stat. Soc. Ser. B Methodol.* 34, 1–18. doi: 10.1111/j.2517-6161.1972.tb00885.x

Mayor-Torres, J. M., Medina - De Villiers, S., Clarkson, T., Lerner, M. D., and Riccardi, G. (2021). Evaluation of interpretability for deep learning algorithms in EEG emotion recognition: a case study in autism. *arXiv* 2111:13208.

Maimaiti, B., Meng, H., Lv, Y., Qiu, J., Zhu, Z., Xie, Y., et al. (2021). An overview of EEG-based machine learning methods in seizure prediction and opportunities for neurologists in this field. *Neuroscience* 481, 197–218. doi: 10.1016/j.neuroscience.2021.11.017

Martis, R. J., Acharya, U. R., Tan, J. H., Petznick, A., Yanti, R., Chua, C. K., et al. (2012). Application of empirical mode decomposition (EMD) for automated detection of epilepsy using EEG signals. *Int. J. Neural Syst.* 22:1250027. doi: 10.1142/S012906571250027X

Ma, J., Tang, H., Zheng, W. L., and Lu, Bao-Liang (2019). "Emotion recognition using multimodal residual LSTM network." in *Proceedings of the 27th ACM international conference on multimedia*. pp. 176–183.

Manor, R., and Geva, A. B. (2015). Convolutional neural network for multi-category rapid serial visual presentation BCI. *Front. Comput. Neurosci.* 9:146. doi: 10.3389/fncom.2015.00146

Mcculloch, W. S., and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* 5, 115–133. doi: 10.2307/2268029

Mehdizadehfar, V., Ghassemi, F., Fallah, A., and Pouretemad, H. (2020). EEG study of facial emotion recognition in the fathers of autistic children. *Biomed. Signal Process. Control* 56:101721. doi: 10.1016/j.bspc.2019.101721

Menezes, M. L. R., Samara, A., Galway, L., Sant'Anna, A., Verikas, A., Alonso-Fernandez, F., et al. (2017). Towards emotion recognition for virtual environments: an evaluation of eeg features on benchmark dataset. *Pers. Ubiquit. Comput.* 21, 1003–1013. doi: 10.1007/s00779-017-1072-7

Mehrabian, A. (1996). Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. *Curr. Psychol.* 14, 261–292. doi: 10.1007/BF02686918

Miranda, J. A., Canabal, M. F., Portela García, M., and Lopez-Ongil, Celia (2018). "Embedded emotion recognition: autonomous multimodal affective internet of things". in *Proceedings of the cyber-physical systems workshop*. Vol. 2208, pp. 22–29.

Murugappan, M. (2011). "Human emotion classification using wavelet transform and KNN". in *2011 international conference on pattern analysis and intelligence robotics*. Vol. 1, pp. 148–153. IEEE.

Murugappan, M., Ramachandran, N., and Sazali, Y. (2010). Classification of human emotion from EEG using discrete wavelet transform. *J. Biomed. Sci. Eng.* 03, 390–396. doi: 10.4236/jbise.2010.34054

Nakisa, B., Rastgoo, M. N., Rakotonirainy, A., Maire, F., and Chandran, V. (2020). Automatic emotion recognition using temporal multimodal deep learning. *IEEE Access* 8, 225463–225461, 225474. doi: 10.1109/ACCESS.2020.3027026

Nath, D., Singh, M., Sethia, D., Kalra, Diksha, and Indu, S. (2020). "A comparative study of subject-dependent and subject-independent strategies for EEG-based emotion recognition using LSTM network." in *Proceedings of the 2020 the 4th international conference on computer and data analysis*. pp. 142–147.

Nie, D., Wang, X. W., Shi, L. C., and Lu, Bao-Liang (2011). "EEG-based emotion recognition during watching movies". in *2011 5th international IEEE/EMBS conference on neural engineering*. pp. 667–670. IEEE.

Otter, D. W., Medina, J. R., and Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. *IEEE Trans. Neural Netw. Learn. Syst.* 32, 604–624. doi: 10.1109/TNNLS.2020.2979670

Payton, L., Szu, Wei F., Syu, Siang W. (2016). Maximum entropy learning with deep belief networks. *Entropy*, 18:251–256. doi:10.3390/e18070251.

Page, A., Shea, C., and Mohsenin, T. (2016). Wearable seizure detection using convolutional neural networks with transfer learning. ISCAS. IEEE.

Pandey, P., and Seeja, K. R. (2019). Subject independent emotion recognition from EEG using VMD and deep learning. *J. King Saud Univ. - Comput. Inf. Sci.* 34, 1730–1738. doi: 10.1016/j.jksuci.2019.11.003

Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q. (2011). Domain adaptation via transfer component analysis. *IEEE Trans. Neural Netw.* 22, 199–210. doi: 10.1109/TNN.2010.2091281

Piana, S., Staglianò, A., Odone, F., and Camurri, A. (2016). Adaptive body gesture representation for automatic emotion recognition. *ACM Trans. Interact. Intell. Syst.* 6, 1–31. doi: 10.1145/2818740

Plutchik, R. (2003). *Emotions and life: Perspectives from psychology, biology, and evolution.* Washington, DC: American Psychological Association.

Purwins, H., Li, B., Virtanen, T., Schluter, J., Chang, S. Y., and Sainath, T. (2019). Deep learning for audio signal processing. *IEEE J. Sel. Top. Signal Process.* 13, 206–219. doi: 10.1109/JSTSP.2019.2908700

Picard, R. W., and Picard, R. (1997). Affective computing (Vol. 252). MIT press Cambridge."EEG-detected olfactory imagery to reveal covert consciousness in minimally conscious state". *Brain Inj.* 29, 1729–1735.

Quinlan, J. R. (1996). Learning decision tree classifiers. *ACM Comput. Surv.* 28, 71–72. doi: 10.1145/234313.234346

Quintana, D. S., Guastella, A. J., Outhred, T., Hickie, I. B., and Kemp, A. H. (2012). Heart rate variability is associated with emotion recognition: direct evidence for a relationship between the autonomic nervous system and social cognition. *Int. J. Psychophysiol.* 86, 168–172. doi: 10.1016/j.ijpsycho.2012.08.012

Rasheed, K., Qayyum, A., Qadir, J., Sivathamboo, S., Kwan, P., Kuhlmann, L., et al. (2020). Machine learning for predicting epileptic seizures using EEG signals: a review. *IEEE Rev. Biomed. Eng.* 14, 139–155. doi: 10.1109/RBME.2020.3008792

Ren, Y., and Wu, Y. (2014). Convolutional deep belief networks for feature extraction of EEG signal. *International joint conference on neural networks*. pp. 2850–2853. IEEE.

Russell, J. A. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol.* 39:1161.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by backpropagating errors. *Nature* 323, 533–536. doi: 10.1038/323533a0

Sakhavi, S., Guan, C., and Yan, S. (2015). Parallel convolutional-linear neural network for motor imagery classification. in *2015 23rd European signal processing conference (EUSIPCO)*. pp. 2736–2740. IEEE.

Salama, E. S., A.el-Khoribi, R., E.Shoman, M., and A.Wahby, M. (2018). EEG-based emotion recognition using 3D convolutional neural networks. *Int. J. Adv. Comput. Sci. Appl.* 9, 329–337. doi: 10.14569/IJACSA.2018.090843

Sakalle, A., Tomar, P., Bhardwaj, H., Acharya, D., and Bhardwaj, A. (2021). A LSTM based deep learning network for recognizing emotions using wireless brainwave driven system. *Expert Syst. Appl.* 173:114516. doi: 10.1016/j.eswa.2020.114516

Schuster, M., and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* 45, 2673–2681. doi: 10.1109/78.650093

Seo, J., Laine, T. H., Oh, G., and Sohn, K. A. (2020). EEG-based emotion classification for Alzheimer's disease patients using conventional machine learning and recurrent neural network models. *Sensors* 20:7212. doi: 10.3390/s20247212

Shamwell, J., Lee, H., Kwon, H., Marathe, A., Lawhern, V., and Nothwang, W. (2016). "Single-trial EEG RSVP classification using convolutional neural networks" in *Micro-and nanotechnology sensors, systems, and applications VIII*, vol. *9836* eds. T. George, A. K. Dutta, and M. S. Islam (Proc. of SPIE), 373–382.

Sharma, R., Pachori, R. B., and Sircar, P. (2020). Automated emotion recognition based on higher order statistics and deep learning algorithm. *Biomed. Signal Process. Control* 58:101867. doi: 10.1016/j.bspc.2020.101867

Soleymani, M. (2012). A multimodal database for affect recognition and implicit tagging. *IEEE Trans. Affect. Comput.* 3, 42–55. doi: 10.1109/T-AFFC.2011.25

Soleymani, M., Asghari-Esfeden, S., Pantic, M., and Fu, Yun2014 IEEE International Conference on Multimedia and Expo (ICME) (2014). "Continuous emotion detection using EEG signals and facial expressions". in *2014 IEEE international conference on multimedia and expo (ICME)*. pp. 1–6. IEEE.

Song, T., Zheng, W., Song, P., and Cui, Z. (2018). EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Trans. Affect. Comput.* 11, 532–541. doi: 10.1109/TAFFC.2018.2817622

Song, T., Zheng, W., Lu, C., Zong, Y., Zhang, X., and Cui, Z. (2019). MPED: a multi-modal physiological emotion database for discrete emotion recognition. *IEEE Access* 7, 12177–12191. doi: 10.1109/ACCESS.2019.2891579

Suhaimi, N. S., Mountstephens, J., and Teo, J. (2020). EEG-based emotion recognition: a state-of-the-art review of current trends and opportunities. *Comput. Intell. Neurosci.* 2020, 1–19. doi: 10.1155/2020/8875426

Stavroulia, K. E., Christofi, M., Baka, E., Michael-Grigoriou, D., Magnenat-Thalmann, N., and Lanitis, A. (2019). Assessing the emotional impact of virtual reality-based teacher training. *Int. J. Inf. Learn. Technol.* 36, 192–217. doi: 10.1108/IJILT-11-2018-0127

Thodoroff, P., Pineau, J., and Lim, A. (2016). "Learning robust features using deep learning for automatic seizure detection". in *Machine learning for healthcare conference*. pp. 178–190. PMLR.

Tuncer, T., Dogan, S., and Subasi, A. (2022). LEDPatNet19: automated emotion recognition model based on nonlinear LED pattern feature extraction function using EEG signals. *Cogn. Neurodyn.* 16, 779–790. doi: 10.1007/s11571-021-09748-0

Topic, A., and Russo, M. (2021). Emotion recognition based on EEG feature maps through deep learning network. *Eng. Sci. Technol. Int. J.* 24, 1442–1454. doi: 10.1016/j.jestch.2021.03.012

Ur, A., Yanti, R., Swapna, G., Sree, V. S., Martis, R. J., and Suri, J. S. (2013). Automated diagnosis of epileptic electroencephalogram using independent component analysis and discrete wavelet transform for different electroencephalogram durations. *Proc. Inst. Mech. Eng. H J. Eng. Med.* 227, 234–244. doi: 10.1177/0954411912467883

Valderas, M. T., Bolea, J., Laguna, P., Fellow IEEEBailón, R., and Vallverdú, M. (2019). Mutual information between heart rate variability and respiration for emotion characterization. *Physiol. Meas.* 40:084001. doi: 10.1088/1361-6579/ab310a

Wang, J., Ju, R., Chen, Y., Zhang, L., Hu, J., Wu, Y., et al. (2018). Automated retinopathy of prematurity screening using deep neural networks. *EBioMedicine* 35, 361–368. doi: 10.1016/j.ebiom.2018.08.033

Wang, Y., Zhang, L., Xia, P., Wang, P., Chen, X., du, L., et al. (2022). EEG-based emotion recognition using a 2D CNN with different kernels. *Bioengineering* 9:231. doi: 10.3390/bioengineering9060231

Wang, Y., Qiu, S., Zhao, C., Yang, W., Li, J., Ma, X., and He, H. 2019). EEG-based emotion recognition with prototype-based data representation. in *2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. pp. 684–689. IEEE.

Wei, C., Chen, L., Song, Z., Lou, X. G., and Li, D. D. (2020). EEG-based emotion recognition using simple recurrent units network and ensemble learning. *Biomed. Signal Process. Control* 58:101756. doi: 10.1016/j.bspc.2019.101756

Wen, G., Li, H., Huang, J., Li, D., and Xun, E. (2017). Random deep belief networks for recognizing emotions from speech signals. *Comput. Intell. Neurosci.* 2017, 1–9. doi: 10.1155/2017/1945630

Wu, C., Zhang, Y., Jia, J., and Zhu, W. (2017). Mobile contextual recommender system for online social media. *IEEE Trans. Mob. Comput.* 16, 3403–3416. doi: 10.1109/TMC.2017.2694830

Wu, X., Sahoo, D., and Hoi, S. (2020). Recent advances in deep learning for object detection. *Neurocomputing* 396, 39–64. doi: 10.1016/j.neucom.2020.01.085

Williams, R. J., and Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.* 1, 270–280. doi: 10.1162/neco.1989.1.2.270

Yi, Z. (2013). *Convergence analysis of recurrent neural networks*, vol. *13* Kluwer Academic Publishers.

Yi, Z. (2010). Foundations of implementing the competitive layer model by lotka–volterra recurrent neural networks. *IEEE Trans. Neural Netw.* 21, 494–507. doi: 10.1109/TNN.2009.2039758

Yuan, S., Zhou, W., and Chen, L. (2018). Epileptic seizure prediction using diffusion distance and bayesian linear discriminate analysis on intracranial EEG. *Int. J. Neural Syst.* 28:1750043. doi: 10.1142/S0129065717500435

Zhang, S., Zhang, S., Huang, T., and Gao, W. (2017). Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Trans. Multimed.* 20, 1576–1590. doi: 10.1109/TMM.2017.2766843

Zhang, S., Zhang, S., Huang, T., Gao, W., and Tian, Q. (2018). Learning affective features with a hybrid deep model for audio–visual emotion recognition. *IEEE Trans. Circuits Syst. Video Technol.* 28, 3030–3043. doi: 10.1109/TCSVT.2017.2719043

Zhang, S., Zhao, X., and Tian, Q. (2019). Spontaneous speech emotion recognition using multiscale deep convolutional LSTM. *IEEE Trans. Affect. Comput.* 13, 680–688. doi: 10.1109/TAFFC.2019.2947464

Zhang, S., Tao, X., Chuang, Y., and Zhao, X. (2021). Learning deep multimodal affective features for spontaneous speech emotion recognition. *Speech Comm.* 127, 73–81. doi: 10.1016/j.specom.2020.12.009

Zhang, T., Zheng, W., Cui, Z., Zong, Y., and Li, Y. (2018). Spatial–temporal recurrent neural network for emotion recognition. *IEEE Trans. Cybern.* 49, 839–847. doi: 10.1109/TCYB.2017.2788081

Zhang, Y., Cheng, C., and Zhang, Y. (2021). Multimodal emotion recognition using a hierarchical fusion convolutional neural network. *IEEE Access* 9, 7943–7951. doi: 10.1109/ACCESS.2021.3049516

Zheng, W. L., Liu, W., Lu, Y., Lu, B. L., and Cichocki, A. (2018). Emotionmeter: a multimodal framework for recognizing human emotions. *IEEE Trans. Cybern.* 49, 1110–1122. doi: 10.1109/TCYB.2018.2797176

Zheng, W. L., and Lu, B. L. (2015). Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Trans. Auton. Ment. Dev.* 7, 162–175. doi: 10.1109/TAMD.2015.2431497

Zhou, H., du, J., Zhang, Y., Wang, Q., Liu, Q. F., and Lee, C. H. (2021). Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* 29, 2617–2629. doi: 10.1109/TASLP.2021.3096037

TYPE Original Research
PUBLISHED 23 March 2023
DOI 10.3389/fnins.2023.1141621

OPEN ACCESS

EDITED BY
Shiqing Zhang,
Taizhou University, China

REVIEWED BY
Hailun Lian,
Southeast University, China
Yang Li,
Xidian University, China
Cheng Lu,
Southeast University, China, in collaboration with reviewer YL
Yuan Zong,
Southeast University, China

*CORRESPONDENCE
Bin Hu
✉ bh@lzu.edu.cn

SPECIALTY SECTION
This article was submitted to Perception Science, a section of the journal Frontiers in Neuroscience

RECEIVED 10 January 2023
ACCEPTED 09 March 2023
PUBLISHED 23 March 2023

CITATION
Liu Z, Yu H, Li G, Chen Q, Ding Z, Feng L, Yao Z and Hu B (2023) Ensemble learning with speaker embeddings in multiple speech task stimuli for depression detection.
Front. Neurosci. 17:1141621.
doi: 10.3389/fnins.2023.1141621

Ensemble learning with speaker embeddings in multiple speech task stimuli for depression detection

Zhenyu Liu[1], Huimin Yu[1], Gang Li[2], Qiongqiong Chen[3,4], Zhijie Ding[2], Lei Feng[5], Zhijun Yao[1] and Bin Hu[1]*

[1]Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, Lanzhou, China, [2]Tianshui Third People's Hospital, Tianshui, China, [3]Second Provincial People's Hospital of Gansu, Lanzhou, China, [4]Affiliated Hospital of Northwest Minzu University, Lanzhou, China, [5]Department of Psychiatry, Beijing Anding Hospital of Capital Medical University, Beijing, China

Introduction: As a biomarker of depression, speech signal has attracted the interest of many researchers due to its characteristics of easy collection and non-invasive. However, subjects' speech variation under different scenes and emotional stimuli, the insufficient amount of depression speech data for deep learning, and the variable length of speech frame-level features have an impact on the recognition performance.

Methods: The above problems, this study proposes a multi-task ensemble learning method based on speaker embeddings for depression classification. First, we extract the Mel Frequency Cepstral Coefficients (MFCC), the Perceptual Linear Predictive Coefficients (PLP), and the Filter Bank (FBANK) from the out-domain dataset (CN-Celeb) and train the Resnet x-vector extractor, Time delay neural network (TDNN) x-vector extractor, and i-vector extractor. Then, we extract the corresponding speaker embeddings of fixed length from the depression speech database of the Gansu Provincial Key Laboratory of Wearable Computing. Support Vector Machine (SVM) and Random Forest (RF) are used to obtain the classification results of speaker embeddings in nine speech tasks. To make full use of the information of speech tasks with different scenes and emotions, we aggregate the classification results of nine tasks into new features and then obtain the final classification results by using Multilayer Perceptron (MLP). In order to take advantage of the complementary effects of different features, Resnet x-vectors based on different acoustic features are fused in the ensemble learning method.

Results: Experimental results demonstrate that (1) MFCC-based Resnet x-vectors perform best among the nine speaker embeddings for depression detection; (2) interview speech is better than picture descriptions speech, and neutral stimulus is the best among the three emotional valences in the depression recognition task; (3) our multi-task ensemble learning method with MFCC-based Resnet x-vectors can effectively identify depressed patients; (4) in all cases, the combination of MFCC-based Resnet x-vectors and PLP-based Resnet x-vectors in our ensemble learning method achieves the best results, outperforming other literature studies using the depression speech database.

Discussion: Our multi-task ensemble learning method with MFCC-based Resnet x-vectors can fuse the depression related information of different stimuli effectively, which provides a new approach for depression detection. The limitation of this method is that speaker embeddings extractors were pre-trained on the out-domain dataset. We will consider using the augmented in-domain dataset for pre-training to improve the depression recognition performance further.

KEYWORDS

depression detection, Resnet x-vectors, speaker embeddings, ensemble learning, speech task stimuli

1. Introduction

Depression is a common and recurrent mood disorder accompanied by functional disability, significantly impacting the individual's physical and mental health and daily activities (Spijker et al., 2004). More than 300 million people worldwide suffer from depression, equivalent to 4.4% of the world's population (World Health Organization, 2017). The latest scientific brief shows a dramatic 25% increase in the global prevalence of anxiety and depression in the first year of the Coronavirus 2019 (COVID-19) pandemic (World Health Organization, 2022). At present, the diagnostic methods for depression detection mainly rely on psychiatrists and scales. The accuracy of diagnostic results is affected by subjective factors such as doctors' clinical experience and whether patients can fully describe their physiological and psychological conditions.

On the other hand, in China, only 7.1% of depression patients who seek treatment in mental health institutions receive adequate treatment (Lu J. et al., 2021). The lack of medical resources leads to many patients being unable to see a doctor in time. Therefore, exploring objective and effective new techniques to identify depression has attracted much attention. Researchers have focused on seeking objective biological markers [i.e., gut hormones (Rajkumar, 2021)], physiological markers [i.e., EEG (Cai et al., 2020)] and eye movement (Shen et al., 2021), and behavioral markers [i.e., speech (Othmani et al., 2021) and facial expressions (Guo et al., 2021)] to aid in the diagnosis of depression. Among these markers, speech signal has become an important research direction for auxiliary diagnosis of depression due to its advantages of acquisition, non-invasion, non-disturbance, low cost, and a large amount of information.

Depression patients are typically sluggish (Beck and Alford, 2009), with longer pauses (Szabadi et al., 1976; Greden and Carroll, 1980) and a lack of rhythm (Alpert et al., 2001). The research showed that the percentage of pause time, the standard deviation of fundamental frequency distribution, the standard deviation of fundamental frequency change rate, and speech speed are correlated with the clinical status of patients with depression (Nilsonne, 1987). There is a strong correlation between speed, percent pause, pitch variation, and scale score (Cannizzaro et al., 2004). Depressed people treated and improved had more significant variation in pitch cycles, fewer pauses, and faster speech (Mundt et al., 2007). Thus, depressed people and healthy people have different pronunciations.

In order to make full use of the influence of speech tasks with different scenes and different emotional stimuli on speech of depressed patients and normal subjects, we designed a multi-task ensemble learning method with speaker embeddings in our depression speech dataset containing 9 speech tasks, and proved the effectiveness of this method from the accuracy, F1-D and F1-H.

The organization of the paper is as follows. The second section briefly reviews some related studies. The two datasets used in this paper are introduced in the third section. Next, the fourth section describes the multi-task ensemble learning method using speaker embeddings for depression recognition proposed in this study. Afterward, in the fifth section, the experimental results are presented. Finally, the conclusions and future works are summarized in the sixth section.

2. Related works

At present, there have been many approaches for depression recognition based on speech processing. Searching for effective acoustic features has always been an important research direction. Manual features such as spectral, source, prosodic, and formant features are commonly employed when analyzing depression and suicidality (Cummins et al., 2015). Moreover, these features are also regarded as inputs to deep neural networks (Lang and Cui, 2018; Lu X. et al., 2021). Studies have shown that the advanced features generated by MFCC feeding into the Short Long-Term Memory (LSTM) can preserve information related to depression (Rejaibi et al., 2022). PLP, and MFCC, called the low-level descriptors, are used to train the multiple classifier systems (Long et al., 2017). The input of the network model is a 3D feature made up of FBANK, the first-order and second-order differences to use the information in speech signals entirely (Wang et al., 2021). The findings of the aforementioned study illustrate that MFCC, PLP, and FBANK as front-end features can refine enough speech details.

Speaker embeddings such as i-vectors, d-vectors, and x-vectors have shown their superiority in speaker recognition (Variani et al., 2014; Wang et al., 2017), and depression detection (Egas-López et al., 2022). Scholars have found that speaker embeddings cannot only solve the variable length problem of frame-level features but also encode the speaker identity and the speech content to a large extent (Wang et al., 2017). In addition, speaker embeddings we extracted are based on the pre-trained speaker recognition model, which can be used for depression recognition tasks. The i-vectors, the low-dimension compact representations, were first proposed for speaker verification (Dehak et al., 2010). Afterward, the i-vector framework was widely applied in speaker recognition (Kanagasundaram et al., 2012), emotion recognition (Vekkot et al., 2019), Alzheimer's disease (AD) detection (Egas López et al., 2019), Parkinson's disease (PD) detection (Garcia et al., 2017), and depression detection (Cummins et al., 2014; Rani, 2017; Afshan et al., 2018; Mobram and Vali, 2022). Furthermore, the correlation between MFCC i-vectors and MFCC features has been determined, and the effectiveness of i-vectors has been examined in diagnosing major depressive disorder (MDD) (Di et al., 2021). A comparison of various i-vectors based on spectral features, prosodic features, formants, and voice quality for clinical depression detection during the interview discovered that spectral feature i-vectors gained the highest accuracy in distinguishing between the speech of depressed and control (Xing et al., 2022). I-vectors can limit speaker and channel variability, which helps the model focus more on depression detection. With the development of the embedding technique, Deep Neural Network (DNN) embeddings, fixed-dimensional speaker embeddings extracted from a feed-forward DNN outperformed i-vectors for text-independent speaker verification on short speech segments (Snyder et al., 2017). X-vectors, the new state-of-the-art speaker embeddings, have been applied in speaker recognition (Snyder et al., 2017, 2018, 2019; Garcia-Romero et al., 2019). The encoder networks of x-vectors include the following categories: TDNN (Waibel et al., 1989), Extended TDNN architecture (E-TDNN) (Snyder et al., 2019), the factorized TDNN (F-TDNN) with skip connections (Povey et al., 2018), and Resnet 2D (He et al., 2016). Experiments show that x-vectors can capture spoken content and channel-related information (Raj et al., 2019). Furthermore, the TDNN x-vectors or F-TDNN x-vectors based on MFCC have demonstrated better performance than PLP i-vectors for the automatic detection of PD (Moro-Velazquez et al., 2020). Besides, the x-vector technique has been used as an advanced method for emotion recognition (Pappagari et al., 2020b), AD detection (Pappagari et al., 2020a), and depression detection (Dumpala et al., 2021, 2022; Egas-López et al., 2022). Consequently, depression detection is carried out in this study using the x-vector approach with the i-vector framework as the baseline.

One unavoidable problem is that the amount of depression data limits that model training. Publicly available and commonly used depression speech datasets are the Audio-Visual Emotion Recognition Challenge and Workshop (AVEC) 2013 (Valstar et al., 2013), including 340 video clips from 292 subjects, and AVEC 2014 (Valstar et al., 2014), including 150 files of 84 speakers. DNN trained on such data would lead to under-fitting; consequently, the classification result needs to be more convincing. One workable solution to the above problem is to pre-train a model on extensive data followed by leveraging the model's knowledge to downstream tasks [e.g., speaker recognition (Snyder et al., 2018), PD detection (Moro-Velazquez et al., 2020), depression detection (Zhang et al., 2021)]. Primarily, results in Zhang et al. (2021) showed that the larger out-domain (e.g., speech recognition) dataset for audio embedding pre-training generally improves performance better than the relatively little in-domain (depression detection) dataset. Therefore, we pre-trained speaker embedding extractors on CN-Celeb (Fan et al., 2020), a large-scale Chinese speaker recognition dataset, followed by extracting corresponding embeddings on our Chinese depression speech dataset.

The method of training models with classification algorithms has occurred frequently in depression detection. SVM and RF were used for depression classification not only on low-level descriptors (LLD) and related functionals in Tasnim and Stroulia (2019) but also on i-vectors in Xing et al. (2022). On the other hand, the results of Saidi et al. (2020), comparing the baseline CNN model with the model combining CNN and SVM, have shown that the SVM classifier improved the classification accuracy. An exploratory study (Espinola et al., 2021), which compared experimental results of MLP, Logistic Regression (LR), RF, Bayes Network, Naïve Bayes, and SVM with different kernels, concluded that RF provided the highest accuracy among all classifiers for MDD detection. Therefore, SVM and RF were preferred as classification algorithms to evaluate speaker embeddings' performance in our study comprehensively.

There have been studies showing that there are differences between depressed and normal subjects' speech under different speech task stimuli. The collection of spontaneous and read speech from 30 depressed and 30 control subjects was used to extract acoustic features (Alghowinem et al., 2013). DEPression and Anxiety Crowdsourced corpus (DEPAC) (Tasnim et al., 2022), which has a diversity of speech tasks (Phoneme fluency, Phonemic fluency, Picture description, Semantic fluency, and Prompted narrative), has been published recently as a depression and anxiety detection corpus. Furthermore, the classification results in Long et al. (2017) based on the corpus of three speech types (reading, picture description, and interview), each of which corresponds to three emotional valences (negative, neutral, and positive), showed that speaking style and mood had a significant influence on depression recognition. From the theory of ensemble learning, combining multiple learners makes a whole's generalization ability usually much more robust than a single learner (Zhou, 2021). Also, multiple speech modes with different affective valence are natural learners. As a result, this study combined the information of nine speech tasks under multiple scenes and emotional valences using the ensemble learning method to improve the depression recognition ability of the model.

The proposed depression detection system was based on the speaker embedding framework and a multi-task ensemble learning approach. The whole process was divided into two stages. The first stage is the process of pre-training speaker embedding extractors. Nine speaker embedding extractors that differed in the front-end features and framework were trained on CN_Celeb. Three front-end feature sets contained MFCC, PLP, and FBANK. Three embedding frameworks contained i-vector, TDNN, and Resnet. In this stage, each speaker embedding extractor could change frame-level features of different lengths into fixed lengths and, more importantly, overcome the challenge of insufficient depression data volume. The second stage is to extract speaker embeddings of the depression dataset and make further classification. The same front-end features were extracted for the depression data of nine tasks, and we obtained the corresponding speaker embeddings using the pre-trained extractors. The depression classification percentage of nine utterances from one subject attained by the SVM classifier were aggregated into integrated features. The final results were then obtained using MLP based on the new features.

The main contributions of this paper are as follows:

1. The speaker embedding extractors were pre-trained on the large-scale out-domain dataset to alleviate the problem of insufficient depression data for depression recognition.
2. We have proved that based on MFCC, PLP, and FBANK, Resnet x-vectors, which are first used to detect depression, outperform TDNN x-vectors, and i-vectors.
3. In the depression detection task, interview speech caught more acoustic differences between depressed and normal subjects than picture description speech. Neutral stimuli performed better compared to positive and negative stimuli.
4. The effectiveness of our multi-task ensemble learning approach was verified on multiple speaker embeddings. Moreover, our multi-task ensemble learning method with Resnet x-vectors can effectively identify depressed patients.

3. Database

Two speech corpora were employed in this study: the first, CN-Celeb, is an extensive Chinese speaker recognition dataset collected 'in the wild' for training i-vector, TDNN x-vector, and Resnet x-vector extractors; the other, the depression speech dataset, is a corpus containing recordings from normal and depressed subjects and was utilized to extract speaker embeddings (i-vectors, TDNN x-vectors, and Resnet x-vectors) and to train back-end classifiers and multi-task ensemble learning models to evaluate their performance in automatic depression detection.

3.1. CN-Celeb

CN-Celeb (Fan et al., 2020) contains more than 130,000 utterances from 1,000 Chinese celebrities, covering 11 different speech scenarios. We chose CN-Celeb for three reasons: its large quantity, which is an indispensable part of the pre-trained model; the language of all recordings is Chinese, which is the same as that of the depression dataset; and its rich speech genres, some of which match the tasks of the depression dataset. Because the task type of the depression speech dataset used in this experiment is interview and picture description, which are all spontaneous speech, the average length of each utterance is longer than 10 s. Based on the comprehensive consideration of speech modes and average duration of each utterance, we select all the speech in the interview and speech scenes of CN-Celeb. The subset includes 67,718 utterances from 902 Chinese celebrities with a total length of 171.99 h. The interview scenario contains 780 subjects with 59,317 utterances and lasts 135.77 h. As for the speech genre, 8,401 utterances from 122 speakers were collected, with a length of 36.22 h. All of them were sampled at 16 kHz.

3.2. Depression speech database

We collected speech data from Beijing Anding Hospital, Lanzhou University Second People's Hospital, and Tianshui Third People's Hospital. All subjects were aged between 18 and 55, native Chinese speakers, and had a primary school education or above. The patients were required to meet DSM-IV criteria (American Psychiatric Association, 1994) with the Patient Health Questionnaire-9 (PHQ-9) (Kroenke et al., 2001) score of 5 or greater and not to have taken any psychotropic drugs during the first 2 weeks of enrollment. In comparison, the control subjects had no definite mental disorder diagnosis and regular mental activity. In order to obtain high-quality speech data, the experiment was conducted in a room with good sound insulation and no electromagnetic interference, and the ambient noise was ensured to be lower than 60 dB. For the purpose of avoiding the distortion of the voice data, a high-precision sound card and microphone were used. The recordings were saved in Waveform Audio File Format (WAV) with a sampling rate of 44.1 kHz and a sampling width of 24 bit. The preprocessing steps of speech signal mainly included pre-emphasis, frame segmentation, and endpoint detection.

TABLE 1 Details of nine tasks.

Task	Genres	Valences	Problems
Task1	Interview	Positive	If you have a vacation to travel, please describe your travel plans.
Task2	Interview	Positive	Please share what you think is a good memory and briefly describe the scene.
Task3	Interview	Neutral	How are you feeling these days? How does this affect your life?
Task4	Interview	Neutral	How is your health these days? How has it affected your life?
Task5	Interview	Neutral	How do you rate yourself?
Task6	Interview	Negative	Describe an event that caused you great pain.
Task7	Picture description	Positive	Describe the positive facial expression, and guess the reason for the expression.
Task8	Picture description	Neutral	Describe the neutral facial expression, and guess the reason for the expression.
Task9	Picture description	Negative	Describe the negative facial expression, and guess the reason for the expression.

This dataset followed two different experimental paradigms whose intersection contained 9 identical speech tasks, including six interview tasks and three picture description tasks with three emotions (positive, neutral, and negative). The specific tasks are listed in **Table 1**. With regard to the evaluation of the valence of interview questions, we recruited 33 volunteers to score the valence and arousal of these questions, respectively, and then divided them into three types according to the degree of pleasure: positive, neutral and negative. The face images displayed in the picture description scene were taken from the Chinese facial affective picture system (CAPS) (Gong et al., 2011), which contains 870 facial images of seven emotions: anger, disgust, fear, sadness, surprise, happiness, and calm. The evaluation is conducted from the three dimensions of pleasure, arousal, and dominance. We selected three female face images of happiness, calm, and sadness as the picture description materials of positive, neutral, and negative stimuli. After Voice Activity Detection (VAD) to all recordings, data from 536 subjects, including 226 normal subjects and 310 depressed subjects, were preprocessed and retained. Each participant contained nine speech segments. Details of the depression speech dataset used in this study are shown in **Table 2**,

TABLE 2 Details of subjects' information.

Subject type	Gender	Subject numbers	Utterance numbers	Age mean (standard deviation)	PHQ-9 mean (standard deviation)	Utterance duration mean(s)
Depression	Male	142	1,278	37.03 (10.88)	14.49 (7.15)	20.74
	Female	168	1,512	38.23 (12.14)	14.85 (8.24)	
Normal	Male	119	1,071	36.00 (10.82)	1.47 (2.31)	15.50
	Female	107	963	33.36 (10.53)	1.42 (0.69)	

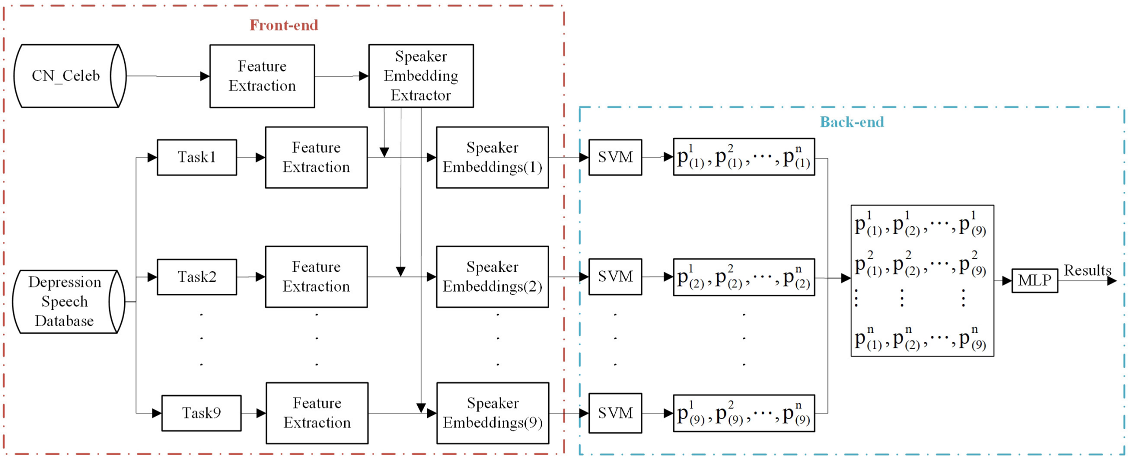

Front-end
CN_Celeb
Feature Extraction
Speaker Embedding Extractor
Depression Speech Database
Task1
Task2
Task9
Feature Extraction
Speaker Embeddings(1)
Speaker Embeddings(2)
Speaker Embeddings(9)
Back-end
SVM
$p^1_{(1)}, p^2_{(1)}, \cdots, p^n_{(1)}$
$p^1_{(2)}, p^2_{(2)}, \cdots, p^n_{(2)}$
$p^1_{(9)}, p^2_{(9)}, \cdots, p^n_{(9)}$
$p^1_{(1)}, p^1_{(2)}, \cdots, p^1_{(9)}$
$p^2_{(1)}, p^2_{(2)}, \cdots, p^2_{(9)}$
$p^n_{(1)}, p^n_{(2)}, \cdots, p^n_{(9)}$
MLP
Results

FIGURE 1
General methodology diagram of the proposed depression detection system. The acoustic feature could be MFCC, PLP, or FBANK. The speaker embedding extractor type can be i-vector, TDNN x-vector, and Resnet x-vector, and the speaker embedding type is derived from the extractor type. *n* represents the number of training subjects. The subjects are divided into 10 folds according to the 10-fold cross-validation rule, in which nine folds are used for training and onefold for testing.

including the subject number, utterance number, age, PHQ-9 score, and the average duration of each utterance in the two groups.

4. Methodology

The method proposed in this paper aims to improve depression classification performance using integrated learning combined with a pre-trained speaker embedding system and multiple speech task stimuli. Figure 1 shows a general block diagram of the depression detection system used in this study.

Firstly, the speech features are extracted from the preprocessed utterances (Section "4.1. Acoustic feature extraction"). Next, the speaker embedding extractors are pre-trained based on acoustic features of the out-domain dataset, and speaker embeddings of the multi-task in-domain dataset are extracted (Section "4.2. Speaker embedding extraction"). In order to take advantage of the effects of nine tasks, the multi-task integrated learning approach is carried out in Section "4.3. Multi-task ensemble learning method." These are described in detail below.

4.1. Acoustic feature extraction

Three acoustic feature sets, including MFCC, PLP, and FBANK, were extracted from each utterance of both CN-Celeb and our depression speech dataset in this study. This process was implemented by Kaldi Toolbox (Povey et al., 2011). We used three kinds of frame-level representations: 60-dimensional MFCCs, 60-dimensional PLPs, and 60-dimensional FBANKs, all with a Hamming window, a frame-length of 25 ms, and a frame-shift of 10 ms.

Mel Frequency Cepstral Coefficients was proposed based on the acoustic characteristics of the human ear, which could be understood as the energy distribution of speech signals in different frequency ranges. MFCC often serves as a standard to fit i-vector models (Di et al., 2021) or x-vector models (Egas-López et al., 2022), or other deep network models (Rejaibi et al., 2022). The literature results convince us that MFCC can contribute to the training of speaker embedding systems.

PLP was proposed using the results obtained from human auditory experiments, and it was beneficial to extract anti-noise speech features. The results of Moro-Velazquez et al. (2020) comparing the i-vector extractors based on PLP and the x-vector extractors based on MFCC showed that the two systems had their advantages in PD detection. Therefore, we extracted PLP for a comparative study of depression recognition.

The response of the human ear to the sound spectrum is nonlinear. FBANK is a front-end processing algorithm that can improve speech recognition performance by processing audio similarly to the human ear. The literature demonstrated that FBANK was more effective than MFCC in x-vector training for the Escalation SubChallenge (José Vicente et al., 2021) and depression assessment (Egas-López et al., 2022). Consequently, FBANK was

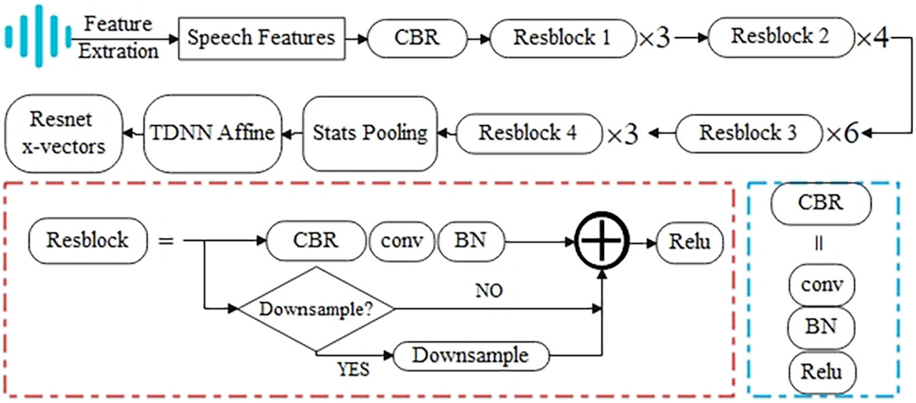


FIGURE 2
The block diagram of the Resnet x-vector extraction process.

also extracted in this study for subsequent training of speaker embedding extractors.

4.2. Speaker embedding extraction

In this study, three frameworks were performed to train different types of speaker embedding extractors based on the acoustic characteristics of CN-Celeb. The task of pre-training is to improve the performance of speaker recognition. We transferred the knowledge learned in the pre-training process to the depression recognition task, that is, to retain the extractors obtained in the upstream task. We applied them to the speaker embedding extraction on phonetic features of the depression speech database. Note that i-vectors served as a classic baseline method without deep learning and TDNN x-vectors served as a DNN baseline. We focused on a new state-of-the-art speaker recognition method: the Resnet x-vectors in depression detection. The procedure of i-vector extraction was carried out using Kaldi. At the same time, the extraction of TDNN x-vectors and Resnet x-vectors was implemented on ASV-Subtools (Tong et al., 2021).

4.2.1. I-vector extraction

The i-vector framework can map speech recordings of arbitrary duration to low dimensional space, and a compact representation of fixed length is obtained. Acquiring the Universal Background Model (UBM) is to train a diagonal covariance matrix and a full matrix on all training subjects' speech data. UBM is a speaker–and channel-independent Gaussian Mixture Model (GMM), which can be regarded as the unified reference coordinate space of the training set. When initializing UBM, the number of Gaussian components, denoted as C, must be set. The ith ($i = 1, 2, \ldots, C$) Gaussian component includes a weight (w_i), a mean vector (μ_i), and a covariance matrix (Σ_i). Thus, the Gaussian mean supervector (m) of UBM can be obtained. Furthermore, the Gaussian mean supervector (M) of the utterance (h) from the speaker (s) is defined as follows:

$$M_{s,h} = m + T\omega_{s,h} \tag{1}$$

Different from the two spaces (a speaker subspace and a session subspace) included in the Joint Factor Analysis (JFA) model, the total variability space (T), which contains the speaker and channel effects simultaneously, is employed in the i-vector model (Dehak et al., 2010). ω is the total variability space factor, and its maximum-a-posteriori (MAP) point estimate is the i-vector. After UBM training, the Baum-Welch statistics of each speaker in the training set are calculated, and T is iteratively estimated by the Expectation-Maximization (EM) algorithm. $M_{s,h}$ is obtained using MAP adaptation followed by the estimation of i-vectors based on $\omega_{s,h}$. More details on the calculation of Baum-Welch statistics and i-vector estimation can be sought out in Dehak et al. (2010).

In this study, we set the number of Gaussian components as 256 and the i-vector dimension as 256.

4.2.2. TDNN x-vector extraction

The TDNN x-vector approach provides a fixed-dimensional utterance-level representation by using a time-delay neural network and the features of variable-length speech. Extracting TDNN x-vectors contains several steps. Firstly, the TDNN architecture runs at the frame level. The current time step is represented by t. The input to the next frame-level layer is concatenated from the current frame and its context of past and future frames. Therefore, the next layer of frame-level representation condenses the temporal context information. As the network deepens gradually, the scope of the temporal context becomes wider. After three time-delay operations, one frame in the fourth layer corresponds to 15 frames in the context of the first layer. The stats pooling layer aggregates all the frames of the speech segment and calculates the mean and standard deviation. Finally, TDNN x-vectors are obtained in the segment-level layer.

Time delay neural network x-vectors and Resnet x-vector extractors were trained on the Pytorch framework. The speech utterances were divided into chunks of 200 frames, and we set the batch size as 64. Moreover, the dimension of TDNN x-vectors and Resnet vectors was 256, the same as that of i-vectors. The process of the Resnet x-vector extraction is detailed in Section "4.2.3. Resnet x-vector extraction." We used a ralamb optimizer containing LookAhead and RAdam optimizer with Layer-wise Adaptive Rate Scaling (LARS). The learning rate was set to 0.001, attenuating every 400 steps and an attenuation factor of 0.7. The number of training sessions was 18.

TABLE 3 Resnet encoder architecture.

Layer	Input	Output	Down sample	Kernel	Stride	Channels	Blocks
Conv1	$F \times T$	$F \times T$	False	7×7	1	32	–
Resblock1	$F \times T$	$F \times T$	False	3×3	1	32	3
Resblock2	$F \times T$	$\frac{F}{2} \times \frac{T}{2}$	True	3×3	2	64	4
Resblock3	$\frac{F}{2} \times \frac{T}{2}$	$\frac{F}{4} \times \frac{T}{4}$	True	3×3	2	128	6
Resblock4	$\frac{F}{4} \times \frac{T}{4}$	$\frac{F}{8} \times \frac{T}{8}$	True	3×3	2	256	3
Stats pooling	$\frac{F}{8} \times \frac{T}{8}$	$\frac{F}{4} \times 1$	–	–	–	256	–
TDNN affine	$\frac{F}{4} \times 1$	1×1	–	$\frac{F}{4} \times 1$	1	256	–

F is the feature dimension (F = 60 for MFCC, PLP, and FBANK), and T is the sequence length.

4.2.3. Resnet x-vector extraction

Residual learning was proposed to simplify training for deeper networks (He et al., 2016). We followed the Resnet34 encoder described by Villalba et al. (2020) to train Resnet x-vector extractors. Figure 2 shows the block diagram of the Resnet x-vector extraction process. Specific architecture of the Resnet encoder is listed as Table 3. The repetition times of the four residual blocks are 3, 4, 6, and 3, respectively, and the number of residual block channels is gradually doubled from 32 to obtain deeper information. The dimension of acoustic features and the number of speech frames are denoted as F and T, respectively. When the stride is set to 2, the dimensions of F and T to the output are halved. Due to the addition operation in residual blocks, the input needs to be downsampled to ensure the same dimensions before adding. Finally, each speech segment can obtain Resnet x-vectors of fixed length after the average pooling layer.

In this study, Adam Weight Decay Regularization optimizer was used in Resnet, and the learning rate was set to 0.001. The attenuation factor was 1.0, and the number of training sessions was 21.

4.3. Multi-task ensemble learning method

In the front-end of the multi-task ensemble learning method, nine speaker embeddings with nine task stimuli were extracted from three acoustic features. The symbolic marks of speaker embeddings are shown in Table 4. The acoustic features can be MFCC, PLP, and FBANK. The types of speaker embedding extractors in pre-training can be i-vector, TDNN x-vector, and Resnet x-vector. Speaker embeddings are extracted according to the speaker embedding extractors. In the back-end part, $Speaker\ Embeddings(j)$ and $\mathrm{p}^{i}_{(j)}$ represent speaker embeddings of the jth speech task and the SVM classification result of the jth speech task from the ith $(i = 1, 2, \ldots, n)$ subject, respectively. Then, all the training set results are spliced and transposed into the matrix. The same operation is performed for the testing set, and the results of this fold are obtained by using MLP.

TABLE 4 The denotation of speaker embeddings.

Denotation	Description
R_m	Resnet x-vectors based on MFCC
R_p	Resnet x-vectors based on PLP
R_f	Resnet x-vectors based on FBANK
T_m	TDNN x-vectors based on MFCC
T_p	TDNN x-vectors based on PLP
T_f	TDNN x-vectors based on FBANK
I_m	I-vectors based on MFCC
I_p	I-vectors based on PLP
I_f	I-vectors based on FBANK

4.4. Combination of different Resnet x-vectors in multi-task ensemble learning method

This study also combined different Resnet x-vectors in our proposed multi-task ensemble learning method. Resnet x-vectors based on different speech features contain different acoustic information, which may play a complementary role in depression recognition. Figure 3 shows that the classification results of three Resnet x-vectors (R_m, R_p, and R_f) on the training partition using SVM are fused into new features in nine tasks, and MLP is carried out to train the optimal model on the training set.$p^{i}R\_m(j)$, $p^{i}R\_p(j)$, and $p^{i}R\_f(j)$ represent the SVM classification result of the jth speech task from the ith $(i = 1, 2, \ldots, n)$ subject based on R_m, R_p, and R_f, respectively. Although Figure 3 shows the fusion process of three Resnet x-vectors, the experiment also carries out fusion cases of two Resnet x-vectors. Additionally, the figure only shows the result of one test fold; the final result is the average of 100 repetitions of 10-fold cross-validation.

5. Experimental results

Our experiments have done the following work: In Section "5.1. Results of nine speaker embeddings for depression detection," we use SVM and RF to compare the performance of nine speaker embeddings in nine tasks. We analyze the performance difference of the Resnet x-vector extractor compared with the TDNN x-vector extractor and the i-vector extractor, the impact of different acoustic features on the three speaker embedding extractors, and the impact of different speech task types and emotional valences on speaker embeddings. In Section "5.2. Results

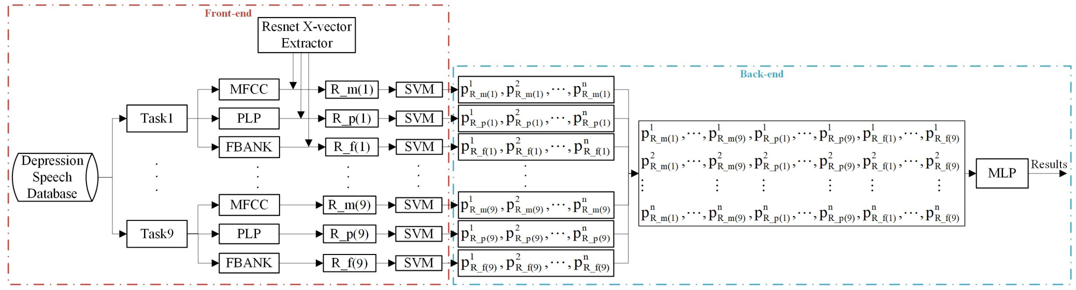


FIGURE 3
Resnet x-vector fusion of the proposed depression detection system. n denotes the number of the training subjects.

of multi-task ensemble learning methods with speaker embeddings for depression detection," we compare the performance of our multi-task ensemble learning method and the other two literature methods in nine speaker embeddings. Moreover, the best effect is obtained by fusing Resnet x-vectors based on different features in the integrated learning method and then compared with the proposed literature studies.

In order to fully evaluate the performance of multiple speaker embeddings and ensemble learning methods in depression detection, we used accuracy, F1-D, and F1-H as performance metrics. F1-D and F1-H are F1 scores of depressed and healthy classes, respectively. For the binary problem in this paper, the four categories in the confusion matrix are True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN). The accuracy, F1-D, and F1-H could be calculated as follows.

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{2}$$

$$F1 - D = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{3}$$

$$F1 - H = \frac{2 \times TN}{2 \times TN + FP + FN} \tag{4}$$

Besides, 100 repetitions of 10-fold cross-validation were employed to examine the algorithm's performance. A total of 536 subjects (310 depressed and 226 normal) in the depression speech dataset were divided into 10 non-overlapping folds according to the proportions of the two classes. Six folds were 54 subjects (31 depressed and 23 normal), and the four folds were 53 subjects (31 depressed and 22 normal). We used ninefolds for training and the remaining fold for testing. This way, the same utterance would not appear in two different folds. The KFold function of the Scikit-learn toolbox (Pedregosa et al., 2011) (sklearn) was performed to partition the training and testing sets. The result of each repetition was an average of 10 test folds. In order to assess the generalizability of our approach, the final result was the average of 10-fold cross-validation for 100 times with different random_state (Mobram and Vali, 2022) has used this experimental scheme.

5.1. Results of nine speaker embeddings for depression detection

After the implementation of the front-end part of the experimental framework in Figure 1, nine speaker embeddings in nine speech tasks were obtained. Two classifiers, SVM and RF, were used to evaluate the depression recognition performance of nine speaker embeddings comprehensively. We trained SVM classifiers with a Gaussian kernel function and tuned the SVR hyper-parameters. Similarly, n_estimators, which represented the number of trees in the forest, were optimized when training RF classifiers. Concerning the experiments of speaker embeddings on each task, the training partition was used to train models, and the results were calculated on the testing partition. The experiments followed the 10-fold cross-validation rule and were repeated 100 times with different randomizations. The accuracies of nine speaker embeddings under nine tasks using SVM and RF were reported in Table 5. The detailed meanings of the nine speaker embedding nicknames in this table are shown in Table 4. We also calculated the corresponding F1-D and F1-H, but they were too long to be listed. However, they would be used in the subsequent comparison of the algorithm's performance.

5.1.1. The effects of different speaker embedding extractors on depression detection system

Figure 4 showed classification accuracy, F1-D, and F1-H of speaker embeddings based on three extractors and the performance differences between SVM and RF. This boxplot was drawn by the results of speaker embeddings under different extractors, as described in Section "5.1. Results of nine speaker embeddings for depression detection." For instance, the accuracy boxplot under the i-vector extractor using SVM in Figure 4A was made based on all results of I_m, I_p, and I_f under nine tasks in Table 5.

The accuracies shown in Figure 4A indicated that the Resnet x-vector extractor provided the best scores, followed by the TDNN x-vector extractor and the i-vector extractor in both SVM and RF. In detail, regardless of whether SVM or RF was used, the upper limit, median and lower limit of the Resnet x-vector extractor were highest, while those of the i-vector extractor were lowest. Although the maximum accuracy of TDNN x-vectors in SVM reached 74.51%, this number was judged as an outlier based on the overall distribution of the boxplot. Additionally, it clearly showed that the box of Resnet x-vectors was overall above the other two. Figure 4B, F1-D of Resnet x-vectors and i-vectors were close, while TDNN x-vectors were slightly inferior. Figure 4C showed that the ranking of F1-H of the three extractors was consistent with that of accuracies.

As could be seen from the results of three assessment criteria under the two classifiers, the Resnet x-vector extractor

TABLE 5 Accuracy comparison of nine speaker embeddings under nine speech tasks using SVM or RF classifier.

SVM	I_m	I_p	I_f	T_m	T_p	T_f	R_m	R_p	R_f
Task1	58.89%	57.71%	57.31%	62.06%	61.26%	62.45%	68.58%	60.67%	67.59%
Task2	61.46%	60.47%	60.67%	62.65%	59.29%	64.62%	63.83%	60.87%	62.45%
Task3	64.62%	65.81%	63.44%	66.01%	67.39%	68.38%	67.79%	62.06%	65.02%
Task4	70.36%	72.33%	67.79%	71.74%	71.34%	74.51%	70.75%	71.74%	71.34%
Task5	57.71%	58.30%	62.25%	67.39%	62.45%	62.85%	65.81%	62.65%	62.06%
Task6	57.71%	60.08%	59.68%	62.06%	61.66%	60.47%	64.82%	60.67%	60.47%
Task7	60.67%	62.25%	60.08%	61.46%	62.06%	61.07%	63.04%	59.88%	65.42%
Task8	62.85%	58.89%	58.70%	59.29%	62.25%	59.29%	64.23%	60.47%	62.06%
Task9	64.43%	61.66%	61.86%	59.68%	59.88%	59.09%	64.23%	61.26%	64.23%
RF	I_m	I_p	I_f	T_m	T_p	T_f	R_m	R_p	R_f
Task1	61.07%	60.28%	58.50%	61.66%	63.04%	61.26%	66.80%	59.29%	65.81%
Task2	60.28%	61.26%	59.68%	60.67%	59.88%	59.68%	64.23%	60.47%	60.67%
Task3	63.64%	62.85%	62.45%	66.40%	67.59%	66.60%	64.43%	62.25%	64.23%
Task4	66.21%	67.98%	64.52%	70.95%	73.72%	72.33%	68.18%	67.79%	68.58%
Task5	58.89%	60.28%	60.28%	62.85%	61.07%	63.83%	63.04%	60.47%	61.07%
Task6	60.28%	59.09%	61.07%	63.44%	58.70%	58.89%	64.82%	59.68%	59.29%
Task7	61.86%	59.49%	60.08%	61.46%	63.64%	61.66%	66.40%	62.85%	66.01%
Task8	59.49%	57.31%	61.07%	59.09%	60.47%	59.49%	64.03%	59.49%	62.25%
Task9	60.67%	60.89%	62.25%	58.30%	60.47%	58.50%	64.43%	61.66%	63.24%

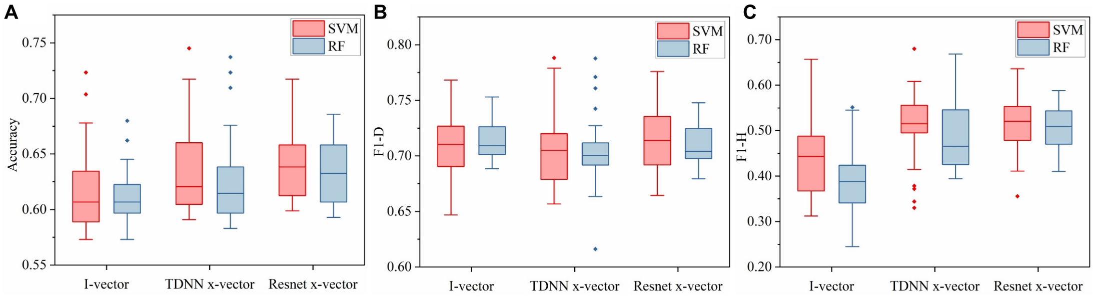


FIGURE 4
The result comparison of speaker embeddings based on three extractors in nine speech tasks between SVM and RF. **(A)** Accuracy boxplot. **(B)** F1-D boxplot. **(C)** F1-H boxplot.

outperformed the TDNN x-vector extractor, which indicated that the ability of upstream knowledge learned by Resnet to transfer to depression screening was stronger than TDNN. Moreover, the DNN embeddings (Resnet x-vectors and TDNN x-vectors) could utilize speakers' traits to build more effective depression models than i-vectors. The results of Egas-López et al. (2022) comparing the performance of DNN embeddings and i-vectors for depression discrimination also supported the above conclusion. It was worth noting that in the three charts of Figure 4, almost all upper limit, upper quartile, and median of the three extractors' whole measurement indicators in SVM were higher than RF. This point was consistent with the deduction of experiments that compared classification results of SVM and RF in various i-vectors (Xing et al., 2022). Consequently, we only contrasted the results of speaker embeddings under SVM in the subsequent analysis. On the other hand, we opted for SVM to train classifiers as the back-end part of the framework displayed in Figure 1 and then integrated nine speech tasks.

5.1.2. The effects of different acoustic features on depression detection system

This part was to find out the most suitable phonetic features for each speaker embedding extractor. The accuracy, F1-D, and F1-H of three speaker embedding extractors based on MFCC, PLP, and FBANK over nine tasks using SVM were plotted in Figure 5. In terms of i-vectors, the medians of three evaluation indicators of the MFCC-based systems exceeded those of systems based on PLP or FANK, and in Figure 5A, the upper limit and upper quartile of the accuracy of MFCC i-vectors were supreme among three i-vector extractors based on different characteristics. Accordingly,

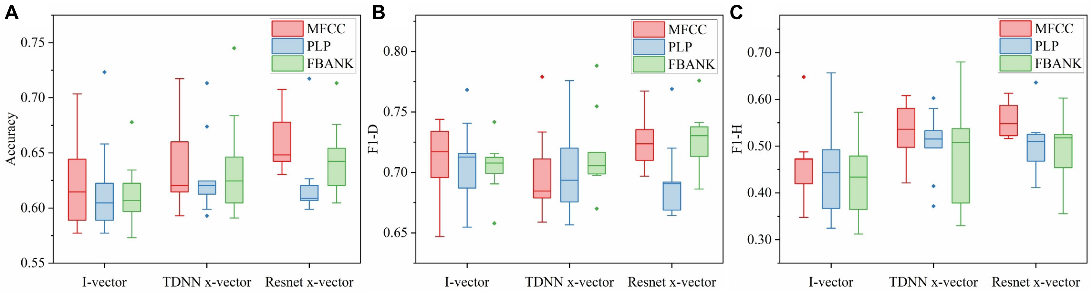


FIGURE 5

The result comparison of speaker embeddings based on different acoustic features in nine speech tasks using SVM. **(A)** Accuracy boxplot. **(B)** F1-D boxplot. **(C)** F1-H boxplot.

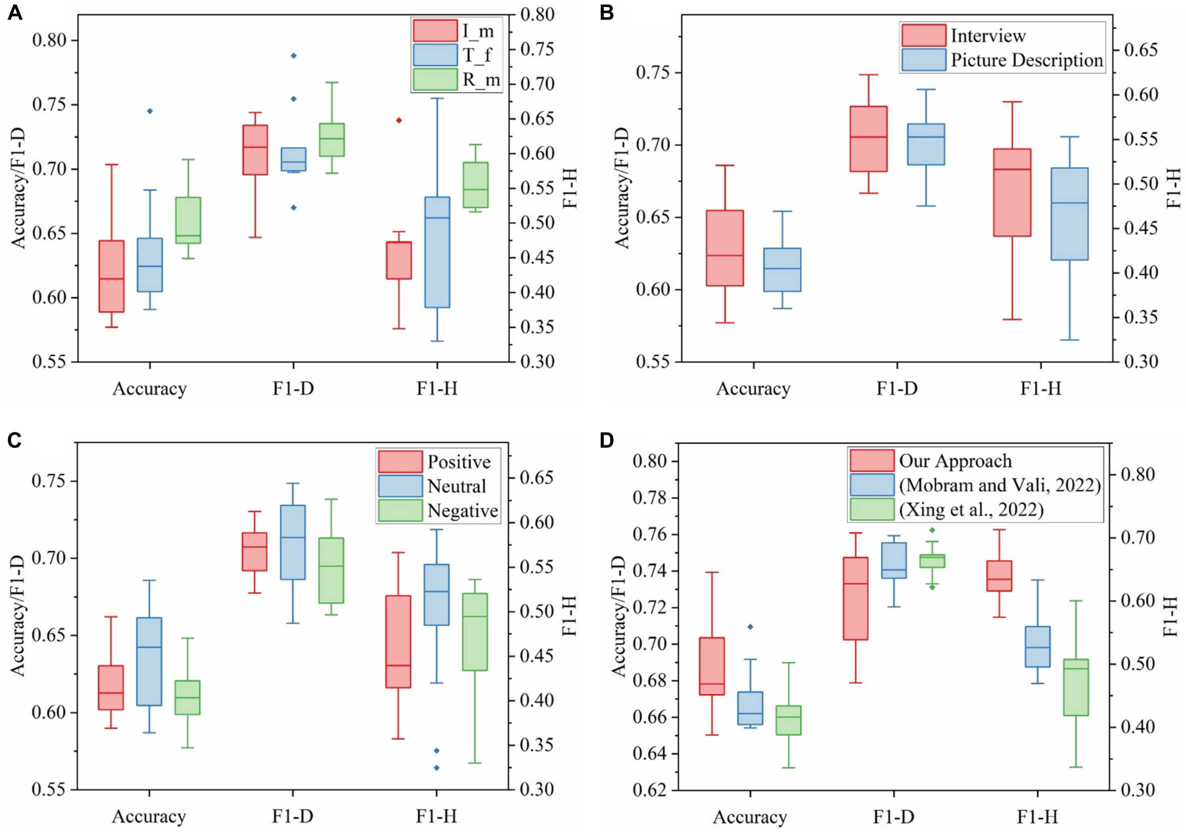


FIGURE 6

(A) The result comparison of each extractor based on the most matching feature set in nine speech tasks using SVM. **(B)** The result comparison of speaker embeddings in speech tasks of different genres using SVM. **(C)** The result comparison of speaker embeddings in speech tasks of different emotions using SVM. **(D)** The result comparison of three ensemble learning methods with speaker embeddings combined with nine tasks.

MFCC was more suitable for i-vectors. In addition, (Di et al., 2021) demonstrated the effectiveness of MFCC i-vectors in the clinical diagnosis of MDD. From the comprehensive analysis of the three boxplots in Figure 5, FBANK outperformed the other feature sets in TDNN x-vectors slightly. Although the accuracies of TDNN x-vectors based on the three feature sets were similar, the median of F1-D and the upper limit of F1-H of FBANK-based systems had advantages. It could also be seen in Egas-López et al. (2022) that TDNN x-vector extractors fitted with FBANK outperformed MFCC, which our results supported. As for the Resnet x-vector extractor, it could be observed that accuracy, F1-D, and F1-H of MFCC-based systems performed better than the other two. As far as we know, there is a lack of research on the befitting phonetic features of these speaker embedding extractors. The results of our experiment can provide some reference for this problem.

Since the i-vector and Resnet x-vector extractors best matched MFCC and the TDNN x-vector extractor best matched FBANK, Figure 6A showed the results of three speaker embeddings (I_m, T_f, and R_m) using SVM in nine tasks for depression classification. It was worth noting that five characteristic values of the accuracy of R_m were optimal, and its data is the most centralized. The upper limit and lower quartile of F1-D of R_m were significantly higher, and other characteristic values were not lower. The characteristic values of F1-H of R_m, except for the upper limit, were obviously better than others. As a result, R_m provided the most vital ability to recognize depression in nine tasks among nine speaker embeddings. However, the accuracy and F1-H of TDNN x-vectors were slightly better than those of i-vectors. Therefore, the performance of the three speaker embeddings was sorted from good to bad: R_m, T_f, and I_m. This conclusion could correspond to the performance ranking of three speaker embedding extractors in Section "5.1.1. The effects of different speaker embedding extractors on depression detection system."

5.1.3. The effects of different speech tasks on depression detection system

This series of analyses were conducted to investigate the influence of different genres and emotions of speech tasks on depression discrimination results of speaker embeddings. As mentioned in Table 1, there were nine tasks of the depression speech database covering two genres and three emotional valences. Table 6 integrated the accuracy of the same emotion in the same scenario in Table 5. Specifically, the accuracies of Int-Pos were the average of those of task1 and task2. The accuracies of Int-Neu were the average of those of task3 to task5. Also, the values of Int-Neg, Pic-Pos, Pic-Neu, and Pic-Neg corresponded to task6 to task9, respectively. F1-D and F1-H of six task types also performed similar operations. This operation ensured that the data volume of the six task types was the same and that the data distribution could be fairly compared through boxplots. Figure 6B and Figure 6C showed the results of nine speaker embeddings in the interview or picture description tasks and positive, neutral, or negative emotions using SVM. Moreover, the accuracy boxplots of both figures were plotted according to Table 6.

The results of Figure 6B presented that the interview scene had more considerable fluctuations of accuracy and F1-D. However, the upper limit, median, and upper quartile of the three assessment criteria were significantly higher than the picture description scene. Even all indexes of the F1-H boxplot of the interview were superior to the picture description. Overall, interview speech performed better than picture description speech using speaker embeddings in depression detection. Although both interview speech and picture description speech were considered as spontaneous voice, we inferred from our experimental results that subjects were more likely to express their true feelings in the interview scene, and interview voice contained more information related to emotional states than picture description. This view coincides with the conclusion of Long et al. (2017).

It could be seen from Figure 6C that the accuracy, F1-D, and F1-H of neutral stimulus materials were evidently superior to positive and negative materials. Although F1-H of positive speech had no advantage over negative speech, all indexes of its accuracy were slightly higher than the negative, and five characteristic values other than the upper limit of F1-D were higher than the negative. In addition, the fluctuation of F1-D of negative speech was also the smallest. Hence, it could be concluded that neutral stimulus materials performed best, followed by positive materials and negative materials. This discovery was consistent with (Liu et al., 2017), which showed that neutral stimuli performed best among three emotional valences when using speaker embeddings for depression detection.

5.2. Results of multi-task ensemble learning methods with speaker embeddings for depression detection

The back-end part of Figure 1 was conducted on nine speaker embeddings, and each integrated nine speech tasks. We implemented MLP using the GridSearchCV function from sklearn, which performed grid optimization of the parameters on the training set and then applied the optimal model on the training partition to the prediction of the testing partition. Note that the result in Figure 1 was just the result of a test fold, and our method's final result was the average of 10 test folds across 100 times.

Our approach was compared with two other ensemble methods. The first method (Mobram and Vali, 2022) was to classify speaker embeddings on nine speech tasks using cosine similarity and then a majority vote based on the results of nine tasks. The second method (Xing et al., 2022) used SVM on speaker embeddings over nine tasks and selected tasks with significant accuracy differences using paired *T*-test. Then the results of the different tasks were integrated into new features for SVM classification. The final results of these two methods were also the average of 100 repetitions of ten-fold cross-validation.

The experimental results in Table 7 indicated that three ensemble learning methods performed best on MFCC-based Resnet x-vectors, which were remarked as R_m among nine speaker embeddings, which illustrated the effectiveness of R_m in depression recognition tasks. In addition, our approach provided the best accuracy (73.94%), F1-D (76.09%), and F1-H (71.30%) on R_m with improvement by 2.99, 0.15, and 7.96% compared with

TABLE 6 Accuracy comparison of nine speaker embeddings under interview or picture description tasks with different emotions using SVM classifier.

Task	I_m	I_p	I_f	T_m	T_p	T_f	R_m	R_p	R_f
Int-Pos	60.18%	59.09%	58.99%	62.36%	60.28%	63.54%	66.21%	60.77%	65.02%
Int-Neu	64.23%	65.48%	64.49%	68.38%	67.06%	68.58%	68.12%	65.48%	66.14%
Int-Neg	57.71%	60.08%	59.68%	62.06%	61.66%	60.47%	64.82%	60.67%	60.47%
Pic-Pos	60.67%	62.25%	60.08%	61.46%	62.06%	61.07%	63.04%	59.88%	65.42%
Pic-Neu	62.85%	58.89%	58.70%	59.29%	62.25%	59.29%	64.23%	60.47%	62.06%
Pic-Neg	64.43%	61.66%	61.86%	59.68%	59.88%	59.09%	64.23%	61.26%	64.23%

TABLE 7 Performance comparison of three methods with speaker embeddings combined with nine tasks.

Speaker embeddings	Accuracy			F1-D			F1-H		
	Our approach	Mobram and Vali, 2022	Xing et al., 2022	Our approach	Mobram and Vali, 2022	Xing et al., 2022	Our approach	Mobram and Vali, 2022	Xing et al., 2022
I_m	65.04%	67.39%	65.22%	67.88%	75.91%	74.79%	61.61%	49.54%	43.95%
I_p	67.83%	65.42%	66.01%	71.25%	74.07%	76.24%	63.37%	48.07%	40.28%
I_f	67.23%	65.61%	63.45%	69.82%	74.56%	74.76%	64.07%	46.95%	33.69%
T_m	67.19%	67.00%	66.20%	73.31%	73.62%	74.21%	57.44%	55.94%	51.00%
T_p	70.36%	66.21%	67.39%	73.40%	73.73%	75.63%	66.52%	52.63%	50.75%
T_f	69.96%	69.17%	66.63%	76.03%	75.55%	74.89%	59.79%	58.29%	50.15%
R_m	73.94%	70.95%	68.99%	76.09%	75.94%	74.64%	71.30%	63.34%	60.05%
R_p	71.15%	65.42%	63.24%	74.74%	72.78%	73.12%	66.36%	52.57%	41.88%
R_f	67.25%	65.81%	65.05%	70.25%	72.05%	73.30%	63.44%	55.98%	49.28%

(Mobram and Vali, 2022) and 4.95, 1.45, and 11.25% over (Xing et al., 2022) on three assessment criteria. Figure 6D was drawn according to the data in Table 7, reflecting the performance of three methods over 9 speaker embeddings. It could be seen that the upper limit, median, and upper quartile of the accuracy of our method were higher than those of the rest two methods. Although F1-D of our approach was slightly lower than others, all indexes of F1-H of our approach were far superior to others. On the whole, the ensemble learning method we proposed performed well.

5.2.1. Combining different Resnet x-vectors in multi-task ensemble learning method

Since the advantages of Resnet x-vector extractors compared to TDNN x-vector and i-vector extractors had been explained in Section "5.1.1. The effects of different speaker embedding extractors on depression detection system," we would fuse different Resnet x-vectors (R_m, R_p, or R_f) in the multi-task integrated learning method as shown in Figure 3. The experiment was to examine the effect of this fusion on the performance of depression detection. It was not difficult to find from Table 8 that when R_m was eliminated from R_m + R_p + R_f, the accuracy, F1-D, and F1-H were reduced by 1.77, 1.19, and 3.17%, respectively. MFCC simulates the audio system of the human ear, which can suppress high-frequency signals, and reduce the interference of environmental noise. Therefore, R_m (MFCC-based Resnet x-vectors) did well in our experiment and provided a significant performance boost during the integration process. Moreover, the results in Table 8 indicated that R_m + R_p provided the highest accuracy (74.72%), F1-D (76.90%), and F1-H (72.05%), with the improvement of 0.78, 0.81, and 0.75% compared with R_m, and with the improvement of 3.57, 2.16, and 5.69% compared with R_p. PLP uses a linear prediction autoregressive model to obtain cepstrum coefficients, which is different from the compression coefficient used by MFCC. PLP also has good noise robustness. The combination of R_m and R_p should have better noise robustness than speaker embeddings before the combination. In this experiment, the speaker embeddings for depression recognition were based on the pre-trained model of out-domain data. It is very important to alleviate the interference of noise for the performance of the depression recognition model.

TABLE 8 Performance of ensemble fusion system of Resnet x-vectors based on different feature sets.

Ensemble fusion	Accuracy	F1-D	F1-H
R_m + R_p	74.72%	76.90%	72.05%
R_m + R_f	73.76%	75.42%	71.76%
R_p + R_f	69.60%	72.60%	65.78%
R_m + R_p + R_f	71.37%	73.79%	68.95%

TABLE 9 Performance comparison of other literature studies on the depression speech dataset.

Method	Accuracy	F1-D	F1-H
Giannakopoulos, 2015	67.98%	74.77%	56.22%
Di et al., 2021	66.40%	72.93%	55.73%
Egas-López et al., 2022	68.18%	75.42%	54.90%
Xing et al., 2022	71.89%	77.27%	63.08%
Our proposed system	74.72%	76.90%	72.05%

5.2.2. Comparison with other proposed methods on the depression speech dataset

This section compares the proposed multi-tasking integrated learning method incorporating different Resnet x-vectors with other literature studies, as shown in Table 9. Since the depression speech dataset used in this study was collected by the Gansu Provincial Key Laboratory of Wearable Computing, the results in Table 9 were obtained by implementing the methods in other papers based on this data. Note that the depression dataset was fairly divided into ten portions. Nine portions were for training, and one portion was for testing, which was unseen data. The final result of each method was the average of 100 repetitions of 10-fold cross-validation.

Our result in Table 9 is the best one of the completed outcomes: the fusion of the MFCC-based Resnet x-vectors and the PLP-based Resnet x-vectors in the multi-task ensemble learning method, with an accuracy of 74.72%, F1-D of 76.90%, and F1-H of 72.05%. Furthermore, our system increases accuracy by 6.74%, F1-D by 2.13%, and F1-H by 15.83% compared to Giannakopoulos (2015), which classified short-term and mid-term voice features from depressed and normal subjects using the SVM classifier with RBF kernel. Also, we improved accuracy by 8.32%, F1-D by 3.97%, and F1-H by 16.32% compared to Di et al. (2021), which only used MFCC i-vectors for depression detection and improved accuracy by 6.54%, F1-D by 1.48%, and F1-H by 17.15% compared to Egas-López et al. (2022) which used pre-trained DNN embeddings based on FBANK for SVM classification. Finally, compared to Xing et al. (2022), which was the hierarchical classification method of combined i-vectors based on several speech features that we published earlier, our accuracy is improved by 2.83% and F1-H by 8.97%, while F1-D is slightly lower.

In general, compared with other literature methods, the accuracy of our method has been improved to some extent, and F1-D, which presents the classification performance of the depressed class, also maintains a reasonable level. Particularly, F1-D, which shows the classification performance of the healthy class, has been significantly improved. This impressive result shows the effectiveness of our proposed method on the gender-independent depressive speech dataset.

6. Conclusion and future works

In order to find the optimal speaker embeddings for depression recognition, this paper compared the performance of three speaker embedding extractors based on different acoustic feature sets for depression detection in a multi-task depression speech database. The comprehensive performance of the new state-of-art Resnet x-vector extractor applied to depression recognition for the first time is better than that of the TDNN x-vector extractor and i-vector extractor, indicating that it can extract more depression-related information than the other two. Finally, nine speaker embeddings on three extractors (Resnet x-vector extractor, TDNN x-vector extractor, and i-vector extractor) based on MFCC, PLP, and FBANK were obtained. We concluded that MFCC was suitable for the i-vector extractor, FBANK for the TDNN x-vector extractor, and MFCC for the Resnet x-vector extractor. Moreover, MFCC-based Resnet x-vectors provided the best recognition among nine speaker embeddings.

Since our depression speech dataset consisted of nine speech tasks covering two genres (interview and picture description), and three emotional valences (positive, neutral, and negative), we explored the effects of different scenes and different emotional stimuli on depression recognition. The conclusion is that the difference in speech information between the two types of subjects in the interview task is more significant than that in the picture description task. The effect of neutral stimulus materials is better than that of positive and negative materials.

To make full use of the information from different scenes and emotions, we designed a multi-task ensemble learning method using speaker embeddings on the depression speech dataset containing nine tasks. The accuracy and F1-H of our method were significantly better than that of the other two literature studies, and F1-D maintained a similar level. In addition, the MFCC-based Resnet x-vectors among nine speaker embeddings performed best in our proposed integration approach. Our multi-task ensemble learning method based on R_m + R_p achieved best results than other literature studies using the depression speech database, indicating that MFCC-based Resnet x-vectors and PLP-based Resnet x-vectors were complementary in depression recognition, and information from 9 speech tasks was also utilized in the integrated system.

In this study, we used the out-domain dataset to train the pre-trained model to alleviate the problem of insufficient data volume in deep learning. We are also constantly collecting the depression speech dataset to expand the data volume. Then we will consider using the augmented in-domain dataset for pre-training to improve the depression recognition performance further.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: Data involves privacy and has not been disclosed. Requests to access these datasets should be directed to ZL, liuzhenyu@lzu.edu.cn.

Ethics statement

The studies involving human participants were reviewed and approved by the Tianshui Third People's Hospital. The patients/participants provided their written informed consent to participate in this study.

Author contributions

ZL, HY, and BH were responsible for the entire study, including study concepts and study design. ZL and HY contributed to the experimental paradigm design and wrote the manuscript. GL, QC, ZD, and LF helped collect data. ZY helped perform the analysis with constructive discussions. All authors agreed to be accountable for the content of the work.

Funding

This work was supported in part by the National Key Research and Development Program of China (Grant No. 2019YFA0706200), the National Natural Science Foundation of China (Grant Nos. 61632014, 61627808, 61802159, and 61802158), and the Fundamental Research Funds for Central Universities (lzujbky-2019-26 and lzujbky-2021-kb26).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Afshan, A., Guo, J., Park, S. J., Ravi, V., Flint, J., and Alwan, A. (2018). Effectiveness of voice quality features in detecting depression. *Proc. Interspeech* 2018, 1676–1680. doi: 10.21437/Interspeech.2018-1399

Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Breakspear, M., and Parker, G. (2013). "Detecting depression: A comparison between spontaneous and read speech," in *Proceedings of the 2013 IEEE international conference on acoustics, speech and signal processing*, (Vancouver, BC: IEEE), 7547–7551. doi: 10.1109/ICASSP.2013.6639130

Alpert, M., Pouget, E. R., and Silva, R. R. (2001). Reflections of depression in acoustic measures of the patient's speech. *J. Affect. Disord.* 66, 59–69. doi: 10.1016/S0165-0327(00)00335-9

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders: DSM-IV*, Vol. 4. Washington, DC: American Psychiatric Association.

Beck, A. T., and Alford, B. A. (2009). *Depression: Causes and treatment*. Philadelphia, PA: University of Pennsylvania Press. doi: 10.9783/9780812290882

Cai, H., Qu, Z., Li, Z., Zhang, Y., Hu, X., and Hu, B. (2020). Feature-level fusion approaches based on multimodal EEG data for depression recognition. *Inf. Fusion* 59, 127–138. doi: 10.1016/j.inffus.2020.01.008

Cannizzaro, M., Harel, B., Reilly, N., Chappell, P., and Snyder, P. J. (2004). Voice acoustical measurement of the severity of major depression. *Brain Cogn.* 56, 30–35. doi: 10.1016/j.bandc.2004.05.003

Cummins, N., Epps, J., Sethu, V., and Krajewski, J. (2014). "Variability compensation in small data: Oversampled extraction of i-vectors for the classification of depressed speech," in *Proceedings of the 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, (Florence: IEEE), 970–974. doi: 10.1109/ICASSP.2014.6853741

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., and Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Commun.* 71, 10–49. doi: 10.1016/j.specom.2015.03.004

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Trans. Audio Speech Lang. Process.* 19, 788–798. doi: 10.1109/TASL.2010.2064307

Di, Y., Wang, J., Li, W., and Zhu, T. (2021). Using i-vectors from voice features to identify major depressive disorder. *J. Affect. Disord.* 288, 161–166. doi: 10.1016/j.jad.2021.04.004

Dumpala, S. H., Rempel, S., Dikaios, K., Sajjadian, M., Uher, R., and Oore, S. (2021). "Estimating severity of depression from acoustic features and embeddings of natural speech," in *Proceedings of the ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, (Toronto, ON: IEEE), 7278–7282. doi: 10.1109/ICASSP39728.2021.9414129

Dumpala, S. H., Rodriguez, S., Rempel, S., Sajjadian, M., Uher, R., and Oore, S. (2022). Detecting depression with a temporal context of speaker embeddings. *Proc. AAAI SAS.*

Egas López, J. V., Tóth, L., Hoffmann, I., Kálmán, J., Pákáski, M., and Gosztolya, G. (2019). "Assessing Alzheimer's disease from speech using the i-vector approach," in *Proceedings of the international conference on speech and computer*, (Berlin: Springer), 289–298. doi: 10.1007/978-3-030-26061-3_30

Egas-López, J. V., Kiss, G., Sztahó, D., and Gosztolya, G. (2022). "Automatic assessment of the degree of clinical depression from speech using X-vectors," in *Proceedings of the ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, (Singapore: IEEE), 8502–8506. doi: 10.1109/ICASSP43922.2022.9746068

Espinola, C. W., Gomes, J. C., Pereira, J. M. S., and dos Santos, W. P. (2021). Detection of major depressive disorder using vocal acoustic analysis and machine learning—an exploratory study. *Res. Biomed. Eng.* 37, 53–64. doi: 10.1007/s42600-020-00100-9

Fan, Y., Kang, J. W., Li, L. T., Li, K. C., Chen, H. L., Cheng, S. T., et al. (2020). "Cn-celeb: A challenging Chinese speaker recognition dataset," in *Proceedings of the ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, (Piscataway, NJ: IEEE), 7604–7608. doi: 10.1109/ICASSP40776.2020.9054017

Garcia, N., Orozco-Arroyave, J. R., D'Haro, L. F., Dehak, N., and Nöth, E. (2017). "Evaluation of the neurological state of people with Parkinson's disease using i-vectors," in *Proceedings of the annual conference of the international speech communication association*, Stockholm, 299–303. doi: 10.21437/Interspeech.2017-819

Garcia-Romero, D., Snyder, D., Sell, G., McCree, A., Povey, D., and Khudanpur, S. (2019). x-vector DNN refinement with full-length recordings for speaker recognition. *Proc. Interspeech* 2019, 1493–1496. doi: 10.21437/Interspeech.2019-2205

Giannakopoulos, T. (2015). Pyaudioanalysis: An open-source python library for audio signal analysis. *PLoS One* 10:e0144610. doi: 10.1371/journal.pone.0144610

Gong, X., Huang, Y. X., Wang, Y., and Luo, Y. J. (2011). Revision of the Chinese facial affective picture system. *Chin. Ment. Health J.* 25, 40–46.

Greden, J. F., and Carroll, B. J. (1980). Decrease in speech pause times with treatment of endogenous depression. *Biol. Psychiatry* 15, 575–587. doi: 10.1007/BF00344257

Guo, W., Yang, H., Liu, Z., Xu, Y., and Hu, B. (2021). Deep neural networks for depression recognition based on 2d and 3d facial expressions under emotional stimulus tasks. *Front. Neurosci.* 342:609760. doi: 10.3389/fnins.2021.609760

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (Las Vegas, NV: IEEE), 770–778. doi: 10.1109/CVPR.2016.90

José Vicente, E. L., Kiss-Vetráb, M., Tóth, L., and Gosztolya, G. (2021). *Identifying conflict escalation and primates by using ensemble x-vectors and Fisher vector features.* Brno: ISCA.

Kanagasundaram, A., Vogt, R., Dean, D., and Sridharan, S. (2012). "PLDA based speaker recognition on short utterances," in *Proceedings of the speaker and language recognition workshop*, Singapore, 28–33. doi: 10.21437/Interspeech.2011-58

Kroenke, K., Spitzer, R. L., and Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *J. Gen. Intern. Med.* 16, 606–613. doi: 10.1046/j.1525-1497.2001.016009606.x

Lang, H., and Cui, C. (2018). Automated depression analysis using convolutional neural networks from speech. *J. Biomed. Inform.* 83, 103–111. doi: 10.1016/j.jbi.2018.05.007

Liu, Z., Hu, B., Li, X., Liu, F., Wang, G., and Yang, J. (2017). "Detecting depression in speech under different speaking styles and emotional valences," in *Proceedings of the international conference on brain informatics*, (Berlin: Springer), 261–271. doi: 10.1007/978-3-319-70772-3_25

Long, H., Guo, Z., Wu, X., Hu, B., Liu, Z., and Cai, H. (2017). "Detecting depression in speech: Comparison and combination between different speech types," in *Proceedings of the 2017 IEEE international conference on bioinformatics and biomedicine (BIBM)*, (Kansas City, MO: IEEE), 1052–1058. doi: 10.1109/BIBM.2017.8217802

Lu, J., Xu, X., Huang, Y., Li, T., Ma, C., Xu, G., et al. (2021). Prevalence of depressive disorders and treatment in China: A cross-sectional epidemiological study. *Lancet Psychiatry* 8, 981–990. doi: 10.1016/S2215-0366(21)00251-0

Lu, X., Shi, D., Liu, Y., and Yuan, J. (2021). Speech depression recognition based on attentional residual network. *Front. Biosci.* 26:1746–1759. doi: 10.52586/5066

Mobram, S., and Vali, M. (2022). Depression detection based on linear and nonlinear speech features in I-vector/SVDA framework. *Comput. Biol. Med.* 149:105926. doi: 10.1016/j.compbiomed.2022.105926

Moro-Velazquez, L., Villalba, J., and Dehak, N. (2020). "Using x-vectors to automatically detect Parkinson's disease from speech," in *Proceedings of the ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, (Barcelona: IEEE), 1155–1159. doi: 10.1109/ICASSP40776.2020.9053770

Mundt, J. C., Snyder, P. J., Cannizzaro, M. S., Chappie, K., and Geralts, D. S. (2007). Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. *J. Neurolinguistics* 20, 50–64. doi: 10.1016/j.jneuroling.2006.04.001

Nilsonne, Å. (1987). Acoustic analysis of speech variables during depression and after improvement. *Acta Psychiatr. Scand.* 76, 235–245. doi: 10.1111/j.1600-0447.1987.tb02891.x

Othmani, A., Kadoch, D., Bentounes, K., Rejaibi, E., Alfred, R., and Hadid, A. (2021). "Towards robust deep neural networks for affect and depression recognition from speech," in *Proceedings of the international conference on pattern recognition*, (New York, NY: Springer International Publishing), 5–19. doi: 10.1007/978-3-030-68790-8_1

Pappagari, R., Wang, T., Villalba, J., Chen, N., and Dehak, N. (2020b). "x-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *Proceedings of the ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, (Piscataway, NJ: IEEE), 7169–7173. doi: 10.1109/ICASSP40776.2020.9054317

Pappagari, R., Cho, J., Moro-Velazquez, L., and Dehak, N. (2020a). *Using state of the art speaker recognition and natural language processing technologies to detect Alzheimer's disease and assess its severity*. Shanghai: INTERSPEECH, 2177–2181. doi: 10.21437/Interspeech.2020-2587

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohamadi, M., et al. (2018). "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proceedings of the annual conference of the international speech communication association*, Hyberabad, 3743–3747. doi: 10.21437/Interspeech.2018-1417

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., et al. (2011). "The kaldi speech recognition toolkit," in *Proceedings of the IEEE 2011 workshop on automatic speech recognition and understanding*, (Waikoloa Village, HI: IEEE Signal Processing Society).

Raj, D., Snyder, D., Povey, D., and Khudanpur, S. (2019). "Probing the information encoded in x-vectors," in *Proceedings of the 2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, (Singapore: IEEE), 726–733. doi: 10.1109/ASRU46091.2019.9003979

Rajkumar, R. P. (2021). Gut hormones as potential therapeutic targets or biomarkers of response in depression: The case of motilin. *Life* 11:892. doi: 10.3390/life11090892

Rani, B. (2017). "I-vector based depression level estimation technique," in *Proceedings of the 2016 IEEE international conference on recent trends in electronics, information and communication technology (RTEICT)*, (Bangalore: IEEE), 2067–2071. doi: 10.1109/RTEICT.2016.7808203

Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S., and Othmani, A. (2022). MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomed. Signal Process. Control* 71:103107. doi: 10.1016/j.bspc.2021.103107

Saidi, A., Othman, S. B., and Ben, S. S. (2020). "Hybrid CNN-SVM classifier for efficient depression detection system," in *Proceedings of the international conference on advanced systems and emergent technologies, IC_ASET*, (Hammamet: IEEE), 229–234. doi: 10.1109/IC_ASET49463.2020.9318302

Shen, R., Zhan, Q., Wang, Y., and Ma, H. (2021). "Depression detection by analysing eye movements on emotional images," in *Proceedings of the ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, (Toronto, ON: IEEE), 7973–7977. doi: 10.1109/ICASSP39728.2021.9414663

Snyder, D., Garcia-Romero, D., Povey, D., and Khudanpur, S. (2017). "Deep neural network embeddings for text-independent speaker verification," in *Proceedings of the annual conference of the international speech communication association, INTERSPEECH*, Vol. 2017-Augus (Stockholm: International Speech Communication Association (ISCA)), 999–1003. doi: 10.21437/Interspeech.2017-620

Snyder, D., Garcia-Romero, D., Sell, G., McCree, A., Povey, D., and Khudanpur, S. (2019). "Speaker recognition for multi-speaker conversations using x-vectors," in *ICASSP 2019-2019 IEEE International conference on acoustics, speech and signal processing (ICASSP)* (Brighton: IEEE), 5796–5800. doi: 10.1109/ICASSP.2019.8683760

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International conference on acoustics, speech and signal processing (ICASSP)* (New York, NY: IEEE), 5329–5333.

Spijker, J., De Graaf, R., Bijl, R. V., Beekman, A. T. F., Ormel, J., and Nolen, W. A. (2004). Functional disability and depression in the general population. Results from the Netherlands mental health survey and incidence study (NEMESIS). *Acta Psychiatr. Scand.* 110, 208–214. doi: 10.1111/j.1600-0447.2004.00335.x

Szabadi, E., Bradshaw, C. M., and Besson, J. A. O. (1976). Elongation of pause-time in speech: A simple, objective measure of motor retardation in depression. *Br. J. Psychiatry* 129, 592–597. doi: 10.1192/bjp.129.6.592

Tasnim, M., and Stroulia, E. (2019). "Detecting depression from voice," in *Proceedings of the Canadian conference on artificial intelligence*, (Kingston, ON: Springer), 472–478. doi: 10.1007/978-3-030-18305-9_47

Tasnim, M., Ehghaghi, M., Diep, B., and Novikova, J. (2022). "Depac: A corpus for depression and anxiety detection from speech," in *Proceedings of the eighth workshop on computational linguistics and clinical psychology*, Seattle, WA, 1–16. doi: 10.18653/v1/2022.clpsych-1.1

Tong, F., Zhao, M., Zhou, J., Lu, H., Li, Z., Li, L., et al. (2021). "ASV-subtools: Open source toolkit for automatic speaker verification," in *Proceedings of the ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, (Toronto, ON: IEEE), 6184–6188. doi: 10.1109/ICASSP39728.2021.9414676

Valstar, M., Schuller, B. W., Krajewski, J., Cowie, R., and Pantic, M. (2014). "AVEC 2014: The 4th international audio/visual emotion challenge and workshop," in *Proceedings of the 22nd ACM international conference on multimedia*, Brisbane, QLD, 1243–1244. doi: 10.1145/2647868.2647869

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., et al. (2013). "Avec 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, Barcelona, 3–10. doi: 10.1145/2512530.2512533

Variani, E., Lei, X., McDermott, E., Moreno, I. L., and Gonzalez-Dominguez, J. (2014). "Deep neural networks for small footprint text-dependent speaker verification," in *Proceedings of the 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, (Florence: IEEE), 4052–4056. doi: 10.1109/ICASSP.2014.6854363

Vekkot, S., Gupta, D., Zakariah, M., and Alotaibi, Y. A. (2019). Hybrid framework for speaker-independent emotion conversion using i-vector PLDA and neural network. *IEEE Access* 7, 81883–81902. doi: 10.1109/ACCESS.2019.2923003

Villalba, J., Chen, N., Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., et al. (2020). State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and speakers in the wild evaluations. *Comput. Speech Lang.* 60:101026. doi: 10.1016/j.csl.2019.101026

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. *IEEE Trans. Acoust. Speech Signal Process.* 37, 328–339. doi: 10.1109/29.21701

Wang, H., Liu, Y., Zhen, X., and Tu, X. (2021). Depression speech recognition with a three-dimensional convolutional network. *Front. Hum. Neurosci.* 15:713823. doi: 10.3389/fnhum.2021.713823

Wang, S., Qian, Y., and Yu, K. (2017). "What does the speaker embedding encode?," in *Proceedings of the annual conference of the international speech communication association*, Stockholm, 1497–1501. doi: 10.21437/Interspeech.2017-1125

World Health Organization (2017). *Depression and other common mental disorders: Global health estimates*. Geneva: World Health Organization.

World Health Organization (2022). *Mental health and COVID-19: Early evidence of the pandemic's impact*. Geneva: World Health Organization.

Xing, Y., Liu, Z., Li, G., Ding, Z., and Hu, B. (2022). 2-level hierarchical depression recognition method based on task-stimulated and integrated speech features. *Biomed. Signal Process. Control* 72:103287. doi: 10.1016/j.bspc.2021.103287

Zhang, P., Wu, M., Dinkel, H., and Yu, K. (2021). "Depa: Self-supervised audio embedding for depression detection," in *Proceedings of the 29th ACM international conference on multimedia*, New York, NY, 135–143. doi: 10.1145/3474085.3479236

Zhou, Z. (2021). "Ensemble learning," in *Machine learning*, (Berlin: Springer), 181–210. doi: 10.1007/978-981-15-1967-3_8

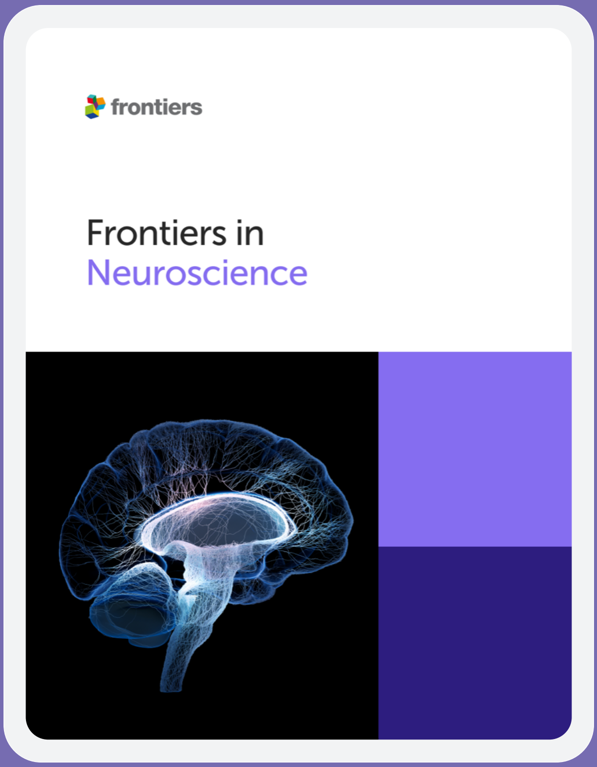
Frontiers in Neuroscience

Provides a holistic understanding of brain function from genes to behavior

Part of the most cited neuroscience journal series which explores the brain - from the new eras of causation and anatomical neurosciences to neuroeconomics and neuroenergetics.

Discover the latest Research Topics

See more →



frontiers

Frontiers in Neuroscience

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact