Design and Evaluation of a Multimodal Emotion Recognition Architecture Using Deep Neural Networks

Dr. E. Srikanth Reddy
Asst.Professor, Department of AI & ML
Vaageswari College of Engineering – Karimnagar, India

Venkatesh Mergu
Project Manager Systems
Mphasis Limited – Hyderabad, India

ABSTRACT:

Affective computing has become increasingly crucial for designing systems that interpret and react to human emotions. This study presents an open-source framework for multimodal emotion recognition, which integrates analysis of facial expressions, speech tone, and text sentiment for precise emotion detection from video content. The proposed solution utilizes advanced methods in computer vision, audio processing, and natural language processing, including deep learning architectures like convolutional neural networks (CNNs) for facial recognition and specialized neural models for other modalities. By combining these diverse features through an effective fusion strategy, the framework achieves improved emotion recognition performance, making it valuable for applications in industries such as film production and mental health assessment.

Keywords: **Affective computing, Multimodal emotion recognition, convolutional neural network, recurrent neural network, Fusion.**

1. INTRODUCTION

In recent years, rapid advancements in artificial intelligence and machine learning have notably improved computational systems' ability to interpret human actions and behavior. Yet, despite these technological leaps, the emotional aspects of human interaction remain largely misunderstood by most intelligent platforms—even though emotional awareness is fundamental for interactions that are truly natural, empathetic, and context-sensitive. The discipline known as affective computing has emerged from the intersection of psychology, neuroscience, and computer science, and its aim is to equip machines with the ability to recognize and respond to human emotions appropriately.

Human emotional expression is inherently multimodal, conveyed simultaneously through facial gestures, voice inflection, and language content. Systems that rely on only one modality to detect emotion tend to be less accurate and miss subtle emotional cues, as feelings are seldom communicated through a single channel alone. To address these shortcomings, multimodal

emotion recognition systems have been developed to integrate diverse sources of information, thereby improving the precision of emotional understanding. This integrative approach has found widespread use in mental health diagnostics, virtual agent technology, adaptive education platforms, and creative industries like film and drama.

Inspired by next-generation frameworks such as Imentiv AI, the proposed open-source multimodal emotion recognition system is designed to analyze video content and extract emotional cues by combining facial expression analysis, vocal tone interpretation, and sentiment detection from text. The system employs deep learning models—including convolutional neural networks and transformer architectures—to extract features and classify emotions from each source. By fusing these modalities, the framework brings artificial intelligence closer to human-like understanding, fostering more authentic and adaptive interactions. Its open development model promotes transparency, collaboration, and reproducibility, enabling researchers to extend its capabilities for diverse practical applications, including performance analytics, digital media production, psychological wellness monitoring, and virtual communications.

2. LITERATURE REVIEW

Understanding human emotions is always considered as a major challenge in making computers emotionally more intelligent. In past few years various methods were explored by researchers to enable machines to understand emotions based on facial expressions, voice tone, and textual patterns. These methods have transformed from depending on a single source of information (unimodal) to combining multiple source (multimodal), which reflects better that how emotions are analyzed int humans' real life.

a. Facial Emotion Recognition

Facial expressions are most direct way humans express their emotions. Early research used simple features of an image, such as edges and texture patterns to identify the emotional states of a person. Later he introduction of deep learning brought a breakthrough with CNNs which could automatically learn features from images and videos to perform the tasks of emotion recognition [1]. FER2013 and AffectNet among other datasets have contributed to the success in modeling happiness, anger, sadness, and much more [2]. More recent research works integrate CNNs with LSTM to capture temporal facial changes over time [3]. Instead of these improvements existing systems continue to face challenges when faces lighting conditions are poor.

b. Audio-Based Emotion Recognition

Emotions are also analyzed through speech and reflected in a person's tone, pitch and rhythm. Earlier approaches depend on traditional audio features such as Mel-Frequency Cepstral Coefficients (MFCCs) whereas recent methods consist deep learning models like CNNs, RNNs, and transformer-based architectures like wav2vec2 and WavLM to achieve higher accuracy and robustness [4]. Datasets such as RAVDESS, CREMA-D and IEMOCAP are commonly used for training and evaluating these models [5], [6]. While audio-based emotion recognition often

struggles with background noise, accent differences and emotional ambiguity which highlights the need for multimodal integration.

c. Text-Based Emotion Recognition

Language provides important cues for understanding emotions. Text-based emotion recognition analyzes words

and phrases to identify the emotional content. Transformer-based models such as BERT, RoBERTa, and DistilBERT have greatly enhanced accuracy in this domain [7]. Datasets like MELD and GoEmotions are commonly used to train and evaluate these models [8]. However, text-only approaches often overlook critical emotional context such as sarcasm, tone or speaker intent which makes the integration of text, audio and visual data essential for achieving more comprehensive emotion recognition.

3. RESEARCH GAP IDENTIFIED

1.Limited Open-Source Frameworks: Most current emotion recognition systems are not open-source, which limits accessibility, reproducibility, and collaborative research.

2.Incomplete Modality Integration: Many existing models combine only two modalities (e.g., audio–text or video–audio) and do not utilize all three — facial, vocal, and textual — together.

3.Basic Fusion Techniques: Fusion in most systems works on simple concatenation or averaging, without using advanced attention-based or transformer fusion methods for better context understanding.

4.Weak Multimodal Synchronization: Current frameworks often process modalities separately rather than synchronizing them from a single video input.

3. METHODOLOGY

This study presents an open-source multimodal emotion recognition framework. The goal of the system is to recognize human emotions from videos by analyzing facial expressions, speech tone, and spoken words together.

By combination of three types of information, the model achieves more accurate emotions similar to Human feelings in real life.

A. Dataset Description

The proposed multimodal emotion recognition system evaluated using the Movie Scene Emotion Detection Dataset a collection of short movie clips gathered from publicly available sources. Each video contains clear visual, vocal and linguistic cues corresponding to different emotional states such as happy, sad, angry confused and surprised.

Each record in the dataset includes:

Clip Path: Google Drive link to the video file.

Duration: The length of the clip in seconds.

Primary Emotion: The dominant emotion annotated for the clip.

Emotion Intensity: Label indicating emotion strength such as Low, Medium or High.

Ambiguity Flag: Indicates whether multiple emotions may overlap in the clip.

B. System Overview

The system begins by taking a video as input and separating it into three parts: visual frames, audio and text. Each of these parts (or modalities) is processed separately using deep learning models specialized for that type of data. Finally, the information from all three modalities is combined using an attention based fusion mechanism and the model predicts the final emotion using a classifier.

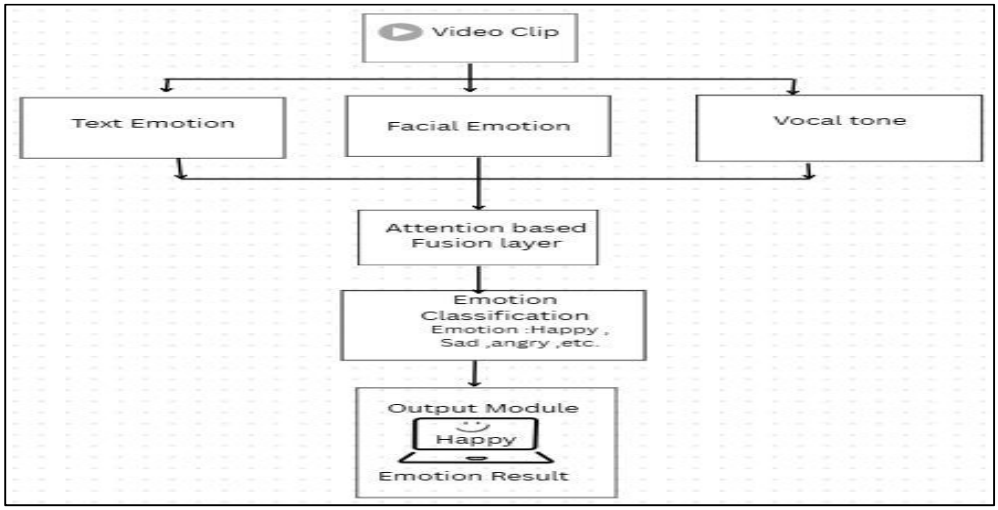

Figure 1. System architecture multimodal emotion recognition using facial, audio, and text cues

C. Data Preprocessing

Before analyzing emotions, the video is preprocessed so that each data type is ready for feature extraction.

1. Visual Preprocessing: The video is divided into individual frames using OpenCV. Each frame is resized and normalized so that the face can be detected easily and uniformly.

2. Audio Preprocessing: The audio is extracted from the video using MoviePy. The extracted audio is converted to .wav format to improve the clarity noise and background sound is reduced.

3. Text Preprocessing: The voice in the video is converted into text using the OpenAI Whisper model. The transcribed text is then cleaned by removing filler words and punctuation.

D. Facial Expression Analysis (Visual Modality)

The facial modality analyzes emotions on identifying through visual cues such as smiles, frowns and eyebrow movements.

1. Feature Extraction: Each frame is analyzed using the Deepface framework which uses the Convolutional Neural Network (CNN) such as ResNet50 or VGGFace is used to extract feature that represent facial muscle movements related to different emotions.

Emotion categories: The system detects seven base emotions such as happy sad, angry, fear, surprise, disgust and neutral.

E. Audio Emotion Analysis (Speech Modality)

This modality focuses on how emotions are expressed through the speaker's tone and voice.

1. Feature Extraction: Audio is extracted using Moviepy. The preprocessed audio is converted into Mel- Frequency Cepstral Coefficients (MFCCs) and spectrograms which represent variations in pitch, energy and rhythm.

2. Emotion Analysis: The SpeechBrain pretrained model wav2vec2 is used for classifying emotional tone. Wav2vec2 is a tranformer based model that learns the speech representations in self-supervised manner and identifies vocal emotion.

F. Text Sentiment Analysis (Text Modality)

The text modality analyzes the emotions expressed through spoken words.

1. Speech-to-Text: The Whisper model converts audio into multilingual text.

2. Text Emotion Classification: The transcribed text is analyzed using the DistilRoBERTa based model (j- hartmann/emotion-english-distilroberta-base), which performs context aware text emotion detection.

G. Multimodal Fusion

Once the facial, audio and text features are extracted they are combined to form a unified emotional understanding.

1. Feature Alignment: The outputs from each modality are converted into a unified representation to easily compared and integrated.

2. Attention Mechanism: The fusion layer use an attention mechanism to assign importance to each modality.

For example:

a. If the person's face is not visible the system Depends more on audio and text.

b. If there's background noise it focuses more on facial and linguistic cues.

4. Fusion Equation: The final combined feature representation is calculated as:

Ffusion=αfFf+αaFa+αtFt

Where

Each α value represents how much weight is given to facial, audio, and text information.

H. Emotion Classification

The fused vector is then passed through a dense layer followed by a SoftMax classifier. This classifier predicts the most likely emotion out of predefined categories such as happy, sad, angry, fearful, or neutral.

E=Softmax(W·Ffusion+b)

The output includes both the predicted emotion label and its confidence score.

3.1 Mathematical Model

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j\in\{F,A,T\}} \exp(e_j)}$$

- V -input video
- {It}-video frames sampled from V.
- a - audio from the video
- s-transcript (text) obtained from the audio (ASR)
- C- number of emotion classes like happy, sad, angry, neutral.
- F,A,T denote vectors (features) for face, audio, and text.
- Scalars αf, αa, αt- fusion weights foreach modality (sum to 1)
- p^- predicted class probabilities

1) Encode each modality

Turn raw inputs into compact feature vectors using pretrained or trainable encoders:

- F=Ef ({It}) (face features) (1)
- A=Ea(a) (audio features) (2)
- T=Et (s) (text features) (3)

Where each E∗ represents a modality-specific encoder:

- Ef: DeepFace (CNN-based)
- Ea: SpeechBrain (wav2vec2 transformer)
- Et: RoBERTa-based text emotion model

Each encoder outputs a vector summarizing emotion cues for that modality.

2) Make dimensions compatible (projection)

Project these vectors to the same size ddd so they can be combined:

$$uF=WFF+bF, uA=WAA+bA, uT=WTT+bT \quad (4)$$

W∗and b∗are small learned matrices/vectors that resize the features to the same length.

3). Compute simple attention weights

Compute how important each modality is for this input. A very simple choice.

1. score each modality (dot with a learnable vector v): ei=vtanh(ui) for i∈{F,A,T}

2. softmax to get normalized weights:

αi are numbers between 0 and 1 that tell the model how much to trust face, audio, or text for this sample.

4). Fuse modalities (weighted sum)

Combine the projected vectors using the attention weights:

$$H=\alpha FuF+\alpha AuA+\alpha TuT$$

H is the final vector that contain information from all the modalities.

5). Classify emotions

Turn H into class probabilities with a small classifier:

$$z=\mathrm{ReLU}(WhH+bh),\ p\hat{}=\mathrm{softmax}(Woz+bo)$$

Predicted emotion: y^=argmaxcP^2

softmax gives a probability for each emotion (happy, sad, angry, …)

6). Training loss

Train the network using cross entropy loss between predicted probabilities and true label y:

4. RESULT AND ANALYSIS

$$\mathcal{L} = -\sum_{c=1}^{C} y_c \log(\hat{p}_c)$$

The system achieves higher accuracy and reliability compared to unimodal models that rely only on facial, audio, or text data. By the combination of three modalities the framework can capture complex emotions that are often lost when a single source is analyzed.

The attention based fusion mechanism is designed to improve performance by randomly assigning higher importance to the most reliable modality in each instance. For example, when facial expressions are not there it mainly focuses on vocal and text. Facial Emotion Recognition

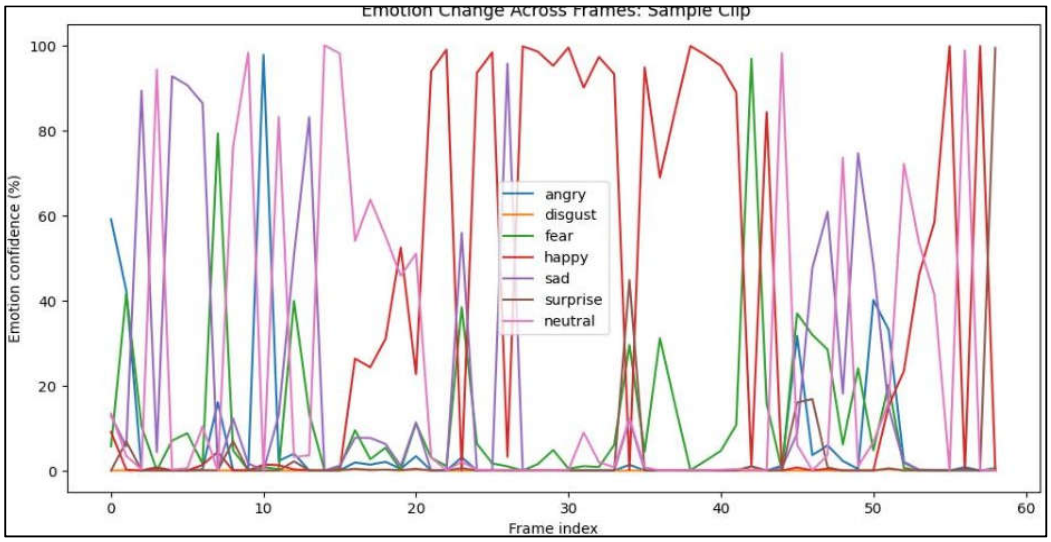

Figure 2. Facial emotion change for sample 1 per frame index

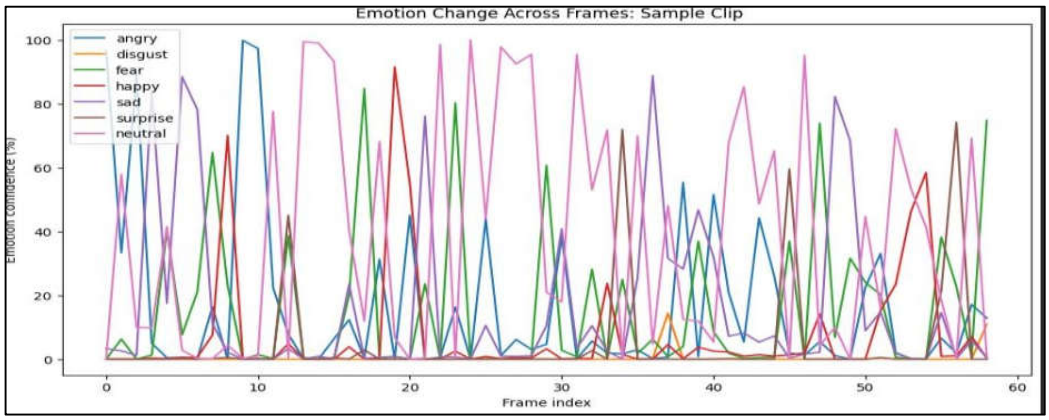

Figure 3. Facial Emotion Change for Sample 2 per frame index

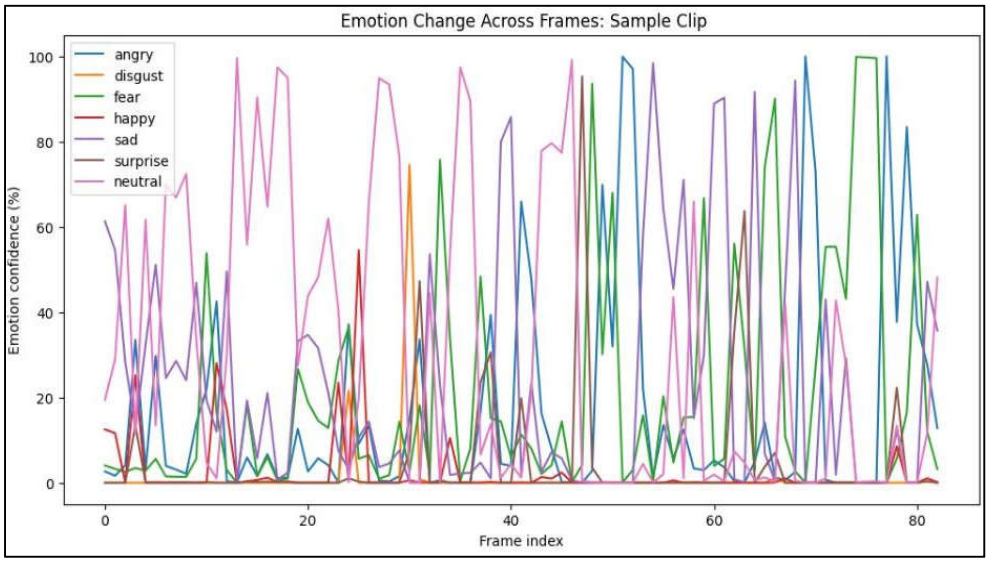

Figure 4. Facial Emotion Change for Sample 2 per frame index

Fig 2,3 4 shows the variation of emotion confidence across video frames for a sample clip. The x-axis represents frame numbers and the y-axis shows the model's confidence for each emotion detected by DeepFace. Each colored line denotes to one emotion angry, disgust, fear, happy, sad, surprise and neutral.

The plot illustrates how emotions changes throughout the clip for example, neutral dominates in early frames, while angry and fear show strong peaks later indicating dynamic emotional changes over time.

Audio Emotion output:

Clip 1:

Clip 0 — Vocal Emotion Result: Emotion: hap | Confidence: 1.00

Clip 2:

Clip 1 — Vocal Emotion Result: Emotion: hap | Confidence: 1.00

Text Emotion output:

Clip 1:

Detected Text Emotion: negative (Confidence: 0.40)

Clip 2:

Detected Text Emotion: neutral (Confidence: 0.45)

The attention-based fusion module proposed in this work has not yet been implemented in the current version of the system. At present the results are generated separately for each modality

facial, audio and text allowing individual evaluation of their performance. In the next stage, the fusion mechanism will be integrated to combine these modalities, assigning weights based on their reliability in each situation. This enhancement is expected to make emotion prediction more consistent, accurate and context aware.

5. CHALLENGES AND LIMITATIONS

Although its potential the framework may face several challenges:

- Data imbalance: Some emotions like fear or disgust may appear less frequently in datasets making training harder.
- Computational complexity: Combining three deep learning branches facial, audio, text can require high processing power.
- Cross-cultural variation: Emotional expression can be different among individuals and cultures which affects generalization.

6.CONCLUSION

This paper presents the conceptual design of an open-source framework for multimodal emotion recognition. The framework designs to strengthen human computer interaction by helping systems to better understand emotions expressed through facial cues, voice tone and language. By combining methods from computer vision, speech analysis and language processing the model aims to offer a deeper more human like understanding of emotional expression. Its architecture combines deep learning methods such as convolutional networks and transformer models.

Though the model some part remains in the conceptual stage it establishes a solid foundation for future research and real-world applications. The open-source design promotes collaboration, knowledge sharing and innovation within the research community. In the future the framework can be expanded with advanced data fusion techniques, real time emotion tracking and domain specific applications in areas such as drama performance, film production, mental health care and virtual communication. In conclusion this work represents a step toward developing emotionally intelligent technologies capable of engaging with people in a more natural, empathetic and meaningful way.

7. FUTURE SCOPE

The framework can be further expanded to recognize complex emotions like sarcasm, confusion or mixed feelings and measure the intensity of emotions. The system can also be used in practical applications such as film and drama analysis, virtual assistants and mental health monitoring systems. Expanding the dataset to include different cultures and languages will improve the systems fairness and usability. Future research could focus on optimizing the model for real time processing either on local devices or through cloud-based services. This will make it easier to access and scale for broader use.

REFERENCES

[1] Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... & Bengio, Y. (2013, November). Challenges in representation learning: A report on three machine learning contests. In International conference on neural information processing (pp. 117-124).Berlin, Heidelberg: Springer berlin heidelberg.

[2] Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1), 18-31.

[3] Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1), 18-31.

[4] Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1), 18-31.

[5] Yu, Z., & Zhang, C. (2015, November). Image based static facial expression recognition with multiple deep network learning. In Proceedings of the 2015 ACM on international conference on multimodal interaction (pp. 435-442).

[6] Baevski, A., Zhou, Y., Mohamed, A., & Auli,M. (2020). wav2vec 2.0: A framework for self- supervised learning of speech representations. Advances in neural information processing systems, 33, 12449-12460.

[7] Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS one, 13(5), e0196391.

[8] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh,

A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4), 335-359.

[9] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).

[10] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. Information fusion, 37, 98-125.

[11] Srivastava, D., Singh, A. K., & Tapaswi, M. How you feelin'? Learning Emotions and Mental States in Movie Scenes SUPPLEMENTARY MATERIAL.

[12] Mocanu, B., & Tapu, R. (2022, June). Audio- video fusion with double attention for multimodal emotion recognition. In 2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP) (pp. 1-5). IEEE.

[13] Georgescu, A. L., Chivu, G. I., & Cucu, H. (2024, October). Exploring Fusion Techniques for Multimodal Emotion Recognition. In 2024 15th International Conference on Communications (COMM) (pp. 1-6). IEEE.

[14] Maham, S., Tariq, A., Tayyaba, B., Saleem, B., & Farooq, M. H. (2024, December). MMER: Mid- Level Fusion Strategy for Multimodal Emotion Recognition using Speech and Video Data. In 2024 18th International Conference on Open Source Systems and Technologies (ICOSST) (pp. 1-6). IEEE.

[15] Gopalakrishnan, S., Eswar, R., Kishoor, T. K., Thamizhmaran, U., & Bharathan, M. (2025, May). Multimodal Emotion Recognition: An Integrated Approach using Facial, Audio and Text Analysis. In 2025 Third International Conference on Augmented Intelligence and Sustainable Systems (ICAISS) (pp. 122-127). IEEE.

[16] Williams, J., Kleinegesse, S., Comanescu, R., & Radu, O. (2018, July). Recognizing emotions in video using multimodal DNN feature fusion. In Grand Challenge and Workshop on Human Multimodal Language (pp. 11-19). Association for Computational Linguistics.

[17] Xie, Y. G., Zhou, N. N., & Zhu, S. Y. (2025).

Multimodal Emotion Recognition Based on Hierarchical Feature Fusion. Journal of Computers, 36(2), 281-296.

[18] Ortiz-Perez, D., Benavent-Lledo, M., Mulero- Pérez, D., Tomás, D., & Garcia-Rodriguez, J. (2024, June). Multimodal Fusion Strategies for Emotion Recognition. In 2024 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

[19] He, J. (2025, January). A Multimodal Approach for Emotion Recognition in Conversations Using the MELD Dataset. In 2025 Asia-Europe Conference on Cybersecurity, Internet of Things and Soft Computing (CITSC) (pp. 54-58). IEEE.

[20] Du, X., Yang, J., & Xie, X. (2023, February). Multimodal emotion recognition based on feature fusion and residual connection. In 2023 IEEE 2nd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA) (pp. 373-377). IEEE.

[21] Wei, Q., Zhou, Y., Xiang, S., Xiao, L., & Zhang, Y. (2024). MEAS: Multimodal Emotion Analysis System for Short Videos on Social Media Platforms. IEEE Transactions on Computational Social Systems.

[22] Ho, N. H., Yang, H. J., Kim, S. H., & Lee, G.(2020). Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network. IEEE Access, 8, 61672-61686.

[23] Chudasama, V., Kar, P., Gudmalwar, A., Shah, N., Wasnik, P., & Onoe, N. (2022). M2fnet: Multi- modal fusion network for emotion recognition in conversation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4652-4661).

[24] Luo, J., Phan, H., & Reiss, J. (2023, June). Cross-modal fusion techniques for utterance-level emotion recognition from text and speech. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

[25] Hsu, J. H., & Wu, C. H. (2023). Applying segment-level attention on bi-modal transformer encoder for audio-visual emotion recognition. IEEE Transactions on Affective Computing, 14(4), 3231-3243.

[26] Alrasheedy, M. N., Muniyandi, R. C., & Fauzi, F. (2022, October). Text-based emotion detection and applications: a literature review. In 2022 international conference on cyber resilience (ICCR) (pp. 1-9). IEEE.

[27] Udahemuka, G., Djouani, K., & Kurien, A. M. (2024). Multimodal Emotion Recognition using visual, vocal and Physiological Signals: a review. Applied Sciences, 14(17), 8071.

[28] Pandey, S., & Handoo, S. (2022, March). Facial emotion recognition using deep learning. In 2022 International Mobile and Embedded Technology Conference (MECON) (pp. 248-252). IEEE.

[29] Jakkani Shravani, Dr.V.Bapuji. “Machine Learning for News Group Text Classification”. Journal of Technology, ISSN. 10123407, Volume 12, Issue 5, 2024. Pages. 143-1440. https://journaloftechnology.org/volume-12-issue-5-2024/ https://drive.google.com/file/d/1RlKl1nNtzluWkRRlie-NHiQf4esZHatg/view

[30] Rangulov, D., & Fahim, M. (2020, December). Emotion recognition on large video dataset based on convolutional feature extractor and recurrent neural network. In 2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS) (pp. 14-20). IEEE.

[31] Kalateh, S., Estrada-Jimenez, L. A., Nikghadam-Hojjati, S., & Barata, J. (2024). A systematic review on multimodal emotion recognition: building blocks, current state, applications, and challenges. IEEE Access, 12, 103976-104019.