Eksplorium *p*-ISSN 0854-1418
Volume 46 No. 1, May 2025: 1068–1081 *e*-ISSN 2503-426X

A Comprehensive Analysis of Integrative Approaches for Emotion Recognition Utilizing Deep Learning Frameworks

[1]Ramakrishna Gandi, [2]Dr. A. Geetha, [3]Dr. B. Ramasubba Reddy

[1]Research Scholar,Computer Science & Engineering Department, Annamalai University, Annamalainagar,Tamil Nadu, 608002, INDIA

E-mail: gandiramakrishna2@gmail.com

[2]Professor, Computer Science & Engineering Department, Annamalai University, Annamalainagar, Tamil Nadu, 608002, INDIA

E-mail: aucsegeetha@yahoo.com

[3]Professor, Department of CSE, MOHAN BABU UNIVERSITY, Sree Sainath Nagar, Tirupati, Andhra Pradesh, 517102, INDIA

E-mail: rsreddyphd@gmail.com

Article Received: 25 Feb 2025, Revised: 22 April 2025, Accepted: 04 May 2025

Abstract: The study of emotion recognition is progressing quickly and holds considerable potential for use in areas such as healthcare,education,and interactions between humans and computers. Integrative emotion recognition systems that combine information from diverse inputs like facial expressions,vocal tones, body language,and physiological indicators provide enhanced precision and reliability when compared to single-modality systems. This systematic review examines recent advancements, methodologies, and challenges in the design and implementation of multi-modal emotion recognition systems. The review highlights key components, including methods for extracting features,fusion strategies,and deep learning frameworks,emphasizing their use in practical situations,. The findings underscore the potential of multi-modal systems to provide context-aware, reliable, and inclusive emotion recognition capabilities while addressing limitations such as data privacy concerns, computational complexity, and integration challenges. The present analysis wraps up by exploring potential avenues for future research, emphasizing the need for standardized datasets, cross-cultural studies, and interpretable models to advance the field and its practical applications.

Keywords- multimodal, emotion recognition, neural networks, fusion, feature extraction, facial, motion dynamics

INTRODUCTION

Communication and engagement are enhanced by emotions, which enable individuals to express themselves nonverbally. They impact personal choices because emotion affects learning, information transfer, perception, and behaviors and provides behavioral flexibility beyond instinct [1][50]. Happy emotions can improve physical,cognitive,and social capabilities,and sharing these emotions through positive interactions with others can increase one’s affinity for them[2]. Negative emotions can weaken the immune system's ability to combat infections and cancerous cells, while expressing feelings of hatred may amplify anger[3]. Recurring emotions may become personality traits, and emotion can be used to assess how an experience and its effects affect an individual. Machines or virtual agents equipped with emotional artificial intelligence are capable of recognizing the emotional and attentional cues of human users during interactions.. Sectors such as healthcare, education, security, safety, consumer marketing, and

entertainment could gain advantages from effective automated emotion recognition. Emotions are dynamic and triggered by evoking stimulation [4]. The process outputs show changes in three major emotion reaction components: an emotion is characterized by unbidden physicalarousal, observable behavioral reactions, and a personal feeling, regardless of the theory of emotion [physiological, neurological, cognitive]. Motor expression communicates between people. Emotions are differentiated by their output signals [5]. Action inclination (a conative response that may affect motor expression) and cognitive assessment of the triggering stimuli are the other two emotion reaction components. Some study psychologists believe cognitive assessment precedes emotions rather than reacting to them. Neurophysiological processes affect subjective emotion, motor expression, cognitive, and conative responses.Machines are capable of identifying a person's emotional state by analyzing motor expression signals such as speech (e.g., [9]), dynamic facial expressions[6], and various visual cues—including posture, gestures, eye movements, and lip or body motions[7]—as well as tactile signals. Speech recognition has been popular in recent decades. Research in human-to-machine relies heavily on these methods,with initial strategies involving manual feature extraction alongside statistical models such as Gaussian mixture models (GMM) [4] and Hidden Markov models (HMM) [8]. Recently, Recurrent neural networks (RNNs) [9] and convolutional neural networks (CNNs) [10] have demonstrated strong performance in voice recognition tasks. Additionally, emotion extraction from text is crucial and a growing field of natural language processing study. Text emotion recognition may increase HCI quality. These findings show that prior surveys have not completely captured deep MER developments [11]. Therefore, a complete overview article on deep multimodal emotion identification is needed to grasp the newest technical advances. A gap in deep learning-based MER is addressed by our survey. Our investigation into deep learning-based MER offers a more comprehensive analysis than previous works; beyond fusion techniques, we also examine multimodal datasets and methods for extracting emotional features. This thorough study distinguishes our research from previous surveys, which focused on certain MER features.

REVIEW OF DIFFERENT TECHNIQUES USED FOR FEATURE EXTRACTION IN MULTIMODAL DATA ANALYSIS

Multimodal emotion recognition systems, which combine inputs like facial expressions,gestures,speech and physiological signals, have become increasingly important because they offer more thorough and precise assessments of emotional states. A key factor in the effectiveness of these systems is feature extraction,which converts raw input into informative representations that can be used for classification and analysis. Different modalities require specialized feature extraction techniques tailored to their unique characteristics; for instance, facial expression analysis relies on image processing and computer vision methods, while speech emotion recognition leverages acoustic and linguistic features.

As reported in [12],the use of Coordinate Attention enhanced the model's ability to detect intricate features of expressions. CReToNeXt, a major improvement, improved the model's

nuanced expression recognition. Performance evaluations demonstrated that CReToNeXt-YOLOv5 achieved a mean average accuracy (mAP) of 89.4%, surpassing YOLOv5 by 6.7%. Xception model from Keras API and wavelet transform were used to extract features for classification using various classifiers in [13]. DL-based EfficinetNetB0 extracts features from pre-processed pictures for further processing in [14]. The Red Fox Optimizer (RFO) optimizes the kernel parameters of the Weighted Kernel Extreme Learning Machine (WKELM) for emotion categorization. In [15] presents a novel attention-based technique to multimodal emotion identification. This method blends face and audio characteristics retrieved by separate encoders to choose the most informative. It processes voice and face characteristics of varied sizes and prioritizes usable information to improve system accuracy. HOG, SIFT, and convolutional neural network feature extraction are employed in [16]. The suggested model can distinguish emotions with 98.48% accuracy for HOG-CNN and 97.96% for SIFT-CNN using CKplus dataset. HOG-CNN has 91.43% accuracy and SIFT-CNN 82.85% accuracy on the Jaffe dataset. Reference[17] presents a method for subject-independent emotion recognition from EEG signals,employing Variational Mode Decomposition (VMD) for feature extraction and a Deep Neural Network as the classifier. Reference[18] presents a technique for selecting key emotion-related subnetworks and analyzes EEG functional connectivity using measures such as network strength, clustering coefficient, and eigenvector centrality. Three network characteristics and eye movement features were sent to a multimodal emotion identification model utilizing deep canonical correlation analysis. The discriminating capacity of EEG connectivity characteristics in emotion detection is examined using SEED, SEED-V, and DEAP public datasets. A tunable Q wavelet transform (TQWT) extracts features in [19]. Dimension reduction uses six statistical approaches. The classification stage uses rotation forest ensemble (RFE) classifier and several techniques. In [20] objected to the feature-based strategy for 2D face photos. SURF and SIFT are utilized for feature extraction. The authors achieved 99.7% identification accuracy using SIFT (64-components) and SURF (32-components). A Multi-learning strategy is utilized to concurrently extract spatially relevant emotional features and model long-term contextual dependencies from voice signals [21]. We utilized a residual blocks with skip connection (RBSC) module to capture correlations and emotional cues, while sequence learning (Seq_L) was applied to model long-term contextual dependencies with in the input features.

Author/year	**Feature extraction method**	**RMSE**	**Reconstruction error**	**Extraction time**
Nie et al., (2024)	Coordinate Attention mechanism	moderate	high	low
Phukan et al., (2024)	wavelet transform	high	moderate	low
Anand et al., (2024)	EfficinetNetB0	low	high	Moderate

Mamieva et al., (2023)	attention-based approach	moderate	high	low
Gautam et al., (2023)	HOG and SIFT	low	low	moderate
Pandey et al., (2022)	Variational Mode Decomposition (VMD	low	moderate	high
Wu et al., (2022)	emotion-relevant critical subnetworks	low	high	low
Subasi et al., (2021)	tunable Q wavelet transform	moderate	low	low
Gupta et al., (2021)	SURF and SIFT	high	high	moderate
Kwon et al., (2021)	used residual blocks with a skip connection	low	moderate	low

SURVEY ON VARIOUS FUSION METHODS

The fusion process integrates data from various modalities to capture complementary and correlated features that are often missed in unimodal systems. Fusion can be carried out at various stages- including data-level, feature-level, and decision-level -each providing distinct advantages and presenting particular challenges. This section explores the landscape of fusion methods, comparing traditional and contemporary approaches, highlighting their strengths and limitations, and discussing their impact on the effectiveness of multimodal emotion recognition systems.

In [22], present a two-stream architecture and two-level spatio-temporal feature. This architecture includes a classification head, a weighted average operator for temporal information aggregation, a spatial encoder (modified ResNet) to extract facial texture features, a temporal encoder (Swin Transformer) to model facial muscle movements, and a feature fusion algorithm to combine multi-level spatio-temporal features. In [23] proposes a multi-scale convolution and TimesNet network fusion model. TimesNet network reconstructs 2D sequences to extract complicated temporal information, and different convolution kernels help extract dynamic spatio-temporal properties. The model enhances emotion recognition. Reference [24] proposes a Coordinated-representation Decision Fusion Network (CoDF-Net) designed to effectively fuse EEG and eye movement data. Subsequently, the Decision-level Fusion Broad Learning System (DF-BLS) constructs multiple subsystems to effectively determine the final emotional states

through robust decision-making. The CoDF-Net achieves accurancy rates of 94.09 and 91.62% in subject-dependent evaluations,and 87.04 and 83.87% in subject-independent evaluations on the SEED-CHN and SEED-GER datasets,respectively.Reference[25] introduces DER-GCN, a Dialogue and Event Relation-Aware Graph Convolutional Neural Network for Multimodal Emotion Recognition. To enhance feature and structure fusion representation, we offer a Self-Supervised Masked Graph Autoencoder (SMGAE). Subsequently, we develop a new Multiple Information Transformer (MIT) to enhance the modeling of relationships and facilitate more effective fusion of multivariate relational data. To conclude, we introduce a loss optimization method based on contrastive learning to strengthen feature representation learning for minority classes. In [26], offer a new attention mechanism-based multiscale feature fusion network (AM-MSFFN) that uses high-level features at multiple scales to enhance model generalization across individuals. Following point-wise convolution, convolutional module attention mechanism (CBAM) is used to address EEG changes across people and classify essential information. To increase training sample quality, we use a data augmentation and alignment preprocessing package. Ablation experiments further reveal that the suggested attention mechanism and multiscale separable convolution improve our AM-MSFFN model performance consistently. In [27],an attention network is developed utilizing complementary modalities,and the reconstructed features are subsequently combined to form a multi-modal representation with enhanced interaction. Reference [28] offers CFNet, a multi-task joint learning network with constraint fusion. CFNet employs a multi-loss strategy and constraint fusion to automatically weight global and local face information,effectively capturing essential characteristics from diverse tasks. CFNet adapts better to FER in complicated scenarios than direct fusion models. In [29] uses Mel-Frequency Cepstral Coefficients (MFCC), mel-scale spectrogram, tonal power, and spectral flux to extract speech features. Deer Hunting with Adaptive Search (DH-AS) selects the appropriate feature to reduce feature size and improve learning efficiency. Hybrid Deep Learning (HDL) using “Deep Neural Network (DNN) and Recurrent Neural Network (RNN)” uses these ideal characteristics to classify emotions. The DH-AS improves these two networks, enabling them to precisely emotions like "happy, sad, anger, fear, calm, etc. In [30], a convolutional neural network (CNN) with four local feature-learning modules and an LSTM layer captures both local and long-term dependencies in the log Mel-spectrogram of the input audio samples. The recently introduced stochastic fractal search (SFS)-guided whale optimization algorithm fine-tunes the learning rate and label smoothing regularization parameter to improve the performance of this deep network. This algorithm's capacity to balance search agent exploration and exploitation ensures the optimum global solution. In [31] introduces GP-FER, a genetic programming framework for facial expression recognition, which selects and fuses features. This approach relies on a tree-based genetic algorithm with three functional levels (feature selection, fusion, and classification). The proposed genetic program is a binary classifier that selects and fuses features differently for each expression class pair. In [32] the fusion model, several GCNNs collect graph domain characteristics, LSTM cells remember the connection between two channels over time and extract temporal data, and Dense layer classifies emotions. A Multiple

Support Vector Neural Network (Multi-SVNN) classifier, utilizing the Whale-Imperialist Optimization algorithm (Whale-IpCA), was introduced in [33] for the task of speech emotion recognition.The Whale-IpCA method trains the Multi-SVNN classifier for emotion recognition by integrating the Whale Optimization Algorithm (WOA) with the Imperialist Competitive Algorithm (IpCA). Additionally, the Whale-IpCA-based Multi-SVNN receives the input signal's spectral feature set for identification. Reference [49] offers the Attention Convolutional Gated Recurrent Neural Network with RMSProp (ACGRNN- RMSProp) classifier aims to achieve high accuracy and robustness in micro facial expression recognition. The attention mechanism sharpens the model's emphasis on important facial regions, while convolutional layers gather spatial features, and GRUs effectively capture temporal dynamics. The RMSProp optimizer ensures efficient and stable training, making this approach well-suited for the complex task of recognizing subtle and fleeting micro expressions.

Author/year	**Fusion method**	**Accuracy**	**Precision**	**RMSE**	**Fusion latency**
Wang et al., (2024)	modified ResNet+swin transformer	moderate	low	moderate	high
Han et al., (2024)	MS-TimesNet	moderate	moderate	high	low
Gong et al., (2024)	CoDF-Net		high	moderate	high
Ai et al., (2023)	SMGAE	moderate	moderate	low	high
Jiang et al., (2023)	AM-MSFFN	low	high	high	moderate
Liu et al., (2023)	attention network	high	low	high	moderate
Xiao et al., (2023)	CFNet	high	moderate	moderate	low
Manohar et al., (2022)	Deer Hunting with Adaptive Search	moderate	moderate	high	low
Abdelhamid et al., (2022)	fractal search (SFS)-guided whale optimization	low	high	moderate	high

	algorithm (WOA).				
Ghazouani et al., (2021)	GP−FER	low	moderate	high	high
Yin et al., (2021)	multiple GCNNs	moderate	high	low	low
Mannepalli et al., (2021)	Whale-IpCA	low	moderate	high	moderate

SURVEY ON CLASSIFIERS FOR MULTIMODAL EMOTION DETECTION

Neural network classifiers have become a cornerstone in multimodal emotion detection, leveraging their ability to model complex, nonlinear relationships across diverse data modalities. Driven by progress in deep learning, neural networks are now widely used to integrate and interpret data from modalities such as facial expressions, speech, physiological signals, and text. These classifiers excel in capturing the intricate dependencies and complementary features between modalities, thereby enhancing emotion recognition accuracy.

In [34], an Efficient Long-distance Latent Relation-aware Graph Neural Network (ELR-GNN) was proposed for identifying emotions in multi-modal conversational data.Initially, we input pre-extracted text, video, and audio features into a Bi-LSTM to capture contextual semantic information and low-level attributes of the utterances. Next, we build a conversational emotion interaction network using low-level utterance attributes.By employing the dilated generalized forward push algorithm to precompute emotional propagation across global utterances and integrating an emotion relation-aware operator to model semantic associations, we effectively capture dependencies between long-distance utterances.In [35], multi-input CNNs. Because various characteristics are derived from diverse inputs, multimodal techniques enable multiple modalities to collaborate to improve performance. Two sets of films depending on expression intensity—acted/strong or spontaneous/normal—will be used to portray the following emotional states: Anger, contempt, fear, joy, and sadness. In [36], a Dialog and Event Relation-aware Graph Convolutional Network (DER-GCN) was introduced for multimodal emotion recognition. It collects latent event relations and models speaker dialog. We provide a contrastive learning-based loss optimization technique to improve minority class feature representation learning. To address these issues, [37] offer a fuzzy temporal convolutional network based on contextual self-attention (CSAT-FTCN) with a membership function modeling diverse fuzzy emotions for deeper emotional comprehension.To enhance emotion understanding, CSAT-FTCN identifies dependency relationships of target utterances based on both internal key features and external contextual cues.In [38],a Transformer-based Deep-Scale Fusion Network (TDFNet) was proposed for multimodal emotion recognition to address existing challenges. The TDFNet multimodal embedding (ME) module combines pretrained models and a lot of unlabeled data to

improve data scarcity by giving the model multimodal information. In [39], offers a manifold learning and CNN-based multimodal emotion detection model.By integrating EEG signals with peripheral physiological and eye movement signals, the Multivariate Synchrosqueezing Transform (MSST) models the joint oscillatory structure of multi-channel signals, enabling the extraction and fusion of feature parameters into comprehensive feature vectors. The COGMEN system [40] uses local (i.e., speaker inter/intra dependence) and global (context) information to recognize emotions. To simulate complicated local and global communication relationships, the suggested model leverages Graph Neural Network (GNN) architecture. In [41], a 3D-Convolutional Neural Network (3D-CNN) architecture is utilized to extract spatio-temporal features from EEG signals, enabling effective prediction of EEG-based responses. Mask-RCNN object detection and OpenCV libraries are used to extract emotional face pixels for the face method. Then, the SVM classifier classifies face chunk 3D-CNN output features. Bagging and stacking are studied for fusion-based emotion identification. In [42] deep canonical correlation analysis (DCCA) and bimodal deep autoencoder (BDAE). DCCA, BDAE, and standard methods are systematically compared on five multimodal data sets.In [43], a pre-trained Spatial Transformer Network utilizing saliency maps and facial images is followed by an attention-based Bi-LSTM. Error analysis revealed that frame-based systems may struggle with video-based tasks, even with domain adaptation, highlighting the need for new research to address this mismatch and effectively leverage the embedded knowledge of pre-trained models.

Author/year	**Classifier**	**Dataset adopted**	**outcome**
Shou et al., (2024)	ELR-GNN	IEMOCAP and MELD	reduced by 52% and 35% respectively for both datasets
Bilotti et al., (2024)	CNN	BAUM-1 and RAVDESS	The highest accuracy obtained on the BAUM-1 dataset is approximately 95%, while the RAVDESS dataset achieved around 95.5%
Ai et al.,(2024)	DER-GCN	IEMOCAP and MELD	96.5% of accuracy
Jiang et al., (2023)	CSAT-FTCN	MELD	98.3% of accuracy
Zhao et al., (2023)	TDFNet	IEMOCAP	82.08% of accuracy
Zhang et al., (2022)	DCNN	DEAP and MAHNOB-HCI	average accuracies of 90.05% and 88.17%

Joshi et al., (2022)	COGMEN	IEMOCAP and MOSEI	7.7% F1-score increase for IEMO
Salama et al., (2021)	3D-CNN	MOSEI dataset	96.13%, and 96.79% of accuracy
Liu et al., (2021)	DCCA+ BDAE	SEED, SEED-IV, DEAP, SEED-V, DREAMER	94.6%, 87.5%, 84.3%, 85.3%. 89% of accuracy
Luna-Jiménez et al., (2021)	bi-LSTM with an attention mechanism	RAVDESS	89.4% of accuracy

SURVEY ON VARIOUS DATASETS

MER has numerous major datasets to aid study and experimentation. Despite the large variety of datasets with their own pros and cons, there are still areas that need additional study. We evaluate these materials in detail here.

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [44], developed by the Signal Analysis and Interpretation Lab (SAIL) at the University of Southern California, serves as a valuable multimodal benchmark for emotion recognition research. Data gathering included 10 actors, equally male and female. This gender-paired cast performed planned and spontaneous talks in five groups. These chats add emotional content to the dataset, making it more reflective of real-world emotional communication. IEMOCAP is a rich source of emotional contexts with 4784 spontaneous and 5255 scripted talks.The inclusion of nine distinct emotions—such as happiness, sadness, anger, surprise, fear, disgust, frustration, excitement, and neutrality—along with continuous emotional dimensions like activation, arousal, and dominance, enables a more dynamic and nuanced emotional analysis.

Youtube Datase t- The YouTube dataset was presented by Morency et al. in 2011 [45]. The dataset includes 47 videos, 20 of which are female and 27 exhibiting male viewpoints. The dataset shows emotional expressions from 14 to 60, adding variety. Each dataset video has people speaking English, despite their diverse ethnic origins. Complexity is essential for developing models that work in noisy real-world circumstances. Emotionally labeled films boost the dataset's worth. All 47 videos are labeled with good, negative, or neutral emotions. Categories give a robust basis for deep learning supervised learning.

MOUD - This dataset was created by Perez-Rosas et al. in 2013 [46]. The dataset provides a distinctive viewpoint by emphasizing statements that are based on opinions, which are inherently emotionally charged. The MOUD has 498 statements meticulously categorized good, negative, or neutral. MOUD is fascinating since it uses YouTube videos, one of the world's largest user-generated content sources. The inclusion of real-world, unscripted emotional expressions

enhances the robustness and real-world relevance of models trained on this dataset. It features speakers aged 20 to 60, capturing a broad range of emotional responses across age groups, and includes 15 female speakers to support gender diversity. All participants provide product reviews in Spanish.

ICT-MMMO - In 2013, Wollmer et al. created the Institute for Creative Technologies' Multi-Modal Movie Opinion (ICT-MMMO) database [47]. It combines ExpoTV with YouTube videos. The dataset uses YouTube's richness of emotionally charged user-generated content, including 370 videos. Sentiment analysis of real user comments distinguishes ICT-MMMO. Real-world context is essential for durable deep learning models. YouTube has 228 positive, 57 negative, and 23 neutral videos, and ExpoTV has 62 unfavorable movie review videos. The ICT-MMMO database classifies videos as strongly positive, weakly positive, neutral, strongly negative, or weakly negative.

CMU-MOSI- Due to its subjective sentiment and emotional intensity annotations, Zadeh et al.'s 2016 Multimodal Opinion-level Sentiment Intensity (CMU-MOSI) dataset [48] is a milestone in multimodal sentiment analysis databases. There are 93 randomly selected YouTube videos with 89 speakers (41 females and 48 males). These films were shot in various places using various mics and cameras. The background, illumination, and user-camera distance varied. Videos retain their integrity. Since user-generated content has diverse audio-visual quality, this realistic method ensures a comprehensive training set for real-world applications. Video, book, and product reviews are among the CMU-MOSI dataset's themes.

CONCLUSION

This review delineates the evolution and present status of deep neural network-based Multimodal Emotion Recognition technology. We have analyzed prominent feature extraction techniques, assessed notable multimodal emotion datasets, and evaluated the impact of deep learning technology in this swiftly evolving domain. We have examined representative fusion techniques, feature extraction methodologies, and classification approaches. This systematic review highlights the advancements in neural network architectures, including CNNs, RNNs, transformers, and hybrid models, which have significantly improved the accuracy, robustness, and scalability of emotion detection systems. Despite their success, challenges such as high computational costs, data scarcity, modality synchronization, and the need for interpretable models remain pressing concerns. Future research should focus on developing lightweight, energy-efficient architectures, leveraging unsupervised and transfer learning for limited data scenarios, and enhancing model interpretability to gain insights into decision-making processes. Additionally, cross-cultural and context-aware studies are critical to improving the inclusivity and generalizability of these systems. By addressing these challenges, neural network-based multimodal emotion recognition systems can achieve their full potential in real-world applications across healthcare, education, and human-computer interaction.

REFERENCE

[1] Li, J.; Mishra, S.; El-Kishky, A.; Mehta, S.; Kulkarni, V. NTULM: Enriching social media text representations with non-textual units. *arXiv* **2022**, arXiv:2210.16586

[2] Graves, A.; Mohamed, A.R.; Hinton, G. Speech recognition with deep recurrent neural networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; IEEE: Piscataway, NJ, USA; pp. 6645–6649. [**Google Scholar**]

[3] Mao, Q.; Dong, M.; Huang, Z.; Zhan, Y. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Trans. Multimed.* **2014**, *16*, 2203–2213.

[4] Kandali, A.B.; Routray, A.; Basu, T.K. Emotion recognition from Assamese speeches using MFCC features and GMM classifier. In Proceedings of the TENCON 2008—2008 IEEE Region 10 Conference, Hyderabad, India, 19–21 November 2008; IEEE: Piscataway, NJ, USA; pp. 1–5

[5] Breuel, T.M. High performance text recognition using a hybrid convolutional-lstm implementation. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; IEEE: Piscataway, NJ, USA, 2017; Volume 1, pp. 11–16

[6] Guo, L.; Wang, L.; Dang, J.; Fu, Y.; Liu, J.; Ding, S. Emotion Recognition with Multimodal Transformer Fusion Framework Based on Acoustic and Lexical Information. *IEEE MultiMedia* **2022**, *29*, 94–103

[7] Sharma, M., & Sharma, S. R. (2025). Advanced hydrological simulation and hybrid CNN-LSTM models for sustainable water resource management in Nepal. Journal of Information Systems Engineering and Management, 10(31s). https://doi.org/10.52783/jisem.v10i31s.5059

[8] Guo, W.; Wang, J.; Wang, S. Deep multimodal representation learning: A survey. *IEEE Access* **2019**, *7*, 63373–63394.

[9] Nwe, T.L.; Foo, S.W.; De Silva, L.C. Speech emotion recognition using hidden Markov models. *Speech Commun.* **2003**, *41*, 603–623.

[10] Graves, A.; Mohamed, A.R.; Hinton, G. Speech recognition with deep recurrent neural networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; IEEE: Piscataway, NJ, USA; pp. 6645–6649. [**Google Scholar**]

[11] Mao, Q.; Dong, M.; Huang, Z.; Zhan, Y. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Trans. Multimed.* **2014**, *16*, 2203–2213

[12] Koromilas, P.; Giannakopoulos, T. Deep multimodal emotion recognition on human speech: A review. *Appl. Sci.* **2021**, *11*, 7962.

[13] Nie, L., Li, B., Du, Y., Jiao, F., Song, X., & Liu, Z. (2024). Deep learning strategies with CReToNeXt-YOLOv5 for advanced pig face emotion detection. *Scientific Reports*, *14*(1), 1679.

[14] Phukan, A., Gupta, D. (2024). Deep feature extraction from EEG signals using xception model for emotion classification. *Multimedia Tools and Applications*, *83*(11), 33445-33463.

[15] Anand, M., Babu, S. (2024). Multi-class facial emotion expression identification using dl-based feature extraction with classification models. *International Journal of Computational Intelligence Systems*, *17*(1), 25.

[16] Mamieva, D., Abdusalomov, A. B., Kutlimuratov, A., Muminov, B., & Whangbo, T. K. (2023). Multimodal emotion detection via attention-based fusion of extracted facial and speech features. *Sensors*, *23*(12), 5475.

[17] Gautam, C., Seeja, K. R. (2023). Facial emotion recognition using Handcrafted features and CNN. *Procedia Computer Science*, *218*, 1295-1303.

[18] Pandey, P., Seeja, K. R. (2022). Subject independent emotion recognition from EEG using VMD and deep learning. *Journal of King Saud University-Computer and Information Sciences*, *34*(5), 1730-1738.

[19] Wu, X., Zheng, W. L., Li, Z., & Lu, B. L. (2022). Investigating EEG-based functional connectivity patterns for multimodal emotion recognition. *Journal of neural engineering*, *19*(1), 016012.

[20] Subasi, A., Tuncer, T., Dogan, S., Tanko, D., & Sakoglu, U. (2021). EEG-based emotion recognition using tunable Q wavelet transform and rotation forest ensemble classifier. *Biomedical Signal Processing and Control*, *68*, 102648.

[21] Gupta, S., Thakur, K., Kumar, M. (2021). 2D-human face recognition using SIFT and SURF descriptors of face's feature regions. *The Visual Computer*, *37*(3), 447-456.

[22] Kwon, S. (2021). MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach. *Expert Systems with Applications*, *167*, 114177.

[23] Wang, Z., Yang, M., Jiao, Q., Xu, L., Han, B., Li, Y., & Tan, X. (2024). Two-level spatio-temporal feature fused two-stream network for micro-expression recognition. *Sensors*, *24*(5), 1574.

[24] Han, L., Zhang, X., Yin, J. (2024). EEG emotion recognition based on the TimesNet fusion model. *Applied Soft Computing*, *159*, 111635.

[25] Gong, X., Dong, Y., Zhang, T. (2024). CoDF-Net: coordinated-representation decision fusion network for emotion recognition with EEG and eye movement signals. *International Journal of Machine Learning and Cybernetics*, *15*(4), 1213-1226.

[26] Ai, W., Shou, Y., Meng, T., Yin, N., & Li, K. (2023). Der-gcn: Dialogue and event relation-aware graph convolutional neural network for multimodal dialogue emotion recognition. *arXiv preprint arXiv:2312.10579*.

[27] Jiang, Y., Xie, S., Xie, X., Cui, Y., & Tang, H. (2023). Emotion recognition via multiscale feature fusion network and attention mechanism. *IEEE Sensors Journal*, *23*(10), 10790-10800.

[28] Liu, S., Gao, P., Li, Y., Fu, W., & Ding, W. (2023). Multi-modal fusion network with complementarity and importance for emotion recognition. *Information Sciences*, *619*, 679-694.

[29] Xiao, J., Gan, C., Zhu, Q., Zhu, Y., & Liu, G. (2023). CFNet: Facial expression recognition via constraint fusion under multi-task joint learning network. *Applied Soft Computing*, *141*, 110312.

[30] Manohar, K., & Logashanmugam, E. (2022). Hybrid deep learning with optimal feature selection for speech emotion recognition using improved meta-heuristic algorithm. *Knowledge-based systems*, *246*, 108659.

[31] Abdelhamid, A. A., El-Kenawy, E. S. M., Alotaibi, B., Amer, G. M., Abdelkader, M. Y., Ibrahim, A., & Eid, M. M. (2022). Robust speech emotion recognition using CNN+ LSTM based on stochastic fractal search optimization algorithm. *Ieee Access*, *10*, 49265-49284.

[32] Ghazouani, H. (2021). A genetic programming-based feature selection and fusion for facial expression recognition. *Applied Soft Computing*, *103*, 107173.

[33] Yin, Y., Zheng, X., Hu, B., Zhang, Y., Cui, X. (2021). EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM. *Applied Soft Computing*, *100*, 106954.

[34] Mannepalli, K., Sastry, P. N., Suman, M. (2022). Emotion recognition in speech signals using optimization based multi-SVNN classifier. *Journal of King Saud University-Computer and Information Sciences*, *34*(2), 384-397.

[35] Shou, Y., Ai, W., Du, J., Meng, T., Liu, H., & Yin, N. (2024). Efficient long-distance latent relation-aware graph neural network for multi-modal emotion recognition in conversations. *arXiv preprint arXiv:2407.00119*.

[36] Bilotti, U., Bisogni, C., De Marsico, M., Tramonte, S. (2024). Multimodal Emotion Recognition via Convolutional Neural Networks: Comparison of different strategies on two multimodal datasets. *Engineering Applications of Artificial Intelligence*, *130*, 107708.

[37] Ai, W., Shou, Y., Meng, T., Li, K. (2024). DER-GCN: Dialog and Event Relation-Aware Graph Convolutional Neural Network for Multimodal Dialog Emotion Recognition. *IEEE Transactions on Neural Networks and Learning Systems*.

[38] Jiang, D., Liu, H., Wei, R., Tu, G. (2023). CSAT-FTCN: a fuzzy-oriented model with contextual self-attention network for multimodal emotion recognition. *Cognitive Computation*, *15*(3), 1082-1091.

[39] Zhao, Z., Wang, Y., Shen, G., Xu, Y., Zhang, J. (2023). TDFNet: Transformer-based deep-scale fusion network for multimodal emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *31*, 3771-3782.

[40] Zhang, Y., Cheng, C., Zhang, Y. (2022). Multimodal emotion recognition based on manifold learning and convolution neural network. *Multimedia Tools and Applications*, *81*(23), 33253-33268.

[41] Joshi, A., Bhat, A., Jain, A., Singh, A. V., Modi, A. (2022). COGMEN: COntextualized GNN based multimodal emotion recognitioN. *arXiv preprint arXiv:2205.02455*.

[42] Salama, E. S., El-Khoribi, R. A., Shoman, M. E., Shalaby, M. A. W. (2021). A 3D-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition. *Egyptian Informatics Journal*, *22*(2), 167-176.

[43] Liu, W., Qiu, J. L., Zheng, W. L., Lu, B. L. (2021). Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, *14*(2), 715-729.

[44] Luna-Jiménez, C., Griol, D., Callejas, Z., Kleinlein, R., Montero, J. M., Fernández-Martínez, F. (2021). Multimodal emotion recognition on RAVDESS dataset using transfer learning. *Sensors*, *21*(22), 7665.

[45] 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. [**Google Scholar**] [**CrossRef**]

[46] Morency, L.P.; Mihalcea, R.; Doshi, P. Towards multimodal sentiment analysis: Harvesting opinions from the web. In Proceedings of the 13th International Conference on Multimodal Interfaces, Alicante, Spain, 14–18 November 2011; pp. 169–176. [**Google Scholar**]

[47] Pérez-Rosas, V.; Mihalcea, R.; Morency, L.P. Utterance-level multimodal sentiment analysis. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Sofia, Bulgaria, 4–9 August 2013; pp. 973–982. [**Google Scholar**]

[48] Wöllmer, M.; Weninger, F.; Knaup, T.; Schuller, B.; Sun, C.; Sagae, K.; Morency, L.P. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intell. Syst.* **2013**, *28*, 46–53. [**Google Scholar**] [**CrossRef**]

[49] Zadeh, A.; Zellers, R.; Pincus, E.; Morency, L.P. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv* **2016**, arXiv:1606.06259.

[50] Ramakrishna, Gandi.; Dr.A.Geetha;Dr.B. Ramasubba, Reddy.Spontaneous Micro-facial Expression Detection using Attention-based Convolutional Gated Recurrent Neural Networks with RMSProp Optimization. *Journal of Information Systems Engineering and Management.* **2025**,*10*(7s),630-642.

[51] Ramakrishna, Gandi.; Dr.A.Geetha;Dr.B. Ramasubba, Reddy.Comprehensive Survey on Recognition of Emotions from Body Gestures.*Journal of Informatics Education and Research.* **2025,***5*(1),545-557.