International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-13 Issue-10, September 2024

Multi-Modal Emotion Recognition Feature Extraction and Data Fusion Methods Evaluation

Sanjeeva Rao Sanku, B. Sandhya

Abstract: ***Research into emotion detection is crucial because of the wide range of fields that can benefit from it, including healthcare, intelligent customer service, and education. In comparison to unimodal approaches, multimodal emotion recognition (MER) integrates many modalities including text, facial expressions, and voice to provide better accuracy and robustness. This article provides a historical and present-day overview of MER, focusing on its relevance, difficulties, and approaches. We examine several datasets, comparing and contrasting their features and shortcomings; they include IEMOCAP and MELD. Recent developments in deep learning approaches, particularly fusion strategies such as early, late, and hybrid fusion are covered in the literature review. Data redundancy, complicated feature extraction, and real-time detection are among the identified shortcomings. Our suggested technique enhances emotion recognition accuracy by using deep learning to extract features using a hybrid fusion approach. To overcome existing restrictions and advance the area of MER, this study intends to direct future investigations in the right direction. Examining various data fusion strategies, reviewing new methodologies in multimodal emotion identification, and identifying problems and research needs to make up the primary body of this work.***

Keywords: Multimodal Emotion Recognition (MER), Speech Analysis, Facial Expression Recognition, MELD, Hybrid Fusion

I. INTRODUCTION

Research in the field of emotion identification is crucial because it allows computers to understand human emotions and respond intelligently to human needs. In the classroom, students' emotional states may have a major impact on their productivity and health. Teachers may get a better picture of their student's emotional health and academic performance with the use of emotion detection technology that tracks their emotional states. Because technology might help physicians comprehend their patients' emotional states, emotion recognition could be useful in healthcare [1]. This would allow for more personalized and adapted medical treatment.

By precisely identifying customers' emotional demands and providing tailored services, emotion identification might improve intelligent customer service systems. With the ability to completely transform the way humans interact with computers, emotion detection systems have quickly become a prominent focus of AI research. In the early stages of emotion recognition research, researchers focused largely on recognizing individual modalities such as voice emotion recognition, text emotion recognition, and facial expression identification [2]. Nevertheless, the restricted precision of emotional assessments based on a single mode is due to inadequate data and vulnerability to interference. As a result, researchers have increasingly turned to using numerous modalities to enhance the accuracy of emotional evaluations. Therefore, in response to this, researchers created multimodal emotion recognition (MER) [3]. In addition, MER may quantify the relationship between distinct aspects of many modalities by using the idea of maximum mutual information [4]. The most useful and distinguishing features of each modality may therefore be extracted in this way. This approach greatly improves the model's ability to differentiate between emotions. Thus, MER has attracted a lot of interest from scholars who are interested in combining data from several modalities. Emotional judgments may be more complete and accurate when these components complement and support one another. Emotional judgments are therefore much improved [5] [6][7]. One way to classify emotions is as neutral or non-neutral. A lack of strong feelings or expressions in reaction to different circumstances is often associated with neutral emotions. Two main types of emotions are not neutral: positive and negative. Negative traits including anxiety, failure, and pessimism have been linked to unpleasant feelings. A person's health can take a hit if they often deal with these unpleasant emotions. It was also shown to correlate directly with the duration that people can focus on a single task. An individual's cardiovascular system may become dysfunctional as a result of these negative feelings. Positive emotions, on the other hand, are seen very differently. Optimal well-being is associated with positive feelings, such as happiness and pleasure [8]. Figure 1 depicts the taxonomy of emotions that was previously discussed.

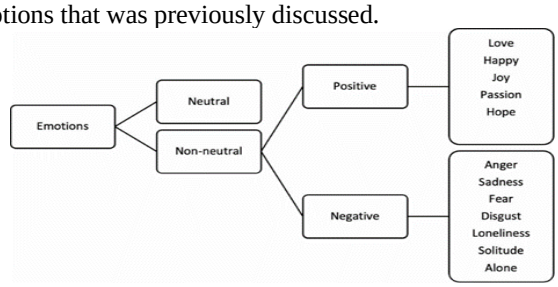

Fig.1: The Categorization of Various Emotions for Multimodal Emotion Identification

Manuscript received on 10 August 2024 | Revised Manuscript received on 20 August 2024 | Manuscript Accepted on 15 September 2024 | Manuscript published on 30 September 2024.
*****Correspondence Author(s)
Sanjeeva Rao Sanku*, Department of Computer Science and Engineering, University College of Engineering, Osmania University, Hyderabad (Telangana), India. Email ID: ssanjeevarao@gmail.com, ORCID ID: 0009-0003-5966-4629
Prof. B. Sandhya, Department of Computer Science and Engineering, MVSR Engineering College, Hyderabad (Telangana), India, Email ID: sandhya_cse@mvsrec.edu.in

A. Components of Multimodal Emotion Recognition

- Speech: Captures vocal expressions of emotion.
- Text: Analyzes written language for emotional content.
- Facial Expressions: Identifies emotions through facial movements.
- Physiological Signals: Includes EEG, heart rate, and skin conductance.

B. Datasets

To train and evaluate emotion identification algorithms, datasets that incorporate multimodal data are essential. Examples include IEMOCAP and the MELD (Multimodal EmotionLines Dataset).

C. Iemocap

An important resource for emotion detection research is the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [9] dataset, which was established by the Signal Analysis and Interpretation Lab (SAIL) at the University of Southern California. A comprehensive investigation of emotional states in interactive scenarios is made possible by the dataset's amalgamation of multiple modalities, including video, voice, facial motion capture, and text data. Ten performers, evenly split between men and women, helped compile the statistics. The performers, who were partnered according to gender and split into five groups, delivered both prepared and spontaneous lines. The diversity of emotional content in this collection of chats makes it a better representation of emotional communication in the actual world. IEMOCAP offers a diverse range of emotional situations for examination, including 4784 spontaneous and 5255 scripted encounters. The conversations include a wide range of emotions, from joy to wrath to surprise to fear to disgust to irritation to excitement, and even neutrality. This allows for a more nuanced examination of emotions by including continuous aspects such as activation, arousal, and dominance. The main advantage of IEMOCAP is that it is authentic; deep learning models may use it to identify real emotional signals across different modalities since the emotions it captures are real and not fake. But IEMOCAP does have certain restrictions. Although the dataset has a wealth of modalities and emotional categories, its relatively modest size may make it unsuitable for deep learning models that need more data. The model's adaptability to many cultures and languages is diminished by its linguistic limitation to English. Because it is only available in English, the model is less flexible and less suited to use with other cultures and languages. The fact that the data was obtained in a controlled laboratory setting could impact the authenticity of the sensations, which might impede the model's capacity to generalize in real-world circumstances. Additionally, the sample has a significant class imbalance, which might lead to erroneous results when it comes to less prevalent emotions.

D. Meld

The Multimodal Emotion Lines Dataset (MELD) [10] provides a fresh viewpoint in the field of MER by concentrating on the emotional complexities inherent in dialogues involving several participants. The dataset is a structured compilation of text from the hit American TV show "Friends," consisting of 1433 conversations with a total of 13,708 words. The MELD dataset offers a wide range of emotions for study, with each phrase labeled with one of seven categories: anger, contempt, sorrow, joy, neutrality, surprise, or fear. Along with these specific designations, every remark also has an emotional categorization of good, negative, or neutral, which helps to comprehend the attitude more broadly. Contextual models for dialogue-based emotion identification may be built and improved with the help of this dataset, whose design is in line with its main goal. An important but so far unexplored facet of human communication, the emotional dynamics within multi-party debate settings, may be uncovered with the use of this dataset. When using the dataset for study or model building, it is important to examine its cultural applicability and realism. The dataset's authenticity is compromised since the conversations are derived from a fictional television series. Consequently, they could fail to capture the nuances of spontaneous, unplanned conversations. Furthermore, the television series is American-made, therefore the conversations mostly reflect American cultural standards and idioms. Understanding cultural variety is crucial in emotion recognition research, since models trained on this dataset may not work as well in other cultural settings. Despite these caveats, the MELD dataset is nevertheless a great resource for research on emotion detection in natural speech. It fills a need in the literature by concentrating on conversations with several participants and may help shape future emotion detection algorithms that are both more complex and sensitive to context.

II. MULTIMODAL EMOTION RECOGNITION

To build multimodal emotion identification frameworks, researchers are investigating several methods. The identification of the facial position and extraction of mathematical selections, visual selections, or a combination of mathematical and visual alternatives on the target face are common features of these systems. The available alternatives are sometimes segregated from the facial region location or entirely extraordinary facial locations including several types of information. Prior research mostly focuses on integrating audiovisual data to facilitate automated emotion identification, such as merging speech and facial expression. Their corpus showed that feature-level fusion was the best method for distinguishing between angry and neutral expressions. Distinguishing between joyful and sad emotions, however, was best accomplished via decision-level fusion. The researchers concluded that the optimal fusion technique is contingent upon the specific application. The multimodal identification system incorporates not only voice and facial emotion but also the thermal distribution of infrared pictures. In decision-level fusion, each modality is trained independently using several unimodal classifiers, and their outputs are combined using predefined weighting algorithms. Many methods for fusing models at the level of individual models have been proposed. Due to its possible applications in healthcare, social robotics, and human-computer interaction, multimodal emotion recognition (MER) has attracted a lot of attention in the last few years.

Emotion recognition systems are made more accurate and reliable by combining several modalities including text, audio, and facial expressions. Focusing on fusion methods, datasets, applications, and the difficulties of model interpretability, this literature review examines the most recent developments in deep learning-based MER approaches. Jiménez-Guarneros and Fuentes-Pineda (2024) [29] introduced CFDA-CSF, a method that integrates cross-subject domain adaptation to address variability in emotion recognition across different individuals. Their approach combines facial and speech features using domain adaptation techniques to enhance generalization across subjects. This multi-modal method improves emotion recognition accuracy by mitigating the effects of individual differences. Sun et al. (2024) [30] proposed a hierarchical knowledge distillation approach for multi-modal emotion recognition. By leveraging a teacher-student framework, their model distills knowledge from a complex, multi-modal teacher network to a simpler student network. This method ensures that the student model retains critical emotional features while being more computationally efficient. Alsaadawi and Daş (2024) [31] utilized a bi-directional LSTM graph convolutional network (Bi-LG-GCN) for the MELD dataset, enhancing emotion recognition by capturing contextual and sequential dependencies in multi-modal data. Their approach demonstrates the potential of graph-based models in handling complex interactions between different modalities. Kumar et al. (2024) [32] explored hybrid fusion methods combining speech and image data to create an interpretable multi-modal emotion recognition system. Their research highlights the need to combine data from many modalities to improve accuracy and interpretability. Umair et al. (2024) [33] developed Emo Fu-Sense, which integrates various multi-modal data streams through a novel fusion technique. Their method emphasizes the coherent integration of emotional signals, leading to a more holistic understanding of the user's emotional state. Makhmudov et al. (2024) [34] improved the handling of multi-modal data by using attention techniques inside BERT and CNN systems. They show how attention processes may improve emotion identification by zeroing down on the most important details in the input data. Wang et al. (2024) [35] introduced Husformer, a multi-modal transformer model designed for human state recognition. This model capitalizes on the transformer's ability to handle sequential data, providing robust multi-modal fusion and state recognition capabilities. Pereira et al. (2024) [36][54][55][56] compiled a thorough synopsis of recent developments and current practices in the area of emotion detection by conducting a systematic analysis of technologies using computer vision and deep learning. Their review highlights the importance of integrating various data sources and the challenges associated with multi-modal emotion recognition. Li et al. (2024) [37] unveiled Magdra, a network for multi-modal attention that uses algorithms for dynamic routing by agreement to control the merging of different modalities. By constantly adjusting to the input data, this model achieves a good balance between the modalities, leading to better recognition performance. To achieve multi-modal emotion identification, Wang et al. (2023) [38] integrate EEG inputs with facial expressions. Their method leverages the complementary nature of physiological and visual data, integrating EEG signals with facial expressions to enhance the emotional context. Better recognition results are achieved by combining EEG and face data, which gives a more complete picture of the underlying emotional states. Lei and Cao (2023) [39] utilized preference learning for audio-visual emotion recognition. Their model incorporates both intended and perceived emotion labels to refine the learning process, aligning the predictions more closely with human perception. Liu et al. (2023) [40] proposed a multi-modal fusion network that focuses on complementarity and the relative importance of each modality. Their model adapts to varying strengths of different modalities, ensuring that the fusion process effectively captures the most salient emotional cues. Hou et al. (2023) [41] proposed a Semantic Alignment Network (SAN) that aligns multi-modal data by learning a shared semantic space. This alignment enhances the integration of modalities by ensuring that the features from different sources contribute coherently to emotion recognition. Shahzad et al. (2023) [42] modified the Xception model to fuse multi-modal CNN features for emotion recognition. Their approach integrates CNN features across modalities, demonstrating significant improvements in accuracy by capturing detailed spatial and temporal patterns. Zhang et al. (2023) [43] developed a Structure-Aware Multi-Graph Network with a primary emphasis on conversational multi-modal emotion identification. Their approach leverages structural information from multi-graph networks, which capture the relational dynamics between different modalities. Zhang et al. (2023) [44] created the M3GAT, an Interactive Graph Attention Network that can handle several tasks and modalities. Through the use of graph attention mechanisms, this model dynamically prioritizes various modalities, resulting in exceptional performance in conversational sentiment analysis and emotion identification. Singh et al. (2022) [45] developed Emoint-trans, a multimodal transformer for identifying social conversational emotions and intentions. Transformers handle sequential data effectively, allowing for a nuanced analysis of conversational context. Zou et al. (2022) [46] improved multimodal fusion with the Main Modal Transformer, which prioritizes the dominant modality in fusion processes for conversation-based emotion recognition. Lian et al. (2022) [47] suggested SMIN—a Semi-supervised Multi-modal Interaction Network—to tackle the problem of lacking labeled data for conversational emotion identification. Their network uses both labeled and unlabeled data to improve model performance. Yoon (2022) [48] investigated cross-modal translation, a method for improving the accuracy and resilience of emotion detection models by combining input from other datasets. Wang et al. (2022) [49] investigated the integration of EEG and speech signals for emotion recognition. Their study highlights the complementary nature of physiological and speech data, providing a comprehensive view of the emotional state. Zheng et al. (2022) [50] introduced a Multi-channel Weight-Sharing Autoencoder with Cascade Multi-Head Attention for emotion recognition. This model effectively fuses features by sharing weights across channels and employing attention mechanisms to prioritize salient features.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Yang et al. (2022) [51] emphasized methods for multi-modal interaction with context and different modes of speech to identify emotions. Their approach enhances the system's emotional understanding and prediction capabilities by capturing the contextual interactions of several modalities. Liu et al. (2021) [52] compared the performance and robustness of various multimodal deep-learning models for emotion recognition. Their comparative study provides valuable insights into how different models handle the complexities of multi-modal data. Guanghui and Xiaoping (2021) [53] fused correlation features of speech and visual data. Their method effectively captures the synchronous and asynchronous interactions between modalities, enhancing emotion recognition accuracy. Table I and Table II provide a thorough overview of the present status and prospects in the domain of multimodal emotion recognition.

A. Identified Gaps

- Data Redundancy and Conflict
- To deal with contradictory or duplicated input in visual and auditory modalities, effective fusion algorithms are needed.

B. Feature Extraction Complexity

- Extracting features from audio and physiological signals remains complex and challenging.

C. Missing Data

- Handling missing data in one or more modalities is crucial to maintaining robustness in emotion recognition.

D. Real-time Detection

- Developing systems capable of real-time emotion detection is necessary for practical applications.

III. DATA FUSION METHODS

When it comes to multimodal emotion identification, data fusion approaches are crucial. By combining data from several modalities, these strategies make emotion detection systems more accurate and resilient. It is possible to broadly classify the methods as either early fusion, late fusion, or hybrid fusion.

A. Early Fusion

Early fusion refers to the process of merging raw data or characteristics from disparate modalities at an early stage, before inputting them into a unified model. This approach leverages the correlation between different modalities from the beginning of the learning process in Fig 2. Although early fusion has been crucial in the field of MER, researchers need to tackle its limits to enhance the optimization of this strategy. This continuous inquiry presents an intriguing subject of study in the ongoing development of MER. In this approach, features extracted from speech, text, and facial expressions are concatenated into a single feature vector, which is then inputted into a unified model for emotion prediction.

B. Late Fusion

Late fusion combines the outputs of individual models trained on different modalities. Each modality is processed independently, and their predictions are fused at a later stage, often using techniques like weighted averaging or voting. On the other hand, late fusion presupposes that each modality works autonomously, which fails to take into account the interdependence of modalities and may lead to imprecise result forecasts. To improve the accuracy of emotion detection (Fig. 3), future studies should focus on creating fusion methods that keep the benefits of late fusion but also include inter-modality interactions. During late fusion, individual models are used to analyze each modality (speech, text, face), and the resulting outputs are merged using a fusion procedure to provide the ultimate prediction.

C. Hybrid Fusion

Hybrid fusion has elements from both early and late fusion. It integrates characteristics at several stages of the processing pipeline, enabling interactions across different modalities at different levels of abstraction. In hybrid fusion, intermediate features from individual modality-specific models are concatenated and then fed into a unified model, which processes these fused features to make the final prediction. The intricacy of these approaches also poses difficulties, most notably in dealing with the increasing processing demands and in finding the optimal mix of early and late fusion algorithms. So, to progress in the area of multimodal effect detection, there has to be continuous study and improvement in hybrid fusion approaches (Fig. 4). These diagrams and descriptions highlight the different stages at which data fusion can occur, demonstrating the versatility and potential of each approach in multimodal emotion recognition systems.

Table- I: Performance of Feature Extraction and Fusion Methods on Most Widely Used Data Sets

S.NO, Author	Modalities Used and Feature Extraction Methods	Fusion policy adopted	Classifier used	Dataset	Performance
1. Wei-Long et al. [11]	EEG, Eye movements, STFT,	Feature level fusion EEG RBM	PCA, Deep Neural Network	EEG data, Eye movement data	Classified 4 emotions happy, sad, fear, and neutral with 72.39% accuracy.
2. Shahla et al. [12]	Audio, visual, text, Fisher vector encoding,	Dempster Shafer, MFA, CFA, CCA	SVM, Naive Bayes	DEAP	Classified 3 emotions happy, sad and fear

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

3.Haiing Huang et al [13]	Audio, video, CNN	GAP	ECNN	DEAP	Classified 4 emotions relaxation, depression, excitement, and fear with 82.92 % accuracy.
4. Hongli Zhang et. Al [14]	Audio, video, BDAE,	Bimodal deep automatic encoder	LIBSVM	Radboud faces database	Classified 4 emotions happy, sad, fear and neutral with 85.71 accuracy.
5.Shamane Siriwardhana et. Al [15]	Text, audio, vision, SSL	Transformers(SSE-FT), Attention(IMA)	Deep Learning	IEMOCAP, MELD, CMU-MOSI, CMU-MOSEI	Classified 5 emotions happy, sad, angry, neutral, excitement. Analyzed sentiment also from conversations.
6. Jinming Zhao et. Al [16]	Text, visual, audio, BERT	Transformers	Self-supervised learning	IEMOCAP, MSP-IMPROV	Classified 4 emotions happy, sad, angry and neutral.
7.Sarala Padi et. Al [17]	Speech, text, BERT, ResNet	Weighted Average(WA)	Residual Networks, Bidirectional Encoder Representations from Transformers	IEMOCAP	Classified 4 emotions angry, happy, neutral, and sad with 70.33% accuracy.
8.Puneet Kumar et. al. [18]	Speech, images, VGG	Gradient Descent	Deep Neural Networks	IIT Roorkee Speech Image Emotion Recognition dataset, IEMOCAP	Classified 4 emotions anger, happy, hate, and sad with 83.29% accuracy.
9.Yucel Cimtay et. Al [19]	Image, GSR, voice, Inception ResnetV2	Hybrid fusion weighted sun	CNN	Cohn Kanade, Faces DB, Radboud, AffectNet, LUMED-2, DEAP	Classified emotions angry, disgust, afraid, happy, neutral, sad and surprised with 74.2% accuracy.
10.Fengmao Lv et. Al [20]	Text, image, attention	Transformers	Unsupervised classifier	CMU-MOSI, CMU-MOSEI, IEMOCAP	Happy, sad, angry, neutral
11.Jiahui Pan et. Al [21]	Image, audio, GhostNet, LFCNN ,tLSTM	Optimal weight distribution	Deep learning	CK++, CMO-DB, MAHNOB-HCI	Classified emotions Anger, disgust, fear, happy, sad, surprise, contempt with 94.36% accuracy.
12.Sanghyun Lee et. Al [22]	Text, audio, visual, SoundNet, VGGish, BERT	SMAF, MAF, and VF transformers	Deep learning	CMU-MOSI, CMU-MOSEI, IEMOCAP	Classified emotions as happy, angry, sad, and neutral with 85.3% accuracy.
13.Dung Nguyen et. Al [23]	Audio, visual, 2DConv-AE, 1DConv-AE	LSTM	Deep Neural Network	RECOLA	Classified emotions happy, angry, excited, depressed, and bored.
14.Yi Yang et. Al [24]	Image, audio, ICA	RBM	Deep Belief Networks, SVM	SEED	Classified 3 emotions Positive, neutral, and negative.
15.Lucas Goncalves et. Al [25]	Audio, visual, VGG	Dot product attention, MHA	Neural Networks	CREMA-D corpus, MSP-IMPROV corpus	Gained 77.2% classification accuracy
16.Ke Zhang et. Al [26]	Text, audio, visual, DDQN, RL	RNN, GRU	RNN	IEMOCAP, MELD	Classified Happy, sad, neutral, angry, excited and frustrated emotions
17.Norbert Braunsch et. Al [27]	Text, audio, BERT, Encoders	LUDWING	Deep learning	IEMOCAP	Classified 4 emotions angry, happy, sad and neutral.
18.Guan-Nan Dong et. Al [28]	Text, audio, CNN_Bi-LSTM, word2vec	AFFM	CNN	IEMOCAP, MELD	Classified 6 emotions anger, happiness, excitement, sadness, frustration, and neutral

Table- II: Key Findings Using Various Feature Extraction and Fusion Methods

S.NO	Author	Methods Used for Feature Extraction	Methods used for Data Fusion	Datasets	Key Findings
1 [29]	Jiménez-Guarneros & Fuentes-Pineda (2024)	SMI BeGaze for eye movement, STFT for EEG features	t-SNE visualization in 2D	SEED, SEED-IV, SEED-V	Improved cross-subject emotion recognition performance
2 [30]	Sun et al. (2024)	VGG, ALBERT, Pretrained MTCNN	HKD- MER(Hierarchical Knowledge Distillation MER)	MOSET, IEMOCAP	Hierarchical knowledge distillation improves multimodal emotion recognition
3 [31]	Alsaadawı& Daş (2024)	K-PCA(Kernel PCA)	Bi-LG-GCN	MELD	Enhanced emotion recognition on MELD dataset using Bi-LG-GCN
4 [32]	Kumar et al. (2024)	Deep Learning based classifiers	ParallelNet, VGG, ResNet	IIT-R SIER Custom dataset	Interpretable emotion recognition through the hybrid fusion of speech and image

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

5 [33]	Umair et al. (2024)	LSTM, CNN, and RNN	Emotion Fusion-Sense (Emo Fu-Sense)	Various datasets	Novel multimodal emotion classification technique
6 [34]	Makhmudov et al. (2024)	Mel Spectogram Features, BERT, Bi-GRU	Attention Mechanisms in BERT and CNN	MELD, CMU-MOSEI	Enhanced emotion recognition through attention mechanisms in BERT and CNN
7 [35]	Wang et al. (2024)	Husformer	Cross-modal Attention Transformers- Husformer	DEAP, WESAD, MOCAS and CogLoad	Multi-modal transformer for human state recognition
8 [36]	Pereira et al. (2024)	Computer Vision and Deep Learning	LSTM	Various computer vision datasets	A systematic review of emotion detection methods
9 [37]	Li et al. (2024)	Multi-modal Attention Graph Network (Magdra)	PAA, ECT	IEMOCAP, CMU-MOSI	Dynamic routing-by-agreement improves multi-label emotion recognition
10 [38]	Wang et al. (2023)	Fusion of EEG signals and facial expressions, CNN, attention, DeepVANet	Weight assignment, AdaBoost	DEAP, MAHNOB-HCI	Multimodal emotion recognition from EEG and facial expressions
11 [39]	Lei & Cao (2023)	OpenSMILE, OFFBA	RankSVM, RankNet, Labmda MART	CREMA-D	Audio-visual emotion recognition with a preference learning
12 [40]	Liu et al. (2023)	Attention, LSTM	Multimodal fusion networks	MELD, IEMOCAP	Network with complementarity and importance enhances emotion recognition
13 [41]	Hou et al. (2023)	Encoders, SMI, attention	SAMS	MELD, IEMOCAP	Semantic alignment network improves emotion recognition
14 [42]	Shahzad, H. M., et al. (2023)	Convolution models	Multi-Modal CNN	M-LFW-F, CREMA-D	The modified Xception model improves emotion recognition
15 [43]	Zhang, Duzhen, et al. (2023)	RoBERTa, openSMILE, 3D-CNN	SAMGN	IEMOCAP, MELD	Efficient emotion recognition in conversations
16 [44]	Zhang, Yazhou, et al. (2023)	ResNet, BERT	Multi-task Interactive Graph Attention Network	MELD, DailyDialog	Effective for sentiment analysis and emotion recognition
17 [45]	Singh, Gopendra Vikram, et al. (2022)	BiLSTM, openSMILE, ResNeXt	MISA	EmoInt-MD	Identifies emotions and intents in social conversations
18 [46]	Zou, ShiHao, et al. (2022)	Main Modal Transformer	MMTr, CMF	MELD	Enhances multimodal fusion in conversations
19 [47]	Lian, Zheng, Bin Liu, and Jianhua Tao (2022)	Bi-GRU, Attention	Semi-supervised Multimodal Interaction Network(SMIN)	IEMOCAP, MELD, CMU-MOSI	Improved conversational emotion recognition
20 [48]	Yoon, Yeo Chan (2022)	FAN, fully connected layer, glove word embedding	Cross-Modal Translator	IEMOCAP, CMU-MOSEI	Utilizes multiple datasets for robust recognition
21 [49]	Wang, Qian, et al. (2022)	AFE, PLDA, ELM	AVER, WAVER, RF, ET	MED4, SEED, DEAP	Uses EEG and speech for emotion recognition
22 [50]	Zheng, Jiahao, et al. (2022)	Encoder, Attention	MCWSA-CMHA	IEMOCAP, MSP-IMPROV	Cascade multi-head attention enhances performance
23 [51]	Yang, Dingkang, et al. (2022)	Transformers	Contextual and Cross-Modal Interaction	IEMOCAP, RAVDESS	Focuses on speech emotion recognition
24 [52]	Liu, Wei, et al. (2021)	STFT	DCCA	SEED-V, DREAMER	Compares performance and robustness of different models
25 [53]	Guanghui, Chen, and Zeng Xiaoping (2021)	AlexNet, C3D-Sports-1M	Correlation Features Fusion	RML, eNTERFACE05 ,BAUM-1	Fuses speech and visual features for emotion recognition

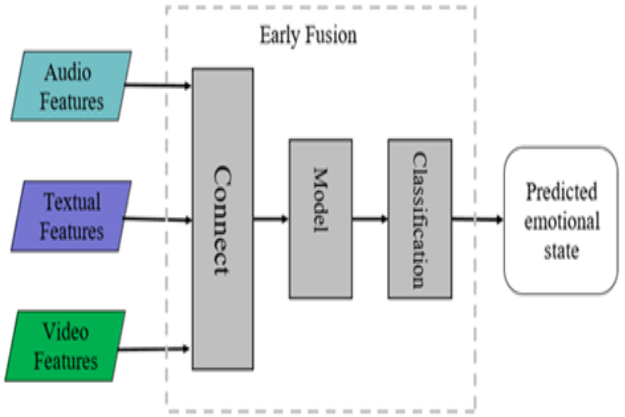

Fig. 2: Framework for Recognizing Emotions Using Several Modes, Based on Early Fusion Techniques

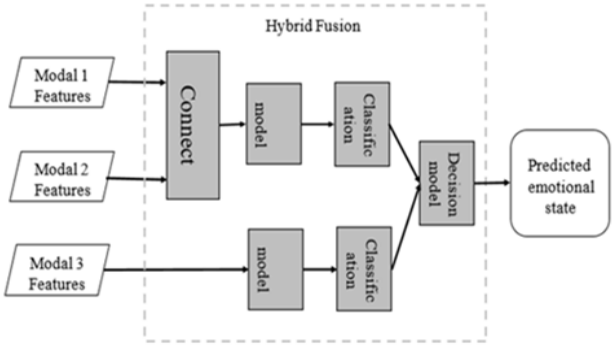

Fig.3: A Framework for Recognizing Emotions Using Several Modes, Based on the Concept of Late Fusions

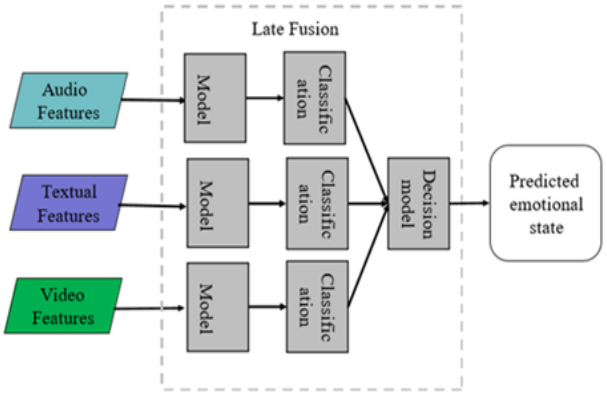

Fig. 4:. A Framework for Recognizing Emotions Using a Combination of Different Modes and Fusion Techniques

IV. EFFECTIVE DATA FUSION FOR MULTIMODAL EMOTION RECOGNITION CHALLENGES

To make emotion recognition systems more accurate and resilient, multimodal emotion identification tries to include additional data from different modalities, such as text, facial expressions, and speech. However, several challenges must be addressed to develop effective data fusion methodologies. These challenges include:

A. Heterogeneity of Data

Different modalities provide diverse types of data, including auditory signals from speech, textual data from transcripts, and visual data from facial expressions. The intrinsic properties of these data types (e.g., temporal vs. spatial, continuous vs. discrete) can make it difficult to combine them effectively.

Temporal Alignment: Speech and facial expressions evolve, requiring synchronization of features extracted from these modalities.

Feature Representation: Speech features (e.g., MFCCs) differ from textual features (e.g., word embeddings) and facial features (e.g., facial landmarks), necessitating a common representation framework.

Fig.4.A framework for recognizing emotions using a combination of different modes and fusion techniques.

B. Modality-Specific Noise and Artifacts

The emotion detection system's performance may be negatively impacted by various forms of noise and artifacts that impact each modality.

Speech: Background noise, microphone quality, and speaker variability can affect the quality of speech features.

Text: Text derived from speech through automatic speech recognition (ASR) may contain transcription errors, especially in noisy environments or with non-native speakers.

Facial Expressions: Variations in lighting, occlusions, and camera angles can affect the accuracy of facial feature extraction.

C. Missing or Incomplete Data

In practical applications, not all modalities may be available at all times due to sensor failures, occlusions, or user preferences.

Handling Missing Modalities: Effective fusion methods must account for situations where one or more modalities are missing and still make accurate predictions.

Imputation Techniques: Methods such as data imputation or estimation of missing features need to be robust and accurate to avoid introducing biases.

D. Computational Complexity

Combining multiple modalities can significantly increase the computational burden, making real-time processing challenging.

Limitations on Available Resources: Real-time multimodal processing may be taxing for devices with little computing power, such as mobile phones.

Algorithm Complexity: Complex fusion algorithms can lead to increased training and inference times, which may not be feasible in time-sensitive applications.

E. Optimal Fusion Strategies

Optimal selection of the fusion technique (early fusion, late fusion, or hybrid fusion) is of utmost importance and may differ based on the specific application and the data that is accessible.

Early Fusion: While it allows for the exploitation of correlations between modalities from the start, it may suffer from high dimensionality and require extensive pre-processing.

Late Fusion: This approach is simpler and more modular but might miss the interdependencies between modalities that could improve performance.

Hybrid Fusion: Combining the strengths of both early and late fusion, hybrid fusion can be complex to design and optimize.

F. Interpretability of Models

Interpreting deep learning models, particularly those used in multimodal fusion, might provide challenges, hence hindering comprehension of the specific contributions of distinct modalities to the ultimate prediction.

Transparency: The trustworthiness and accountability of the model depend on its decision-making process being open and understandable.

Explainability Techniques: Developing methods to explain the contribution of each modality and feature can help in debugging and improving the model.

G. Generalization Across Domains

Emotion recognition models trained on specific datasets may not generalize well to other domains or environments.

Domain Adaptation: Ensuring that the model can adapt to different speaking styles, languages, and cultural expressions of emotion is essential for real-world applications.

Transfer Learning: However, a substantial quantity of labeled data from the target domains may be necessary to fine-tune pre-trained models.

H. Evaluation Metrics and Benchmarking

Standardized evaluation metrics and benchmark datasets are needed to compare the performance of different fusion methodologies effectively.

Performance Metrics: Accuracy, F1-score, confusion matrices, and other metrics should be used to evaluate and compare models.

Benchmark Datasets: Comprehensive datasets that include synchronized multimodal data and cover a wide range of emotional expressions are essential for benchmarking.

V. METHODOLOGY

A. Semantic Diagram

The proposed methodology involves extracting features from text, audio, and video modalities and then using a hybrid fusion approach to combine these features. A robust classification model will be developed to accurately identify emotions. There is no text provided. Fig. 5. depicts the all-encompassing structure of a standard deep learning-powered MER (Machine Emotion Recognition) system. Emotion recognition (MER) is now a hot subject in artificial intelligence (AI) because of the fast development of deep learning methods, which have played a major role in its progress.

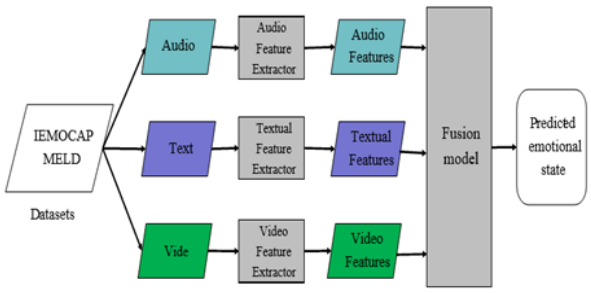

Fig.5: An Outline of a Standard Multimodal Emotion Identification System

The essence of this lies in the creation of accurate network architectures and their accompanying measures of error. Many modern loss functions are inspired by concepts related to entropy, and one noteworthy example is the cross-entropy loss function. Essentially, several deep learning architectures enhance their models by optimizing certain loss functions associated with entropy, either by maximizing or decreasing them. The flexibility of deep learning algorithms allows for their precise customization to leverage the interactions between many senses, resulting in the extraction of emotionally rich characteristics for accurate emotion identification. Researchers are seeing incredible results when they use deep learning algorithms to MER. Firstly, we present commonly used MER datasets and perform a comprehensive study of their intrinsic properties and possible problems. In addition, we analyze the advantages and disadvantages of various datasets, assisting researchers in choosing the most appropriate datasets for their particular investigations. Furthermore, we thoroughly examine several strategies for extracting emotional features from many modes of communication, with a particular focus on highlighting the advantages and limitations of each technique. Finally, we provide a comprehensive review of MER algorithms, focusing on fusion methods for early, late, hybrid, and intermediate layers. We examine the pros and cons of various fusion methods, providing useful perspectives to aid researchers in choosing the most suitable fusion approaches. We anticipate that our study will address the existing deficiency in the area and function as a beneficial resource for academics seeking to get a thorough comprehension of the latest accomplishments and future research opportunities in the domain of MER.

B. Implementation Steps

- Data Collection: Use datasets like IEMOCAP and DEAP.
- Feature Extraction: Employ NLP techniques for text, MFCCs for speech/audio, and FACS for facial/video expressions.
- Data Fusion: Efficient data fusion methods are essential to combine features from different modalities. There are three main types of fusion methods: early, late, and hybrid.
- Classification: Common classifiers used in emotion recognition include Convolutional Neural Networks (CNNs), Support Vector Machines (SVMs), and recurrent neural networks (RNNs).

VI. CHALLENGES AND FUTURE DIRECTIONS

An important step in comprehending and responding to human emotions in a variety of settings is multimodal emotion recognition. In comparison to single-modal techniques, MER systems provide more accurate and robust emotion recognition by combining voice, text, and facial expressions. The exploration of deep learning techniques and fusion strategies reveals the potential for substantial improvements in emotion recognition systems. However, challenges such as handling heterogeneous data, modality-specific noise, and real-time processing must be addressed.

To make MER systems more accurate and useful, we suggest a paradigm that combines enhanced feature extraction with hybrid fusion. Emotion recognition models should be generalized across domains and cultural settings, and future research should concentrate on enhancing model interpretability and creating more efficient fusion approaches. This comprehensive study lays the groundwork for further research to improve multimodal emotion recognition.

VII. CONCLUSION

In this paper, we presented various feature extraction methods of different modalities like audio, video, and text. Various feature fusion methods are presented with their evaluation. This is very useful for researchers who are working in multimodal data fusion and classification.

DECLARATION STATEMENT

After aggregating input from all authors, I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** This article has not been funded by any organizations or agencies. This independence ensures that the research is conducted with objectivity and without any external influence.
- **Ethical Approval and Consent to Participate:** The content of this article does not necessitate ethical approval or consent to participate with supporting documentation.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.

- **Authors Contributions:** The authorship of this article is attributed equally to all participating authors.

REFERENCES

1. Swain, M.; Routray, A.; Kabisatpathy, P. Databases, features and classifiers for speech emotion recognition: A review. Int. J. Speech Technol. 2018, 21, 93–120. https://doi.org/10.1007/s10772-018-9491-z
2. Zong, Y.; Lian, H.; Chang, H.; Lu, C.; Tang, C. Adapting Multiple Distributions for Bridging Emotions from Different Speech Corpora. Entropy 2022, 24, 1250. https://doi.org/10.3390/e24091250
3. Li, S.; Deng, W. Deep facial expression recognition: A survey. IEEE Trans. Affect. Comput. 2020, 13, 1195–1215. https://doi.org/10.1109/TAFFC.2020.2981446
4. Yang, H.; Xie, L.; Pan, H.; Li, C.; Wang, Z.; Zhong, J. Multimodal Attention Dynamic Fusion Network for Facial Micro-Expression Recognition. Entropy 2023, 25, 1246. https://doi.org/10.3390/e25091246
5. Zeng, J.; Liu, T.; Zhou, J. Tag-assisted Multimodal Sentiment Analysis under Uncertain Missing Modalities. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, 11–15 July 2022; pp. 1545–1554. https://doi.org/10.1145/3477495.3532064
6. Shou, Y.; Meng, T.; Ai, W.; Yang, S.; Li, K. Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis. Neurocomputing 2022, 501, 629–639. https://doi.org/10.1016/j.neucom.2022.06.072
7. Li, Y.; Wang, Y.; Cui, Z. Decoupled Multimodal Distilling for Emotion Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 6631–6640. https://doi.org/10.1109/CVPR52729.2023.00641
8. Shirke, B., Wong, J., Libut, J. C., George, K., & Oh, S. J. (2020). Brain-iot-based emotion recognition system. In Proceedings of the 10th annual Computing and Communication Workshop and Conference (CCWC) (pp. 0991–0995). IEEE. https://doi.org/10.1109/CCWC47524.2020.9031124
9. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. Lang. Resour. Eval. 2008, 42, 335–359. https://doi.org/10.1007/s10579-008-9076-6
10. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 527–536. https://doi.org/10.18653/v1/P19-1050
11. Wei-Long Zheng , Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki."EmotionMeter: A Multimodal Framework for Recognizing Human Emotions" Volume: 49, Issue: 3, March 2019, Pages: 1110 – 1122, February 2018, DOI:10.1109/TCYB.2018.2797176. https://doi.org/10.1109/TCYB.2018.2797176
12. SHAHLA NEMATI ,REZA ROHANI , MOHAMMAD EHSAN BASIRI , MOLOUD ABDAR , NEIL Y. YEN , AND VLADIMIR MAKARENKOV. "A Hybrid Latent Space Data Fusion Method for Multimodal Emotion Recognition, IEEE Access Volume: 7, Pages: 172948 – 172964, ISSN:2169-3536, November 2019, DOI:10.1109/ACCESS.2019.2955637. https://doi.org/10.1109/ACCESS.2019.2955637
13. HAIPING HUANG, ZHENCHAO HU, WENMING WANG, AND MIN WU"Multimodal Emotion Recognition Based on Ensemble Convolutional Neural Network", IEEE Access Volume: 8, Pages:3265 – 3271, December 2019, DOI:10.1109/ACCESS.2019.2962085. https://doi.org/10.1109/ACCESS.2019.2962085
14. HONGLI ZHANG "Expression-EEG Based Collaborative Multimodal Emotion Recognition Using Deep AutoEncoder", IEEE Access Volume: 8, Pages: 164130 – 164143, ISSN: 2169-3536, September 2020, DOI:10.1109/ACCESS.2020.3021994. https://doi.org/10.1109/ACCESS.2020.3021994
15. SHAMANE SIRIWARDHANA , THARINDU KALUARACHCHI , MARK BILLINGHURST , AND SURANGA NANAYAKKARA. "Multimodal Emotion Recognition With Transformer-Based Self Supervised Feature Fusion", IEEE Access Volume: 8, Pages: 176274 – 176285, ISSN: 2169-3536, September 2020, DOI:10.1109/ACCESS.2020.3026823, https://doi.org/10.1109/ACCESS.2020.3026823
16. Jinming Zhao, Ruichen Li, Qin Jin, Xinchao Wang, Haizhou Li, "MEMOBERT: PRE-TRAINING MODEL WITH PROMPT-BASED LEARNING FOR MULTIMODAL EMOTION RECOGNITION", 27oct 2021.
17. Sarala Padi, Seyed Omid Sadjadi, Dinesh Manocha, and Ram D. Sriram. "Multimodal Emotion Recognition using Transfer Learning from SpeakerRecognition and BERT-based models", 16 Feb 2022.
18. Puneet Kumar, Sarthak Malik, and Balasubramanian Raman, "Interpretable Multimodal Emotion Recognition using Hybrid Fusion of Speech and Image Data", Springer Nature 2023. https://doi.org/10.1007/s11042-023-16443-1
19. YUCEL CIMTAY, ERHAN EKMEKCIOGLU, AND SEYMA CAGLAR-OZHAN, "Cross-Subject Multimodal Emotion Recognition Based on Hybrid Fusion", September 14, 2020, DOI: 10.1109/ACCESS.2020.3023871 https://doi.org/10.1109/ACCESS.2020.3023871
20. Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, Guosheng Lin. "Progressive Modality Reinforcement for Human Multimodal EmotionRecognition from Unaligned Multimodal Sequences"
21. Jiahui Pan, Weijie Fang, Zhihang Zhang, Bingzhi Chen, Zheng Zhang, Shuihua Wang, "Multimodal Emotion Recognition based on Facial Expressions Speech, and EEG", DOI 10.1109/OJEMB.2023.3240280
22. SANGHYUN LEE, DAVID K. HAN, AND HANSEOK KO, "Multimodal Emotion Recognition Fusion Analysis Adapting BERT With Heterogeneous Feature Unification", June 2021, Digital Object Identifier 10.1109/ACCESS.2021.3092735.
23. Dung Nguyen, Duc Thanh Nguyen , Rui Zeng, Thanh Thi Nguyen , Son N. Tran, Thin Nguyen, Sridha Sridharan, "Deep Auto-Encoders With Sequential Learning for Multimodal Dimensional Emotion Recognition" IEEE, VOL. 24, 2020, pages:1313-1324. https://doi.org/10.1109/TMM.2021.3063612
24. Yi Yang, Qiang Gao, Yu Song, Xiaolin Song, Zemin Mao, and Junjie Liu, "Investigating of Deaf Emotion Cognition Pattern By EEG and Facial Expression Combination", IEEE, VOL. 26, FEBRUARY 2022, pg.589-599 https://doi.org/10.1109/JBHI.2021.3092412
25. Lucas Goncalves, Carlos Busso, "Robust Audiovisual Emotion Recognition: Aligning Modalities, Capturing Temporal Information, and Handling Missing Features", IEEE, VOL. 13, OCTOBER-DECEMBER 2022, pg. 2156- 2169 https://doi.org/10.1109/TAFFC.2022.3216993
26. Ke Zhang, Yuanqing Li, Jingyu Wang, Erik Cambria, Xuelong Li, "Real-Time Video Emotion Recognition Based on Reinforcement Learning and Domain Knowledge", IEEE Vol 32, MARCH 2022, pg. 1034- 1047 https://doi.org/10.1109/TCSVT.2021.3072412
27. Norbert Braunsch weiler , Rama Doddipatla ,Simon Keizer, and Svetlana Stoyanchev, "Factors in Emotion Recognition With Deep Learning Models Using Speech and Text on Multiple Corpora",IEEE , VOL. 29, 2022, pg. 722-726 https://doi.org/10.1109/LSP.2022.3151551
28. Guan-Nan Dong, Chi-Man Pun, and Zheng Zhang, "Temporal Relation Inference Network for Multimodal Speech Emotion Recognition" IEEE, VOL. 32, SEPTEMBER 2022, pg.6472- 6485 https://doi.org/10.1109/TCSVT.2022.3163445
29. Jiménez-Guarneros, Magdiel, and Gibran Fuentes-Pineda. "CFDA-CSF: A Multi-modal Domain Adaptation Method for Cross-subject Emotion Recognition." IEEE Transactions on Affective Computing (2024). https://doi.org/10.1109/TAFFC.2024.3357656
30. Sun, Teng, et al. "Muti-modal Emotion Recognition via Hierarchical Knowledge Distillation." IEEE Transactions on Multimedia (2024). https://doi.org/10.1109/TMM.2024.3385180
31. Alsaadawı, Hussein Farooq Tayeb, and Resul Daş. "Multimodal Emotion Recognition Using Bi-LG-GCN for MELD Dataset." Balkan Journal of Electrical and Computer Engineering 12.1 (2024): 36-46. https://doi.org/10.17694/bajece.1372107
32. Kumar, Puneet, Sarthak Malik, and Balasubramanian Raman. "Interpretable multimodal emotion recognition using a hybrid fusion of speech and image data." Multimedia Tools and Applications 83.10 (2024): 28373-28394. https://doi.org/10.1007/s11042-023-16443-1
33. Umair, Muhammad, et al. "Emotion Fusion-Sense (Emo Fu-Sense)–A novel multimodal emotion classification technique." Biomedical Signal Processing and Control 94 (2024): 106224. https://doi.org/10.1016/j.bspc.2024.106224
34. Makhmudov, Fazliddin, Alpamis Kultimuratov, and Young-Im Cho. "Enhancing Multimodal Emotion Recognition through Attention Mechanisms in BERT and CNN Architectures." Applied Sciences 14.10 (2024): 4199. https://doi.org/10.3390/app14104199
35. Wang, Ruiqi, et al. "Husformer: A multi-modal transformer for multi-modal human state recognition." IEEE Transactions on Cognitive and Developmental Systems (2024). https://doi.org/10.1109/TCDS.2024.3357618

36. Pereira, Rafael, et al. "Systematic Review of Emotion Detection with Computer Vision and Deep Learning." Sensors 24.11 (2024): 3484. https://doi.org/10.3390/s24113484
37. Li, Xingye, et al. "Magdra: a multi-modal attention graph network with dynamic routing-by-agreement for multi-label emotion recognition." Knowledge-Based Systems 283 (2024): 111126. https://doi.org/10.1016/j.knosys.2023.111126
38. Wang, Shuai, et al. "Multimodal emotion recognition from EEG signals and facial expressions." IEEE Access 11 (2023): 33061-33068. https://doi.org/10.1109/ACCESS.2023.3263670
39. Lei, Yuanyuan, and Houwei Cao. "Audio-visual emotion recognition with preference learning based on intended and multi-modal perceived labels." IEEE Transactions on Affective Computing (2023). https://doi.org/10.1109/TAFFC.2023.3234777
40. Liu, Shuai, et al. "Multi-modal fusion network with complementarity and importance for emotion recognition." Information Sciences 619 (2023): 679-694. https://doi.org/10.1016/j.ins.2022.11.076
41. Hou, Mixiao, et al. "Semantic alignment network for multi-modal emotion recognition." IEEE Transactions on Circuits and Systems for Video Technology (2023). https://doi.org/10.1109/TCSVT.2023.3247822
42. Shahzad, H. M., et al. "Multi-Modal CNN Features Fusion for Emotion Recognition: A Modified Xception Model." IEEE Access (2023). https://doi.org/10.1109/ACCESS.2023.3310428
43. Zhang, Duzhen, et al. "Structure Aware Multi-Graph Network for Multi-Modal Emotion Recognition in Conversations." IEEE Transactions on Multimedia (2023). https://doi.org/10.1109/TMM.2023.3238314
44. Zhang, Yazhou, et al. "M3GAT: A Multi-modal, Multi-task Interactive Graph Attention Network for Conversational Sentiment Analysis and Emotion Recognition." ACM Transactions on Information Systems 42.1 (2023): 1-32 https://doi.org/10.1145/3593583
45. Singh, Gopendra Vikram, et al. "Emoint-trans: A multimodal transformer for identifying emotions and intents in social conversations." IEEE/ACM Transactions on Audio, Speech, and Language Processing 31 (2022): 290-300. https://doi.org/10.1109/TASLP.2022.3224287
46. Zou, ShiHao, et al. "Improving multimodal fusion with Main Modal Transformer for emotion recognition in conversation." Knowledge-Based Systems 258 (2022): 109978. https://doi.org/10.1016/j.knosys.2022.109978
47. Lian, Zheng, Bin Liu, and Jianhua Tao. "Smin: Semi-supervised multi-modal interaction network for conversational emotion recognition." IEEE Transactions on Affective Computing (2022). https://doi.org/10.1109/TAFFC.2022.3141237
48. Yoon, Yeo Chan. "Can we exploit all datasets? Multimodal emotion recognition using cross-modal translation." IEEE Access 10 (2022): 64516-64524. https://doi.org/10.1109/ACCESS.2022.3183587
49. Wang, Qian, et al. "Multi-modal emotion recognition using EEG and speech signals." Computers in Biology and Medicine 149 (2022): 105907. https://doi.org/10.1016/j.compbiomed.2022.105907
50. Zheng, Jiahao, et al. "Multi-channel weight-sharing autoencoder based on cascade multi-head attention for multimodal emotion recognition." IEEE Transactions on Multimedia (2022). https://doi.org/10.1109/TMM.2022.3144885
51. Yang, Dingkang, et al. "Contextual and cross-modal interaction for multi-modal speech emotion recognition." IEEE Signal Processing Letters 29 (2022): 2093-2097. https://doi.org/10.1109/LSP.2022.3210836
52. Liu, Wei, et al. "Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition." IEEE Transactions on Cognitive and Developmental Systems 14.2 (2021): 715-729. https://doi.org/10.1109/TCDS.2021.3071170
53. Guanghui, Chen, and Zeng Xiaoping. "Multi-modal emotion recognition by fusing correlation features of speech-visual." IEEE Signal Processing Letters 28 (2021): 533-537. https://doi.org/10.1109/LSP.2021.3055755
54. Kanani, P., & Padole, Dr. M. (2019). Deep Learning to Detect Skin Cancer using Google Colab. In International Journal of Engineering and Advanced Technology (Vol. 8, Issue 6, pp. 2176–2183). https://doi.org/10.35940/ijeat.f8587.088619
55. Sultana, N., Rahman, Md. T., Parven, N., Rashiduzzaman, M., & Jabiullah, Md. I. (2020). Computer Vision based Plant Leaf Disease Recognition using Deep Learning. In International Journal of Innovative Technology and Exploring Engineering (Vol. 9, Issue 5, pp. 622–626). https://doi.org/10.35940/ijitee.e2486.039520
56. Radhamani, V., & Dalin, G. (2019). Significance of Artificial Intelligence and Machine Learning Techniques in Smart Cloud Computing: A Review. In International Journal of Soft Computing and Engineering (Vol. 9, Issue 3, pp. 1–7). https://doi.org/10.35940/ijsce.c3265.099319

AUTHORS PROFILE

Sanjeeva Rao Sanku, is pursuing a Ph.D in Computer Science and Engineering from Osmania University, Hyderabad. He completed his M.Tech and B.Tech in CSE from JNTU Hyderabad University. His research areas include Data mining, Machine Learning, and Deep Learning. He has 17+ years of teaching and research experience.

Prof. B. Sandhya, was awarded a Ph.D in Computer Science from the University of Hyderabad in the year 2011. She is currently working as a Professor in the CSE department at MVSR Engineering College, Hyderabad. She has 22+ years of teaching, research, and consultancy experience. Her principal areas of research include Image Processing, Machine learning, Deep learning, and Computer Vision. She has authored/co-authored around 40 research publications in international conferences and journals, with total citations of about 170.