IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 5 June 2024, accepted 7 July 2024, date of publication 10 July 2024, date of current version 14 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3425953

RESEARCH ARTICLE

Multimodal Emotion Recognition Using Feature Fusion: An LLM-Based Approach

OMKUMAR CHANDRAUMAKANTHAM[1], **N. GOWTHAM**[1], **MOHAMMED ZAKARIAH**[2], **(Member, IEEE), AND ABDULAZIZ ALMAZYAD**[3]

[1]School of Computer Science and Engineering, Vellore Institute of Technology- Chennai Campus, Chennai 600127, India
[2]Department of Computer Sciences and Engineering, College of Applied Science, King Saud University, Riyadh 11543, Saudi Arabia
[3]Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

Corresponding author: Omkumar Chandraumakantham (omkumar.cu@vit.ac.in)

The research is funded by the Researchers Supporting Program at King Saud University, (RSPD2024R809).

ABSTRACT Multimodal emotion recognition is a developing field that analyzes emotions through various channels, mainly audio, video, and text. However, existing state-of-the-art systems focus on two to three modalities at the most, utilize traditional techniques, fail to consider emotional interplay, lack the scope to add more modalities, and aren't efficient in predicting emotions accurately. This research proposes a novel approach using rule-based systems to convert non-verbal cues to text, inspired by a limited prior attempt that lacked proper benchmarking. It achieves efficient multimodal emotion recognition by utilizing distilRoBERTa, a large language model fine-tuned with a combined textual representation of audio (such as loudness, spectral flux, MFCCs, pitch stability, and emphasis) and visual features (action units) extracted from videos. This approach is evaluated using the datasets RAVDESS and BAUM-1. It achieves high accuracy (93.18% in RAVDESS and 93.69% in BAUM-1) on both datasets, performing on par with the SOTA (state-of-the-art) systems, if not slightly better. Furthermore, the research highlights the potential for incorporating additional modalities by transforming them into text using rule-based systems and utilizing them to refine further pre-trained large language models, giving rise to a more comprehensive approach to emotion recognition.

INDEX TERMS Multimodal models, emotion recognition, large language models, feature extraction with rule-based systems, early fusion strategies.

I. INTRODUCTION

"Human sentiment in natural language is generally an intricate combination of emotions, which can sometimes be indeterminate, neutral, or ambiguous" [1]. Accurately deciphering and understanding these complex emotional states is most important in building meaningful connections, empathizing, and navigating social landscapes. However, the capacity to decode emotions is limited despite being remarkable. Subtle cues can vanish in the blink of an eye, cultural nuances are easily misconstrued, and personal biases can cloud judgment. This is where emotion recognition models emerge, offering a computational lens to understand the emotional undertones hidden within human expression.

However, traditional emotion recognition models aren't entirely accurate in their predictions since they depend on the verbal utterances while neglecting the non-verbal cues involved, such as facial expressions, vocal inflections, tones, etc. Even if non-verbal cues are utilized, the model usually includes only one of those modalities, generally audio [2], [3], [4], [5] or visual [6], [7], [8]. These modalities are not isolated entities but interdependent elements of a complex emotional state. Limiting emotional recognition to text is like understanding a symphony by listening to a single instrument, depriving it of richness, resulting in misinterpretations and missed nuances. The introduction of Multimodal Emotion Recognition (MER) [9] represents a

The associate editor coordinating the review of this manuscript and approving it for publication was Wenbing Zhao.

paradigm shift in emotion recognition, aiming to capture the complete interplay of human emotion by harnessing the power of multiple modalities. MER offers a more comprehensive understanding of emotional states by interconnecting information from facial expressions, speech, body language, and even physiological signals (such as heart rate and skin conductance). However, current MER systems focus on only 2 or 3 modalities, failing to capture the full spectrum of emotional cues present, leading to a lesser understanding of features. Some MER systems may involuntarily incorporate irrelevant features like gender or race into their classification process [6], which can lead to biased and inaccurate results that do not reflect the true emotional state of the individual. A few of the systems use computationally expensive and time-consuming methods [10], [11], potentially limiting their real-world feasibility. Most of the existing systems fail to take emotional interdependency into account, potentially missing out on the combined nuances [12], [13], [14]. This research intends to address these limitations and allow future advancements by providing a novel approach to solving emotion recognition with the following contributions:

- Introduction of rule-based logic specific to each modality (audio and visual) to convert the non-verbal features into a unified text format using predefined thresholds, which are further combined into prompts. The rule-based systems are at the core of this research, which facilitates the seamless integration of additional modalities in the future while maintaining consistency.
- Utilization of a pre-trained Large Language Model (LLM), fine-tuned on the verbal prompts generated to generate a user's emotion. This fine-tuning process leverages the LLM's ability to understand complex relationships within text data, leading to more accurate emotion recognition.

Simultaneously, this manuscript addresses the following questions vital to research:

- Does adding more modalities improve the performance of MER systems?
- Is it possible to create multimodal emotion recognition models with a unified method to utilize the different modalities while maintaining the scope to add more?
- Can the proposed system be benchmarked with the best publicly available datasets containing videos for emotion recognition? Is it better than the current state-of-the-art (SOTA) systems?

The research manuscript has been divided into multiple sections for better viewability. The research work is organized as follows: Section II provides a literature review of the traditional and SOTA systems perfected in the field of MER, resulting in the identification of the research gap and the creation of the problem statement. Section III introduces the proposed system with a detailed description of its architecture, workings, and the dataset utilized. Section IV provides information on the results obtained and the comparison of the benchmarks, along with their implications. Section V provides the conclusion along with the scope for future work in this research.

II. LITERATURE REVIEW

A. RELATED WORKS

Traditional models have typically focused on single modalities, such as video or audio, limiting their ability to capture the complexity of human emotions expressed through multiple channels. However, recent advancements have explored integrating multiple modalities to improve the accuracy and granularity of emotion recognition. This research expands the concept of multimodality by considering the textual domain as a distinct modality. Textual data, when analysed with large language models, offers a unique perspective as it inherently integrates aspects of other modalities, representing a promising new direction in emotion recognition.

1) UNIMODAL EMOTION RECOGNITION

Traditional emotion recognition methods, as shown in Tables 1 and 2, tend to prioritise a single modality, such as audio or visual cues, neglecting the potential benefits retrieved from complementary modalities, but they tend to be less complex in terms of implementation, making them much more reproducible. While data augmentation is present in some works, further development is necessary, particularly for handling real-world variations in data such as occlusions in images and noises in audio. Feature extraction methods often rely on handcrafted approaches like using HOG features in facial frames or the extraction of MFCCs from raw audio signals, which capture the best representations of such features (such as action units for facial expressions) but may not fully capture the temporal nature of emotions, while the usage of deep-learning approaches isn‘t completely efficient and may not capture the right set of features like the former approach. Most of the approaches only considered scenarios with limited data, leaving behind vulnerabilities in the computational demand and feasibility of the approach in real-time applications. Additionally, concerns persist regarding generalizability to diverse scenarios and the accuracy achieved in classifying specific emotions.

2) MULTIMODAL EMOTION RECOGNITION

Existing multimodal emotion recognition methods, as shown in Table 3, while leveraging mainly audio-visual cues, exhibit limitations. Modality-specific data augmentation techniques, though explored, require further development to handle real-world variations like noise and occlusions. Feature extraction, whether handcrafted or via deep learning architectures, might not fully capture the temporal nature of emotions, especially for large datasets. A few of the approaches use graphs, potentially allowing them to capture the dynamic nature of emotions. Most of the approaches focusing on visual modality use facial frames for training, which isn‘t the best method as there is a vulnerability to capturing unwanted features. At the same time, most approaches just

TABLE 1. Literature review on audio-based emotion recognition.

Ref.	Methodology	Pros	Cons
[2]	Audio features such as Mel Frequency Cepstral Coefficients (MFCCs) are extracted after applying windowing, data normalisation, and noise reduction using Librosa.	Audio features such as Mel Frequency Cepstral Coefficients (MFCCs) are extracted after applying windowing, data normalisation, and noise reduction using Librosa.	Audio features such as Mel Frequency Cepstral Coefficients (MFCCs) are extracted after applying windowing, data normalisation, and noise reduction using Librosa.
[3]	1. Variants of MFCCs (MFCC, delta MFCC, and del-delta MFCC) are extracted and stacked to make a 1D feature vector. 2. The feature vector is passed through a 4-layer Convolutional Neural Network (CNN) using gender-dependent training to get 16 classes (8 for each gender).	1. The proposed methodology performs better than baseline methodology (audio features like MFCCs, Mel spectrogram, chromogram, spectral contrast, and Tonnetz through a 6-layer CNN) in terms of predicting emotions like calm, happy, sad, and disgust, and similar performance in terms of surprise.	1. The proposed methodology performs better than baseline methodology (audio features like MFCCs, Mel spectrogram, chromogram, spectral contrast, and Tonnetz through a 6-layer CNN) in terms of predicting emotions like calm, happy, sad, and disgust, and similar performance in terms of surprise.
[4]	1. The Wav2vec2 model is used to obtain the masked representation of the input speech signal for latent space quantization to understand contextualized representations. 2. K-means clustering is applied to 13 MFCC features with their first and second-order differences to produce labels for the model using the HuBERT framework. 3. Linear Support Vector Machine (SVM) is used to streamline the set of features to classify emotions.	1. Satisfactory performance was achieved using HuBERT (Hidden unit BERT (Bidirectional Encoder Representations from Transformers)) X-large and it performs better than SOTA approaches. 2. The need for preprocessing raw audio signals becomes invalid as the framework takes care of it.	1. It uses only a single modality (audio). 2. Misclassifications between neutral and calm, sadness and calm, and happiness and surprise were observed. 3. The extraction of embedded features from the HuBERT and Wav2vec2 models is time-consuming, posing a significant computational burden. The features are also of considerable size. 4. The scope for adding more modalities isn't defined.
[5]	1. Audio is loaded using Librosa, augmented using noise injection, and cut and padded to make the recordings uniform. 2. Audio features such as MFCCs and Root Mean Square Energy (RMSE) are extracted with hop-length as 512 to get 157 t time-steps per recording. 3. SVM or 3-block CNN (two 1D convolutional layers, a ReLU activation function, followed by 1D pooling and dropout layers) for classification. 4. Best results were observed in the averaged probability voting strategy.	1. Consistent performance was observed across gender-specific data, indicating robustness and gender-independence. 2. It can perform well with multiple voices involved. 3. It has a lower complexity while being able to achieve satisfactory performance.	1. It uses only 2 classes: Disruptive (Anger, sadness, fear, disgust) and non-disruptive (happy, neutral, surprise, calm), which makes the classification less comprehensive. 2. Data augmentation with noise-corrupted data has almost no impact on model performance. 3. Vulnerability of SVMs in large-scale datasets will be observed as computational demands increase. 4. The study does not address multi-cultural, multi-lingual scenarios, and has not been tested in real-world deployments. 5. It uses a single modality (audio) and the scope for other adding modalities isn't defined.

TABLE 1. *(Continued.)* Literature review on audio-based emotion recognition.

use Mel spectrograms for audio classification, leaving behind potential characteristics such as MFCCs, spectral flux, etc. Additionally, performance inconsistencies across datasets and difficulties with specific emotion classifications raise concerns about generalizability. Most approaches use late fusion, which avoids the potential inter-modal interactions such as in early fusion. Furthermore, interpretability issues with techniques like VAEs and overreliance on facial features highlight the need for more robust and comprehensive approaches. Cross-dataset performance variance is also observed in many of the approaches.

3) TEXT-BASED SOLUTIONS IN EMOTION RECOGNITION

Existing text-based multimodal methods for emotion recognition (Table 4) have shortcomings, such as fine-tuning large language models (LLMs) for this task, which can be time-consuming and computationally expensive, and their performance can be inconsistent. Since the data used is limited, one-shot learning is a huge difficulty. The scope for adding modalities wasn't explored in most of the approaches. Their reliance on converting a few audio and visual features directly into text without capturing temporal dynamics using rule-based systems, although a new approach to solving emotion recognition problems with the scope of adding more modalities, may not capture the subtleties of emotional expression inherent in these modalities. Furthermore, a lack of proper benchmarking makes it difficult to assess their effectiveness compared to other methods, and they exhibit significant performance variations across datasets, raising

TABLE 2. Literature review on visual-based emotion recognition.

TABLE 2. *(Continued.)* Literature review on visual-based emotion recognition.

Ref.	Methodology	Pros	Cons
[6]	1. Preprocessing and data augmentation by rescaling, shifting, rotating, and cropping are done, and each of the pixels is normalized. 2. Visual Geometry Group Network (VGGNet) architecture (3 convolutional stages and 3 fully connected layers) is used to classify the emotions.	1. Satisfactory performance was observed in terms of accuracy. 2. Data augmentation performed was performed which led to an increase in robustness.	1. It is unimodal (just visual modality). 2. Usage of facial frames makes it vulnerable to learning unnecessary facial characteristics such as gender, race, etc., instead of the actual facial feature alignments. 3. Misclassification of anger as sadness and neutral, disgust as anger, fear as sadness, and sadness and neutral were observed. 4. It is not as good as the current SOTA as it doesn't involve better extraction of facial features. 5. The scope for adding other modalities isn't defined.
[7]	Introduction of PyFEAT 1. Facial feature detection of Action Units (AU) and emotions from images or video is accomplished using detection overlay. 2. Preprocessing is applied to align the images with the Histogram of Oriented Gradients (HOG) features and to extract multi-wavelet frequencies of each AU. 3. Analysis was done using statistical methods such as t-tests and inter-subject correlations. 4. Visualization to show vector fields and muscle heatmaps on the face was achieved.	1. It supports facial-landmarking, action unit detection, emotion detection, and head pose detection along with preprocessing and analysis. 2. It is open source. 3. Usage of real-world data collection to improve robustness while subjected to variations in luminance, occlusions of specific regions of the face, and also head rotation was observed. 4. Occlusion of specific facial structures provided consistent results for higher-level facial feature extraction such as action units. 5. AU detection on PyFEAT performs almost on par with other methods like OpenFace and FACET iMotions in terms of F1 score.	1. Shallow learning detectors that rely on HOG features are more dramatically impacted by high and low levels of variance of luminance. 2. Face detection substantially drops when the nose is removed. 3. The emotion models are even more dramatically affected by the occlusion of specific facial structures. Anger, fear, sadness, and surprise detection are substantially impacted by occlusion of the eyes, while disgust, happiness, and neutral detection drop when the mouth is blocked, and anger, disgust, fear, and sadness degrade with occlusions to the nose. 4. AU detection performance tends to decrease as head rotation angles increase. 5. The performance of emotion detection is average.
[8]	1. Facial features are extracted using HOG (to determine the alignment of facial features) or Scale-Invariant Feature Transform (SIFT-identification of local facial features) including jawline structure, winking eyes, and opened mouth. 2. These features are passed through a CNN with 3 convolutional layers, and a pooling layer, and they are passed through a fully connected layer of 7 neurons for classification.	1. Good performance was observed in terms of accuracy in both the variants (HOG-CNN and SIFT-CNN) and HOG-CNN performs better than SOTA models on both datasets (CK+ and Jaffe). 2. Usage of handcrafted features significantly improved the performance.	1. It is unimodal (Just visual modality). 2. Cross-dataset performance drop was observed in the SIFT-CNN variant indicating overfitting towards the CK+ dataset. 3. Misclassification of happy with surprise was observed in SIFT-CNN. 4. Handcrafted feature extraction is not viable and is time-consuming when the data involved is huge. 5. The scope for adding other modalities isn't defined.

concerns about generalizability and robustness in real-world scenarios with noise and variations.

4) UNIQUE ALTERNATIVES

Contrary to the conventional approaches taken, [10] proposes an approach that fuses electroencephalogram (EEG) signals along with speech modals implemented with Grey Wolf Optimisation (GWO), which is an innovative algorithm inspired by wolf pack hunting behaviour that efficiently selects the most relevant features from both audio and EEG data, focusing on those that contribute most to emotion recognition, and passes it through a CNN network to obtain excellent results despite using real-time data, making it much more robust. However, only three target classes were involved, and the extraction of EEG signals requires hardware like the Emotiv device. Reference [11] introduces various methods to extract emotional responses unconsciously by monitoring various methods using direct sensors such as EEG, EMG (electromyography), EOG (electrooculogram), modulating sensors like GSR (galvanic skin response), and measurements like RR (respiration rate) and HRV (heart rate variability). Even though these are unique alternatives to

TABLE 3. Literature review on multimodal emotion recognition.

TABLE 3. *(Continued.)* Literature review on multimodal emotion recognition.

Ref.	Methodology	Pros	Cons
[12]	1. Residual Networks 50s (ResNet50s) are used for the backbone architecture of both the audio and video network to get audio features from Mel Spectrograms and visual features from facial frames. 2. The features from both modalities are fused using a Fusion Network consisting of fully connected layers. 3. Correlation loss (based on HGR maximal correlation) is used to extract common information between the different modalities (visual features, audio features, and emotion labels), and classification loss is used to capture discriminative information from each modality (audio and visual features) for emotion prediction. 4. Further generalization is done using a semi-supervised learning scenario.	1. Better performance than SOTA models (handcrafted features and shallow DNN-based features) when singular modalities are considered as well as multimodality (with a significant growth in performance) is observed. 2. Significant performance improvement is observed when the computed correlation loss is considered for different weight coefficients, proving it to be stable in multimodal recognition.	1. Subpar performance is observed in singular modalities and no comparison with SOTA is given under the BAUM-1s dataset. 2. The proposed system finds it difficult to recognize fear and surprise in eNTERAFACE and RAVDESS due to less emotional information. Anger and fear have lower recognition accuracies on the BAUM-1s dataset due to class imbalances. There are slightly more misclassifications among the negative emotions across the dataset. 3. Real-time scenarios aren't considered as they include various noises and occlusions. 4. The scope to add other modalities could potentially be difficult if proceeded in the same fashion.
[13]	1. Facial landmark recognition is accomplished using Dlib to get 68 landmarks. They are converted into mm from pixel values using OpenCV, and Subnasale to soft tissue over Pogonion (Sn-sPog/Gn) distance is computed to relate between the jaw movement and various phoneme types to get video output files. 2. Audio feature extraction of the Mel spectrogram was done. Amplitude vs Time graph using Librosa was achieved after converting video to audio and trimming to match phonetic classes to get SnP output files.	1. It is an open-source tool. 2. It is a unique approach to classifying emotion using jaw alignment making it more robust to non-verbal cues closely related to verbal cues. 3. Usage of two modalities (audio-visual) was observed.	1. Usage of only a few AUs (jaw movements) doesn't make the detection very robust. 2. The scope for increasing more modals is difficult with this approach. 3. The absence of benchmarks on publicly available datasets to compare with other SOTA systems is observed.
[14]	1.1. Extraction of head pose using facial landmarking and an algorithm for head pose estimation is done. 1.2. Facial frames are extracted using Multi-Task Cascaded Convolutional Neural Networks (MTCNN). 2. Evaluation of emotion involved using Deep CNN such as VGG19 and ResNet50 models was observed. 3. User engagement was classified based on the emotion extracted and the head pose.	1. Computation of engagement along with emotion recognition using custom algorithms is a new approach. 2. Best performance is achieved in the combination of FER and CK + dataset.	1. Although it is bimodal, it potentially comes under the same category, i.e., facial features, leaving it oblivious to other models like audio and text. 2. Cross-dataset performance isn't consistent demonstrating a dramatic drop in taking the individual datasets. 3. Misclassification of sadness with fear and surprise is observed in both variants. Misclassification of anger with neutral and fear with sadness is observed in the ResNet50 variant.
[15]	1.1. EfficientFace layers (Except the classification layer) are utilized to extract facial features after facial frame detection using MTCNN. The frames are then passed through a 4-layer 1D CNN to reduce computational complexity and latency while capturing the temporal dynamics of facial features. 1.2. Audio features such as MFCC, Mel spectrogram, spectral contrast, and Tonnetz are extracted and stacked into a single vector. This vector is passed through another 1D CNN with 3 convolutional layers, 2 max-pooling layers, a dropout layer, and an adaptive average pooling layer to generate the relevant audio features. 2. The TFusion module (based on a transformer architecture using an attention mechanism) is used to fuse the audio and video features.	1. The multimodal variant proved to be much better than the unimodal variants in terms of performance. 2. The proposed system outperforms SOTA models in the RML dataset and is on par with RAVDESS. 3. Usage of TFusion provided robustness and the advantage of handling missing modalities. 4. The scope for adding additional modalities is possible considering the TFusion module.	1. The audio modality's performance is extremely poor when compared to the video modality due to the difficulty in distinguishing emotions based on audio signals. It often misclassified happiness and disgust. 2. The proposed system performed average on eNTERAFACE'05 and BAUM-1 datasets. 3. Slight misclassifications were observed such as sad as fearful, surprised as fearful, and disgust as angry and fearful.

TABLE 3. ***(Continued.)*** **Literature review on multimodal emotion recognition.**

TABLE 3. ***(Continued.)*** **Literature review on multimodal emotion recognition.**

	3. Fully connected layers are used to classify the emotions.		
[16]	Introduction of SMFNM (Semi-supervised Multimodal Fusion Network with Main-modal) 1. Semi-supervised Intra-modal Interaction (SII) module is implemented to extract intra-modality interactions using a semi-supervised end-to-end structure. 2. Considering text as the main modal and audio as the auxiliary modal, the Main-modal Cross-modal Interaction (MCI) module is used to implement cross-modal learning. 3. Multimodal fusion of three-utterance level based on the SII and MCI modules. 4. Usage of Directed Acyclic Graph (DAG) to extract conversational context based on the multimodal fusion features and the contextual information is fused. 5. Final emotional classification is done using a fully connected layer and Softmax layer.	1. Best performance is achieved when compared to the baseline models trained on MELD and IEMOCAP considered. 2. Intra-modal and cross-modal interactions are captured using SII and MCI. 3. DAG is used to grab conversational contextual information as the conventional ones can aggregate neighboring information from the previous layer but can't propagate it to the next layer. 4. The usage of two modalities (audio-text) is identified. The usage of text as the main modal and audio as the auxiliary modal resulted in an increase in performance rather than weighing them equally.	1. Happiness and sadness confusion in IEMOCAP (4-way) is observed. Confusion with happiness and excitement, and misclassification of sadness, neutrality, and anger as frustration in IEMOCAP (6-way) is discerned. Due to the underrepresentation of disgust and fear, there are misclassifications in those categories as well. 2. Environmental occlusions such as noise affect the performance, indicating the loss of robustness. 3. Great performance variance is observed in each emotion category in MELD. 4. Visual modality hasn't been utilized. 5. Difficulty in capturing the accurate conversational context was observed because SMFNM considers both speaker-sensitive and context-sensitive modalities, especially with multiple speakers in MELD.
[9]	1.1. Facial frames are extracted using the RGB channel. (3 frames of highest effect) 1.2. Mel spectrogram is extracted using Librosa and data augmentation (using pitch shift, and noise injection). (3 spectrograms) 2. Each segment is passed through 5 convolutional layers and 3 dense layers (2 for Mel) and combined to find results. (VGG-16-like model with early fusion applied) (Only the best model is considered)	1. Usage of properly benchmarked datasets resulting in efficient performance in lesser epochs (40) is observed. 2. Early fusion is used to capture interdependencies of the features involved. 3. Random sampling of test data is used to see at least one sample for each individual. 4. The usage of 2 modalities (audio-visual) is identified. 5. Usage of weight frame segmentation gets the most effective representation under video frames and Mel Spectrogram.	1. Only 5 emotions, namely sadness, happiness, fear, disgust, and anger, are used. 2. The dynamic nature of emotional phenomena is still not effectively reflected in the designed architectures. 3. Currently, it doesn't have a structure to channel a greater number of data modalities. 4. Cross-dataset performance isn't consistent demonstrating a dramatic drop in performance when training on one dataset and testing on the other one. 5. Facial frames, although a good method of classifying expressions have a vulnerability to involuntarily classify emotions about other unnecessary variables such as gender, race, etc.
[17]	1.1. Usage of MTCNN for facial detection and a CNN to predict the emotion involved is observed. 1.2. The Speech Emotion Recognition model is used to predict emotions from audio. 2. Fuzzy inference logic is created to define the overall emotional intensity using audio and video emotional intensity.	1. Overall emotional intensity and stability could be discerned using this new approach (fuzzy logic). 2. Usage of two models (audio-visual) is observed.	1. The absence of benchmarks on publicly available datasets and comparison with SOTA systems is observed. 2. Performance metrics (accuracy) are identified to be subpar. 3. Older and conventional approaches are used to predict emotions. 4. The difference between adult faces and children's faces might affect the system. Children of different ages might have varying emotional responses and cognitive abilities. 5. The scope of the current method is difficult to extend to other modalities involved.
[18]	Introduction of Multimodal Information Bottleneck Sentiment Analysis (MIBSA) 1.1. Textual features are extracted using the BERT encoder and the word embedding is considered as the utterance-level representation. 1.2. BiLSTM encoders with 128 hidden states are utilized to get the unimodal representations of audio and video. 2. Projection of the 3 unimodal	1. The information bottleneck is used to capture the most predictive factors and discard the irrelevant information for compression while maintaining information integrity. 2. Usage of three modalities (audio-text-visual) is observed. 3. It is on-par performance with SOTA models under the same datasets.	1. Facial frames, although a good method of classifying expressions have a vulnerability to involuntarily classify emotions about other unnecessary variables such as gender, race, etc. 2. Modality-specific data augmentation techniques aren't applied, especially for audio. 3. The scope of adding more modalities while converting low-level features into high-level representation is

TABLE 3. ***(Continued.)*** **Literature review on multimodal emotion recognition.**

	representations into two subspaces for consistency (shared among each modality) and specificity (private to each modality) decomposition by VAE-based encoders and concatenated is done. 3. The concatenated representation is passed through an encoder for learning the information bottleneck representation and a decoder to recover the unimodal representation.		still an interpretability problem.

emotion recognition, they aren't as economically feasible, accessible, or user-friendly as conventional approaches.

B. RESEARCH GAP ANALYSIS

Current emotion recognition methods struggle to capture the full picture. While some leverage multiple modalities (audio, visual, text), they often focus on individual modalities, ignoring the potential benefits of combined analysis. Data augmentation needs improvement to handle real-world variations like noise and occlusions. Feature extraction methods, both handcrafted and deep learning-based, might not fully grasp the temporal aspects of emotions, especially in large datasets. Existing approaches often rely on limited data, hindering generalizability and real-time application feasibility. Additionally, performance inconsistencies across datasets and difficulties with specific emotions highlight limitations. Multimodal approaches using audio-visual cues also have shortcomings. Both modality-specific data augmentation and feature extraction techniques might not fully capture the temporal dynamics of emotions. Most approaches use late fusion, neglecting potential inter-modal interactions. Techniques like VAEs lack interpretability, and overreliance on facial features raises concerns about robustness. Cross-dataset performance variance is another concern. Fine-tuning large language models used by text-based multimodal methods is computationally expensive and time-consuming, and performance can be inconsistent. Limited data hinders one-shot learning, and the scope for adding modalities is often unexplored. These methods might miss emotional subtleties by directly converting features into text without considering temporal dynamics. The lack of proper benchmarking makes it difficult to assess effectiveness, and significant

TABLE 4. **Literature review on text-based solutions in emotion recognition.**

Ref.	Methodology	Pros	Cons
[19]	1.1. Pitch and energy are extracted from the audio using MATLAB. 1.2. AUs are extracted using OpenFace and the AU is treated as 'appeared' only if it appears in over half of the total frames considered. 2. Non-textual modalities into text using rule-based systems and paragraphs are constructed with the textual representations of them with the utterance. 3. Usage of the paragraphs on discriminative LLMs like BERT and RoBERTa (for freezing and finetuning) and ChatGPT is observed.	1. Satisfactory performance is obtained by using all three modalities in the discriminative LLMs on a custom dataset. 2. Usage of a unique methodology that has the potential to cover a greater number of modalities is discerned. 3. Usage of natural communication in combination with modalities for the LLMs proved to be much more effective than using separators.	1. ChatGPT's sentiment prediction was subpar. 2. Benchmarking with other SOTA methods and publicly available datasets is absent. 3. Usage of only pitch and energy fluctuations isn't optimal as there are other audio characteristics such as MFCCs and spectral flux, which can potentially be used to detect emotions. 4. Working with BERT and RoBERTa is quite time-consuming when it comes to finetuning or predicting when compared to other neural networks.
[1]	1. Features are generated using PLMs like BERT, RoBERTa, MPNet, XLNet, and ELECTRA (with dot-product attention and sentence embedding). 2. Features are classified using intermediate layers. 3. Features are inferenced using Neutrosophic Iterative Neutral Clustering. 4. Cluster centers are evaluated using cosine distance to get a Refined Neutrosophic Set (REN).	1. Usage of multiple PLMs resulted in improved reproducibility. 2. Ability to cluster neutral emotions with minor undertones of other emotions as the other emotion is possible. 3. Sentiment classes with few samples are also separable from large clusters. 4. Neutral emotion forms a central cluster, negative emotions are grouped separately from positive emotions, showing that features of similar sentiments are more related to each other. 5. The main idea is quantifying the relation between different sentiments using their high	1. One-shot learning is a challenge since the datasets are small and finetuning LLMs with them is difficult. 2. Highly imbalanced data is challenging for clustering algorithms. 3. Only the text-based modality is considered. 4. While providing a different perspective to sentiment analysis tasks by understanding the human language better, it hasn't provided any proper benchmarks.

TABLE 4. ***(Continued.)*** **Literature review on text-based solutions in emotion recognition.**

		dimensional representation in the latent space.	
[20]	1. Preprocessing is done to extract information-rich semantic features from utterances using RoBERTa-Large (excellent zero-shot feature extraction capabilities). 2. Custom context filters are used for denoising when considering the semantic relevance and the informativeness of the utterances to get an adjacency matrix. 3. DAG is used to depict the flow of emotions in the conversation resulting in the context-embedded emotional features for the entire conversation. 4. Semantic and emotional features are corrected using gates and concatenated to give the final fused feature. 5. Classification is done using fully connected layers and the feature with the highest probability is considered.	1. Better performance than SOTA models concerning all 4 datasets involved in terms of f1 is observed. 2. Usage of context filters and GNN capabilities make the model more robust when compared to SOTA. 3. Directed graphs are used to manage the emotional flow rather than just the states as it is.	1. Weak performance was observed when trying to classify similar emotions such as happy and excited (positive), sadness, frustration, and anger (negative), and neutral misclassified as frustrated in IEMOCAP. Class imbalance in MELD causes poor performance in classifying disgust and fear, and misclassification of positive and negative emotional clusters. Underrepresentation of disgust classes leads to misclassification of disgust as anger and surprise being confused with the other emotions in Dailydialog. Least performance is observed in EmoryNLP with misclassifications between powerful and joy, peaceful and neutral, and powerful and peaceful. 2. Cross-dataset performance is very variant. 3. It is unimodal (Only text is used) and the scope of adding more modalities is difficult (Could add audio/video features as additional parameters to the context filters). 4. Data augmentation techniques aren't utilized.

performance variations across datasets raise concerns about generalizability and robustness in real-world scenarios. Recent approaches to exploring brain signals or physiological responses often require specialised hardware or lack user friendliness compared to conventional methods. There is a need for more robust and comprehensive approaches that effectively combine modalities, capture temporal dynamics, and achieve generalizability across datasets and real-world variations.

C. PROBLEM STATEMENT

The core challenge lies in developing robust, efficient, and generalizable multimodal emotion recognition techniques. These techniques should effectively capture the complexity of human emotions expressed through multiple channels, such as audio, video, and potentially text, while allowing scope to add many more modalities. However, they must also address the limitations of the literature reviewed. Real-world variations like noise and occlusions necessitate further development in data augmentation techniques. Feature extraction methods need to evolve beyond handcrafted approaches or limited deep learning architectures to capture the full spectrum of emotions, especially for large datasets. Additionally, methods should achieve consistent performance across diverse datasets and improve accuracy in classifying specific emotions. Furthermore, interpretability is crucial; techniques should allow us to understand how features contribute to emotion recognition. Ideal methods will be feasible and accessible for practical implementation, considering economic and user-friendliness aspects. Emotional recognition systems could be more robust and comprehensive by addressing these research gaps.

III. METHODOLOGY

This section consists of the proposed multimodal early-fusion-based LLM-based approach with novel rule-based systems for textual conversions.

A. PRELIMINARIES

The preliminaries provide the necessary context for a better understanding of the methodology by establishing the foundational concepts.

1) FACIAL FEATURES

Facial Action Units (AUs), as defined by the Facial Action Coding System (FACS), as mentioned in [21], are anatomically based descriptors corresponding to the contraction of specific facial muscles or muscle groups. These AUs provide a standardised and objective way to capture minute facial movements, including wrinkle formation, eyebrow raising, and lip corner depression. This detailed mapping allows researchers to deconstruct complex emotional expressions into their constituent AUs, enabling a deeper understanding of the relationship between facial muscle activity and underlying emotional states, as shown in Fig. 1.

2) AUDIO FEATURES

The audio features that are to be extracted when it comes to the aspect of emotion recognition are as follows:

a) *Equivalent sound level (equivalent loudness)* is a single numerical value representing the average sound pressure level over a defined period. It signifies the

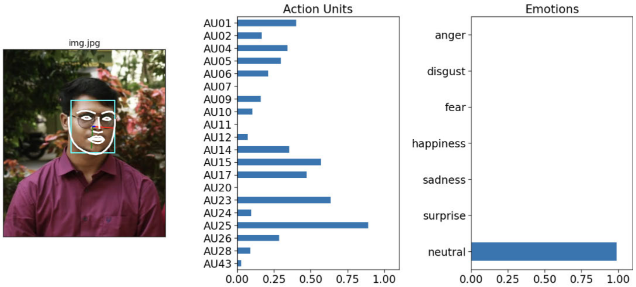


FIGURE 1. Action units detected by PyFEAT.

overall loudness of the audio, similar to how decibels (dB) measure sound intensity. It's a basic but effective indicator of emotional arousal.

b) *A semitone* is a musical term representing the smallest interval between two adjacent notes in a standard Western musical scale. It doesn't directly relate to an audio feature but describes the pitch difference between notes. However, in some audio analysis applications, semitones can be used to represent the change in fundamental frequency (F0) between consecutive sound frames. F0 is the perceived pitch of a sound, so analysing semitone changes can reveal how the pitch varies over time, from which you can determine pitch stability and pitch emphasis.

c) *Spectral flux* captures the rate of change in the frequency spectrum of an audio signal. The spectrum shows the distribution of energy across different frequencies at a specific time. Spectral flux indicates how quickly the sound's "colour" or timbre is changing.

d) *Mel-Frequency Cepstral Coefficients (MFCCs)* are a popular feature set derived from the audio signal's Mel-Frequency Cepstrum. The Mel scale approximates human auditory perception, where sounds with similar perceived pitch are grouped closer together. MFCCs capture the envelope of the sound spectrum on this Mel scale. They are particularly useful for representing the characteristics of speech and speaker identification because they are less sensitive to variations in pitch while encoding the formants (resonant frequencies) that contribute to the sound quality.

3) EASY DATA AUGMENTATION

Easy data augmentation (EDA) in text augmentation involves a set of simple yet powerful techniques to artificially increase the size and diversity of the training data. Some of the common EDA techniques are as follows:

a) *Synonym Replacement* replaces a word in the sentence with a synonym while aiming to preserve the meaning. For example, "The movie was terrible" could be augmented to "The movie was awful". There are various ways to find synonyms, including pre-built wordnet libraries like NLPAug or online thesauruses.

b) *Random Insertion* has a random word (or synonym) inserted into the sentence at a random location. This can introduce slight variations in sentence structure and vocabulary. For instance, the sentence "We went to the park" could become "We often went to the park". However, it's important to maintain grammatical coherence to avoid nonsensical sentences.

c) *Random Swap* involves swapping the order of two randomly chosen words within the sentence. This can introduce minor phrasal variations without significantly altering the meaning. An example could be "I love to eat pizza" becoming "I eat to love pizza" (although the latter might not be grammatically ideal for this specific case).

d) *Random Deletion* has a random word chosen and deleted from the sentence in this method. This can simulate natural speech disfluencies or typos. For example, the sentence "The quick brown fox jumps over the lazy dog" could be augmented to "The quick brown fox jumps over lazy dog". It's crucial to ensure the remaining sentence retains meaningfulness.

These EDA techniques are relatively easy to implement and can be effective in boosting the performance of Natural Language Processing (NLP) models, especially when dealing with limited datasets. They are particularly useful when combined with other augmentation strategies or more complex techniques.

4) MULTIMODALITY AND LLMS

Multimodal models are models that integrate information from multiple modalities (e.g., audio, visual) to improve performance. In the context of emotion recognition, multimodal models combine features from different modalities to capture complementary information and enhance the accuracy of emotion recognition. LLMs, or large language models, are essentially powerful machine learning models trained on massive amounts of text data. They can understand and generate human language, allowing them to perform a variety of tasks like translation, writing different kinds of creative content, and answering questions in an informative way. A pre-trained LLM, DistilRoBERTa (referenced in [22]), is used to understand the semantic meaning within the text in this context. DistilRoBERTa is a transformer-based model known for its effectiveness in various natural language processing tasks. DistilRoBERTa is a "distilled" version of RoBERTa-base. This means it retains the core functionality of the original model with significantly fewer parameters (around 82 million compared to 125 million for the RoBERTa base). This reduction in size translates to faster training and inference times.

5) FUSION STRATEGIES

Fusion strategies, the architects of multi-modal information, orchestrate the seamless integration of data from diverse sources to forge a more accurate and comprehensive representation. According to [16], these strategies fall into three primary categories, each wielding distinct strengths and weaknesses.

1. In *Early Fusion*, raw data or extracted features from different modalities are directly concatenated at the beginning of the processing pipeline. This forms a unified feature vector fed into a single processing unit. It can capture holistic inter-modality relationships and is computationally efficient for simpler tasks, but the increased dimensionality can lead to computational burden with complex data and can result in the potential loss of modality-specific information.
2. In *late fusion*, independent models are trained on each data modality separately. Their predictions are then aggregated at the end using techniques like weighted averaging or majority voting to produce a consolidated result. They leverage specialised models for each modality and are readily interpretable due to modularity. On the contrary, they are susceptible to inconsistencies between models and their decision boundaries and may not capture complex inter-modality relationships.
3. In *joint fusion*, data modalities are intertwined within the hidden layers of a single, powerful neural network. Instead of merging at the input, they interact and influence each other's learning process at various intermediate layers. They lead to superior and more nuanced representations by capturing complex inter-modality relationships and feature interactions, but they are computationally expensive, and model interpretability can be challenging due to their complex architecture.

B. LIMITATIONS OF PAST ARCHITECTURES AND ETHICAL IMPLICATIONS

Although PyFEAT proved to be a good open-source tool to depict facial landmarks, detect action units, etc., it only included visual modality, and it wasn't efficient when it came to emotion detection as it resulted in an average score of only 0.55 on the AffectNet dataset. Other alternatives to such approaches included (again, they only possessed a visual modality):

- Noldus FaceReader, which wasn't open source and only performed slightly better than PyFEAT, It didn't provide facial landmarking and wasn't robust in its performance variance towards various head poses.
- iMotions, which also wasn't open source (as it required the purchase of modules like AFFDEX and FACET), but it did provide a significant increase in efficiency from PyFEAT in terms of all the modules present, such as facial landmarking, action unit detection, emotion detection, head pose, and gaze detection.

On the contrary, considering speech emotion recognition, there weren't a lot of open-source tools with satisfactory efficiency found. However, OpenEAR [23] and Vokaturi stood out. OpenEAR is very old software but was able to capture satisfactory performance in recognising emotions on publicly available benchmarked datasets such as EmoDB and Enteraface'05. Vokaturi was also able to achieve good performance in real-time scenarios, as mentioned in [12], but had problems with gender independence. Moreover, there was no provision for adding visual modalities. The multimodal emotion recognition approach by [9] proved to be the best among other alternatives by providing excellent performance on benchmarked datasets such as RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) [24] and BAUM-1 (Bahçeşehir University Multimodal Affective Database-1) [25]. However, this approach faced difficulties such as cross-dataset performance variance, usage of facial frames, and a few audio features, and a classification variance of emotions was also observed. To tackle the problem of unifying the modalities, [19] presented a unique approach by introducing text-based representations of modalities by using rule-based systems. Still, it used only a few audio features, such as pitch and energy, and it didn't provide any performance benchmarks on publicly available datasets. The proposed architecture is taken to simplify the emotion recognition process while allowing interdependencies between emotions, presenting a methodology to unify the addition of more modalities into a text-based format using rule-based systems, benchmarking the performance on publicly available datasets such as RAVDESS and BAUM-1, and providing a less-complex yet customisable architecture.

Multimodal emotion recognition systems have some serious ethical issues because they can intrusively monitor and interpret people's emotional states. This raises a lot of concerns about how much personal and sensitive information is being collected, stored, and used. These technologies can be invasive, capturing data in places where people might not even realise they're being monitored, like in public spaces or during private moments. This kind of surveillance can make people feel like they're always being watched, which can mess with their sense of autonomy and discourage them from expressing themselves freely. Plus, the ways these systems get consent from users often aren't great. People might not fully understand what they're agreeing to, how much data is being collected, or how it could be used or misused. There's a real risk of this data being shared without permission, used to profile people, or even leading to discrimination based on emotional responses. This can make existing biases and inequalities even worse. To deal with these issues, there is a need for strong privacy protections, transparent data handling practices, and clear, informed consent protocols, making it important for developers, policymakers, and stakeholders to keep talking about these things to create guidelines and regulations that protect people's rights while still allowing users to benefit from these advanced technologies.

C. TOOLSET DESCRIPTION

1) PYFEAT

PyFEAT (mentioned in [7]) is a state-of-the-art deep-learning library that excels at extracting facial features, especially

AUs, from images and videos. PyFEAT's RetinaFace is considered in this research for fast and accurate face detection, even in challenging scenarios where lighting or pose variations might pose obstacles. Once faces are reliably detected, PyFEAT's MobileFaceNet model excels at locating key facial landmarks and pinpointing specific points like the corners of the lips, eyebrows, and eyes. This precise landmark detection assists in analysing subtle facial movements associated with different emotional states. Once the landmarks are detected, PyFEAT's SVM model is utilised to detect the 20 AUs based on the Facial Action Coding System (FACS). By analysing these AUs, PyFeat provides a highly detailed and accurate representation of facial features, crucial for precise emotion recognition. This combination creates a synergy that is both effective and computationally efficient, making it well-suited for real-time emotion recognition tasks.

2) LIBROSA

Librosa is a popular open-source Python library specifically focused on audio and music analysis. It provides a comprehensive set of functionalities for various audio processing tasks, including feature extraction, signal processing, audio manipulation, and music content analysis. This research utilises its audio-loading functionality to load audio files from various formats and decode them for further analysis.

3) OPENSMILE

OpenSMILE [26] is an open-source feature extraction system specifically designed for speech and audio processing. It focuses on extracting features relevant to human speech analysis, including prosodic features (related to intonation and rhythm), spectral features, and voice quality features. In this research, the eGeMAPSv02 feature set with functional mapping is applied to extract the necessary features with the best accuracy.

D. *DATASET DESCRIPTION*

The datasets utilised in this study are RAVDESS and BAUM-1, which offer two distinct sets of videos based on different intensity levels of expression: acted/strong or spontaneous/normal. These datasets provide representations of various emotional states, including anger, surprise, disgust, fear, happiness, and sadness. The primary objective associated with the dataset involves emotion classification and prediction tasks.

1) RAVDESS

RAVDESS comprises 7356 recordings from 24 professional actors (12 female and 12 male). Each actor vocalises two lexically identical statements in a neutral North American accent, expressing various emotions: calm, happy, sad, angry, fearful, surprise, and disgust (speech) and calm, happy, sad, angry, and fearful (song). Each expression is performed at normal and strong intensity levels, with an additional neutral rendering. Recordings encompass three modalities: audio-only, audio-video (full-AV), and video-only. For this research, the dataset focuses solely on the "full-AV" (audio and video) speech recordings, excluding the "calm" emotion. This subset comprises 1380 mp4 files where actors express happiness, sadness, anger, fear, and surprise. Each file is named using a standardised code that specifies:

- Modality: (01 = full-AV, 02 = video-only, 03 = audio-only)
- Vocal Channel: (01 = speech, 02 = song)
- Emotion: (01 = neutral, 02 = calm, 03 = happy, etc.)
- Intensity: (01 = normal, 02 = strong)
- Statement: (01 = "Kids are talking…", 02 = "Dogs are sitting…")
- Repetition: (01 = 1st, 02 = 2nd)
- Actor: (01-24, odd = male, even = female)

This refined dataset enables researchers to delve into the nuanced expression of emotions through both auditory and visual cues.

2) BAUM-1

- BAUM-1 contains 1184 multimodal facial video clips collected from 31 subjects, containing acted and spontaneous facial expressions and speech of 13 emotional and mental states such as happiness, anger, sadness, disgust, fear, surprise, boredom, contempt, confusion, neutrality, thinking, concentrating, and bothered. The dataset is issued with two annotation files, one acted and the other spontaneous, corresponding to each of the folders containing the sessions. The annotation files marked each session directory's video files (.mp4) and subtitle files (.srt) with a label among the 13 emotions. Only the video files corresponding to the 7 emotions as suggested in RAVDESS are considered. The files follow a specific naming convention, which consists of a 2-part numerical identifier (e.g., S001_005.mp4). These identifiers define the stimulus characteristics:
- Session (S001 to S031 for each of the 31 actors)
- File number (e.g., 005 for the 5th file)
- The annotation files contain the following columns:
- Number (Serial Number of each entry)
- Subject (Actor number)
- Clip (Video file number)
- Clip Name (Video file identifier)
- Emotion (Name of one of the 13 emotions involved)
- Emotion Code (Corresponding code of emotion)
- Gender (M = Male, F = Female)

3) UTILIZED DATASET

The study leverages a combined dataset of the RAVDESS and BAUM datasets, resulting in approximately 2,285 audio-visual recordings. Following pre-processing to ensure proper file path alignment with corresponding emotions, the audio data was processed using the openSMILE library. This audio processing step eliminates 51 entries due to unidentified audio segments (mean voiced segment length of 0).

The proposed approach uses a subset of features for further analysis, which include measures of loudness (equivalentSoundLevel_dBp), fundamental frequency stability (F0semitoneFrom27.5Hz_sma3nz_stddevNorm), and emphasis (F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2), spectral flux (rate of frequency content change) (spectralFlux_sma3_amean), and the first four Mel-frequency cepstral coefficients (mfcc1_sma3_amean, mfcc2_sma3_amean, mfcc3_sma3_amean, mfcc4_sma3_amean) which reflect spectral energy distribution patterns. Since the feature extraction of openSMILE and Librosa is excellent and has been perfected over the years, barely any limitations are found.

Simultaneously, facial feature extraction is performed on the video segments using PyFEAT. One video segment failed to render properly, resulting in 2,233 entries for facial analysis. PyFEAT identified the presence or intensity of 20 distinct action units (AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU11, AU12, AU14, AU15, AU17, AU20, AU23, AU24, AU25, AU26, AU28, and AU43) corresponding to various facial muscle movements. PyFEAT faces the major limitation of providing lesser performance in recognising the action units when compared with its alternatives such as OpenFace and FaceReader, but it is implemented since it's consistent over various head poses, the addition of occlusions, etc., while providing user-friendly APIs to use.

Additionally, after the features were converted into text by using a rule-based system based on various viable thresholds in regards to the dataset and both textual representations were combined into a prompt-like structure using string interpolation, text augmentation was applied to enhance the diversity and robustness of the data, and NLPAug was utilised to augment the prompts in the dataset using EDA, giving rise to a larger dataset with a much more enhanced vocabulary.

E. DESIGN ARCHITECTURE

The first and foremost procedure is the acquisition of MP4 files from RAVDESS and BAUM-1 datasets, vast repositories that encapsulate diverse emotional expressions captured in both audio and video. The data entries are merged to make it a generalised dataset. To further prepare this data for meaningful learning, a series of transformations are applied. Librosa is utilised to extract the audio data from the raw audio signals of the MP4 using the loading API. OpenSMILE was employed to generate all the necessary audio features. On the contrary, PyFEAT was utilised to detect and capture the action units (visual features).

With both audio and visual features extracted, the data undergoes a conversion to text, which involves the core of the system, i.e., rule-based systems based on various numerical thresholds. The textual representations of the audio and video features are combined into a prompt-like structure with their corresponding emotion labels, encoded for efficient processing. Additionally, the prompts are augmented using EDA to produce a bigger dataset with a higher vocabulary and a broadened scope. This augmented dataset is then carefully divided into three distinct groups: the training set, where the system learns its skills; the validation set, where it fine-tunes its abilities; and the testing set, where it demonstrates its true understanding of unseen data, ensuring robust evaluation and generalisation.

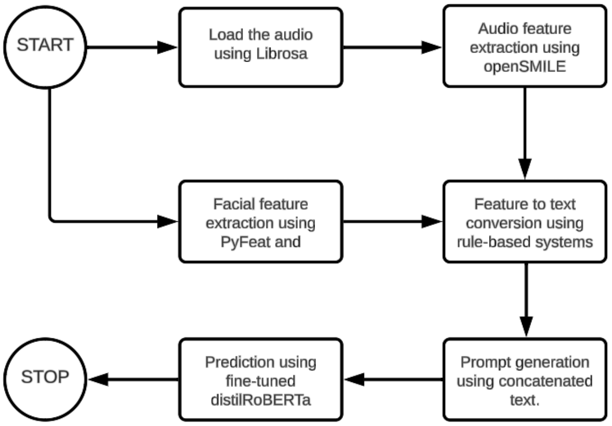

START
Load the audio using Librosa
Audio feature extraction using openSMILE
Facial feature extraction using PyFeat and
Feature to text conversion using rule-based systems
Prompt generation using concatenated text.
Prediction using fine-tuned distilRoBERTa
STOP

FIGURE 2. **Proposed design architecture.**

Then, the pre-trained Large Language Model (LLM), distilRoBERTa, is fine-tuned with all its layers trainable on this dataset using those prompts. This process allows distilRoBERTa to learn the intricate relationships between the textual prompts (encapsulating audio and video features) and their corresponding emotions. By fine-tuning all layers, the model leverages the dataset to improve its ability to recognise emotions from the combined audio and video data. This fine-tuned distilRoBERTa becomes a powerful multimodal model for emotion recognition. Fig. 2 shows the design architecture of the entire proposed system.

F. FEATURE EXTRACTION

1) FACIAL FEATURE EXTRACTION

For facial feature extraction (as shown in Table 5), PyFEAT's SVM model is employed to extract facial Action Units (AUs) from each frame of the facial videos. The extracted AUs are then cast into tensors for further processing. Additionally, average pooling is used to capture the AUs that occurred in more the half of the total frames involved in the video.

2) AUDIO FEATURE EXTRACTION

For audio feature extraction (as shown in Table 6), openSMILE is utilized to extract the comprehensive set of audio features from the MP4 files loaded using Librosa in the dataset. These features include:

- Loudness (Equivalent Sound Level)
- Pitch Variance (Semitone from 27.5Hz F0 frequency)
- Spectral Flux
- MFCCs (MFCC1, MFCC2, MFCC3 and MFCC4)

TABLE 5. Algorithm for facial feature extraction using PyFEAT.

```
Input
    •   Input Data: video_path
    •   Facial Detector: PyFEAT Detector (PF)
Output
    •   Output Data: facial_features
Initialization
    •   Facial Detector: PyFEAT Detector (PF)
            ○   face_model ← “retinaface”
            ○   landmark_model ← “mobilefacenet”
            ○   au_model ← “svm”
    •   Average Pooling Layer: Pool
Start Algorithm
Compute video ← PF.detect_video(video_path)
Let aus ← video.aus
Compute pooled_output ← Pool(aus)
Let facial_features ← If values over 0.5 in pooled_output Then set to 1
Else 0
End Algorithm
```

TABLE 6. Algorithm for audio feature extraction using opensmile.

```
Input:
    •   Input Data: audio_data, sample_rate (sr)
    •   Audio Processor: openSMILE (SMILE)
Output:
•   Output Data: audio_features
Initialization:
•   Audio Processor:  openSMILE.Smile (SMILE)
            ○   feature_set ← eGeMAPSv02,
            ○   feature_level ← Functionals
Start Algorithm
Compute all_features ← SMILE(audio_data, sr)
Let audio_features ← all_features[“equivalentSoundLevel_dBp”,
“F0semitoneFrom27.5Hz_sma3nz_stddevNorm”,
“F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2”,
 “spectralFlux_sma3_amean”, “mfcc1_sma3_amean”,
“mfcc2_sma3_amean”, “mfcc3_sma3_amean”, “mfcc4_sma3_amean”]
End Algorithm
```

3) SIGNIFICANCE OF THE EXTRACTED FEATURES

The accurate detection of emotions from human behaviour is a complex task, but by analysing the extracted features, significant progress was achieved. AUs, a standardised system for describing minute facial movements, provide a granular view of emotional expression. Each of the 20 AUs corresponds to a specific muscular movement, such as AU 6 (cheek raising), often associated with happiness; AU 4 (brow lowering), linked to anger; and AU 1 (eyebrow raising), potentially indicating surprise or fear.

On the audio side, the equivalentSoundLevel_dBp feature directly measures loudness, with higher values potentially indicating strong emotions like anger, surprise, or happiness. Conversely, lower volumes might suggest sadness or neutrality. F0semitoneFrom27.5Hz_sma3nz_stddevNorm and F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2 analyse the fundamental frequency (F0), which corresponds to the perceived pitch of a sound. The standard deviation of the semitone-converted F0 (F0semitoneFrom27.5Hz_sma3nz_stddevNorm) reflects pitch stability, with higher values potentially indicating vocal wavering often linked to emotions like anger or surprise. Conversely, lower values suggest a steadier pitch, potentially associated with sadness or neutrality. The F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2 feature focuses on the range of F0 values within a short time window, potentially revealing emphasis through stressed syllables or changes in speaking rate. Strong emphasis can be indicative of anger or surprise, while weaker emphasis might be associated with sadness or neutrality. SpectralFlux_sma3_amean captures the rate of change in the frequency spectrum of the audio data. This feature is particularly sensitive to rapid changes in sound characteristics, such as percussive events or sudden shifts in instrumentation, often observed during moments of fear or surprise. In contrast, lower spectral flux values are typically associated with calmer emotional states. Finally, MFCCs represented by features mfcc1_sma3_amean, mfcc2_sma3_amean, mfcc3_sma3_amean, and mfcc4_sma3_amean provide a more detailed analysis of the distribution of audio energy across the frequency spectrum. These features capture how much sound energy is present at different frequency bands. Generally, a distribution skewed towards higher frequencies might be associated with anger or frustration, while a distribution with more emphasis on lower frequencies could indicate happiness or neutrality.

TABLE 7. Action unit mapping to textual representations.

Action Unit (AU)	Description
AU1	Raises inner eyebrows
AU2	Raises outer eyebrows
AU4	Lowers brows
AU5	Raises upper eyelid
AU6	Raises cheeks
AU7	Tightens eyelids
AU9	Wrinkles nose
AU10	Raises upper lip
AU11	Deepens lines around nose and mouth
AU12	Pulls corners of lips outward (smile)
AU14	Forms dimples
AU15	Depresses corners of lips (frown)
AU17	Raises chin
AU20	Stretches lips
AU23	Tightens lips
AU24	Presses lips together
AU25	Parts lips slightly
AU26	Lowers jaw significantly
AU28	Sucks lips
AU43	Closes eyes completely

G. DESIGN IMPLEMENTATION

1) CONVERSION TO TEXTUAL REPRESENTATIONS

Under Facial Feature Representation, Table 7 details the mapping between Action Units (AUs) extracted by PyFEAT and their corresponding textual descriptions for facial expression analysis. This mapping allows for the creation of human-readable interpretations of facial movements within a dataset. For each frame analysed by PyFEAT, the corresponding AU values are examined. If a specific AU value is equal to one, indicating its activation, the corresponding textual description from the table is appended to a main string.

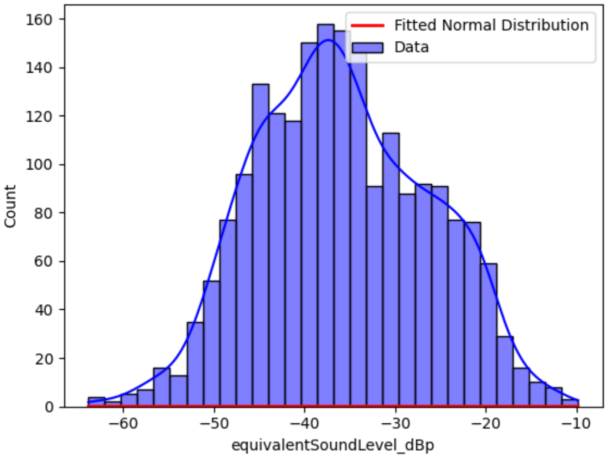


FIGURE 3. **Non-normal distribution of Loudness.**

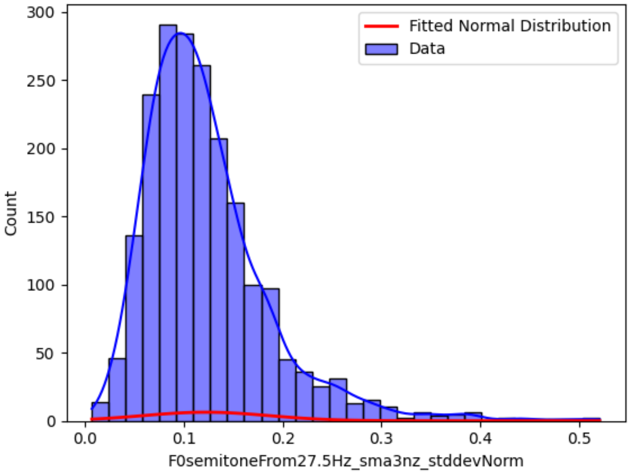


FIGURE 4. **Non-normal distribution of Pitch Stability.**

This process results in a textual representation that captures the active facial features within that frame. For example, if the action units AU1 and AU6 were recorded as one over half of the frames, then the main string would be "Raises inner eyebrows. Raises cheeks."

Under Audio Feature Representation, audio features extracted from an audio signal (loudness, pitch, spectral flux, and Mel-Frequency Cepstral Coefficients) undergo a rule-based conversion process (Table 8) to generate human-readable descriptions that are appended to a main string that marks the textual representation of those audio features based on certain bounds. To figure out those bounds, the research uses Chebyshev's inequality [27] of non-normal distributions (as shown in Fig. 3, 4, 5, and 6) of the audio features other than MFCCs to estimate that at least 50% of the entries lie beyond the thresholds at a constant value of the square root of 2. Subsequently, the lower and higher thresholds are obtained using the formulas given below:

$$\text{higher\_threshold} = \text{mean} + \sqrt{2}\text{x std}$$

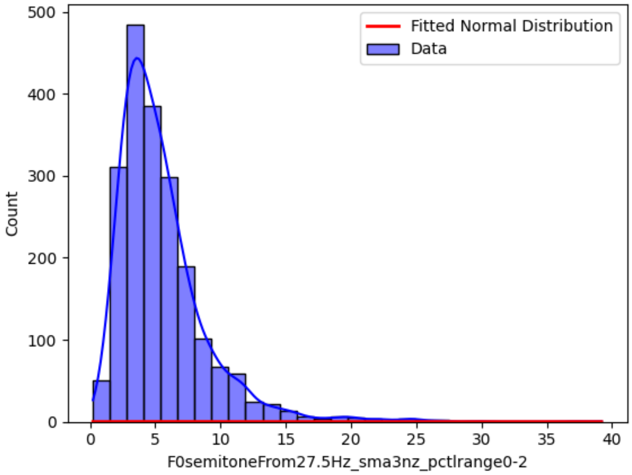


FIGURE 5. **Non-normal distribution of Pitch Emphasis.**

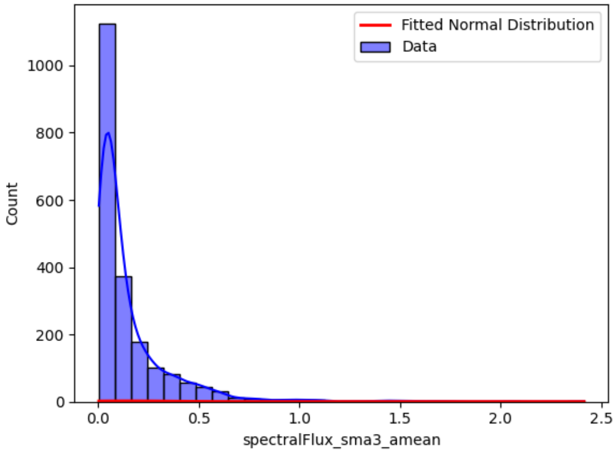


FIGURE 6. **Non-normal distribution of Spectral Flux.**

$$\text{lower\_threshold} = \text{mean} - -\sqrt{2}\text{x std}$$

where std is the standard deviation of the non-normally distributed data.

However, spectral graphs are used to visualise the energy distribution with the different scenarios about the 4 MFCC values to convert them to text, which are mentioned below:

When mfcc1 is greater than mfcc2 and mfcc3 is greater than mfcc4, the energy distribution is dominant at the lower frequencies, as shown in Fig. 7.

When mfcc2 is greater than mfcc1 and mfcc3 is greater than mfcc4, the energy is distributed across different frequencies, with more emphasis on the lower-mid frequencies, as shown in Fig. 8.

When mfcc3 is greater than mfcc1 and mfcc2, the energy distribution is observed more towards the higher frequencies, as shown in Fig. 9.

In the rest of the cases, energy is comparatively less distinct, as shown in Fig. 10.

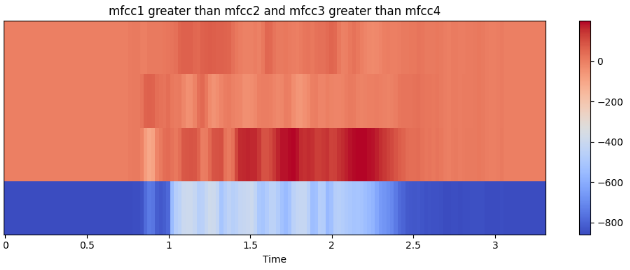


FIGURE 7. Energy distribution when mfcc1 > mfcc2 and mfcc3 > mfcc4.

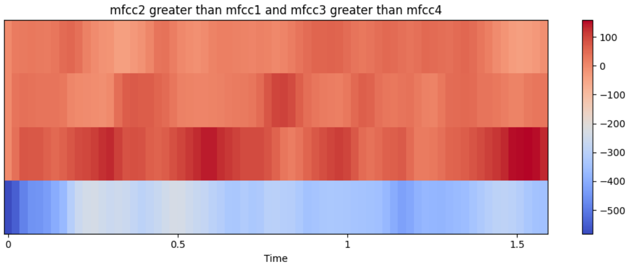


FIGURE 8. Energy distribution when mfcc1 < mfcc2 and mfcc3 < mfcc4.

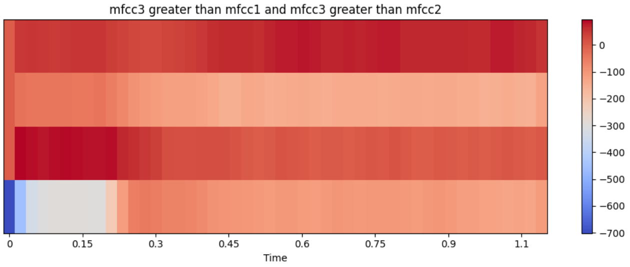


FIGURE 9. Energy distribution when mfcc3 > mfcc2 and mfcc3 > mfcc1.

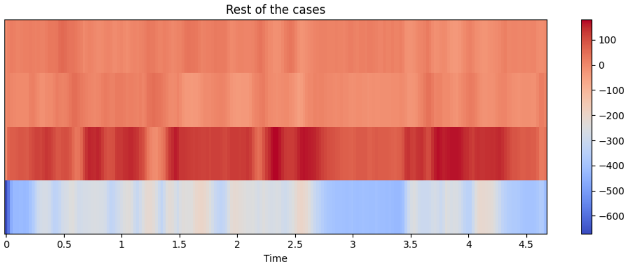


FIGURE 10. Energy distribution in the rest of the cases.

TABLE 8. Audio features mapping to textual representations.

Feature	Thresholds	Textual Descriptions
Loudness	Greater than -30	The speaker has a loud overall volume.
	Between -50 and -30	The speaker has a moderate overall volume.
	Lesser than -50	The speaker has a quiet overall volume.
Pitch Stability	Greater than 0.5	The speaker has an unstable pitch.
	Between 0.18 and 0.5	The speaker has a moderately stable pitch.
	Lesser than 0.18	The speaker has a relatively stable pitch.
Pitch Emphasis	Greater than 5	The speaker emphasizes strongly.
	Between 2 and 5	The speaker emphasizes moderately.
	Less than 2	The speaker emphasizes weakly.
Spectral Flux	Greater than 0.5	The audio has highly dynamic frequency content.
	Between 0.1 and 0.5	The audio has moderately dynamic frequency content.
	Less than 0.1	The audio has relatively stable frequency content.
MFCCs	mfcc1 greater than mfcc2 and mfcc3 greater than mfcc4	The audio has a dominant spectral peak in the lower frequencies.
	mfcc2 greater than mfcc1 and mfcc3 greater than mfcc4	The audio has its energy distributed across various frequencies, with a slight emphasis on the lower-mid frequencies.
	mfcc3 greater than mfcc1 and mfcc3 greater than mfcc2	The audio has its energy distributed across various frequencies, with a potential emphasis on the higher frequencies.
	The rest of the case	The audio has a less distinct spectral distribution pattern.

2) TEXT AUGMENTATION

To enrich the dataset and improve model robustness, this approach utilises Natural Language Augmentation (NLA) techniques. Specifically, the textual descriptions of both facial expressions and audio features undergo augmentation using the Synonym Augmentation function within the NLPAug library. This function leverages WordNet, a lexical database of English, to identify synonyms for the existing words. It then generates five variations of the original text, where each variation replaces more than five words with their synonyms while preserving the overall meaning. As a result, this process creates 25 (5 variations for facial features and 5 variations for audio features) distinct combinations of augmented text descriptions for every single instance of facial and audio features within the dataset. This strategy effectively expands the dataset size 25 times and introduces variations in how the same information is expressed, which can ultimately lead to a more robust model that generalises better to unseen data.

3) EARLY FUSION AND PROMPT GENERATION

The system leverages the generated textual descriptions of both facial expressions and audio features to create prompts for further analysis using early fusion. These descriptions, potentially augmented using synonyms to increase variation, are combined into a single prompt following a specific format.

$$\begin{aligned}\text{Prompt} = {}& \text{“audio :}'' + \text{audio\_text} + \text{“facial :}'' + \text{facial\_text} \\ & + \text{“} < |\text{endoftext}| >'' \end{aligned}$$

where audio_text is the textual representation of audio features after augmentation and facial_text is the textual representation of facial features after augmentation, The prompt structure utilises "audio_text" and "facial_text" placeholders, which are replaced with the actual descriptions of the sounds and expressions observed in that specific instance. This approach allows for the creation of a rich and informative prompt that captures the interplay between the visual and auditory aspects of the data. The limitations of the early fusion technique only become a viable threat if the number of dimensions is large, which is not a problem here on the fusion angle since the features are converted to text and just concatenated with each other.

4) MODEL FINE-TUNING

The pre-trained text-processing model, DistilRoBERTa [22], is used to understand the semantic meaning within the text. The AutoTokenizer function from the Transformers library [28] is used to perform this tokenization based on the pre-trained DistilRoBERTa model's vocabulary. The code employs padding and truncation techniques during tokenization. The padding ensures that all input sequences have the same length, which is necessary for efficient batch processing during model training. Truncation addresses situations where text entries exceed a predefined maximum length. In such cases, the tokenizer shortens the text by removing characters from the end while attempting to preserve the core meaning. Following tokenization, the code creates separate datasets for training, validation, and testing. Each dataset is a collection of text inputs (tokenized text) paired with their corresponding labels (the categories the model needs to predict). These datasets are further divided into batches for efficient training.

This research leverages a pre-trained DistilRoBERTa model working on 7 classes (as there are 7 emotions) specifically designed for sequence classification tasks. Pre-trained models offer a significant advantage as they have already been trained on massive amounts of text data, allowing them to capture complex language patterns. Fine-tuning a pre-trained model for a specific task involves adjusting its internal parameters to adapt to the characteristics of the new dataset and classification problem. The model training process involves an optimisation algorithm to find the best configuration of the model's internal parameters that minimises the difference between the model's predictions and the actual labels (ground truth). The Adam optimizer with a learning rate of 2e-5 is used for this purpose, along with a sparse categorical cross-entropy loss function. The loss function measures the discrepancy between the model's predicted probability distribution for each class and the one-hot encoded ground truth label. To assess the model's performance during training, the code monitors both the training and validation losses. The validation data provides an unbiased estimate of how well the model generalises to unseen data. Early stopping, a regularisation technique, at a patience of 5 is implemented to prevent overfitting. Overfitting occurs when the model memorises the training data too well and performs poorly on unseen data. Early stopping monitors the validation loss and halts training if it stops improving for a predefined number of epochs (training cycles). This helps ensure the model learns robust patterns that generalise well to new data.

IV. RESULTS AND DISCUSSION

A. EXPERIMENTAL SETUP

The empirical analysis of the fine-tuned model, incorporating feature extraction and early-fusion methods, was conducted using Python in a Windows 11 environment, supported by a 16GB RAM system with a Nvidia RTX 4060 series GPU. The experiments were executed with Python version 3.10 and used the TensorFlow framework. However, the minimum requirements to create this system (especially, finetune distilRoBERTa) are having a modern multicore processor, at least 8GB of RAM (preferably, 16GB), and a GPU with a computing power of at least as much as the Nvidia RTX 3060.

B. EVALUATION AND IMPLICATIONS

The proposed LLM-based approach is evaluated on two benchmark datasets for emotional expression recognition: RAVDESS and BAUM-1. Generally, most of the other works utilised metrics such as accuracy, f1 score, and confusion matrix to determine the performance of their model. This research considers the following various metrics to comprehensively assess the model's performance:

Accuracy: To find the overall correctness of the classification of emotions.

Precision: To find the proportion of correctly classified instances of each emotion.

Recall: To find out if the model can identify all instances of a particular emotion.

F1-Score: To find the reliability of the model in predicting each instance of emotion.

Error Rate Graph, Accuracy Rate Graph, and Loss Rate Graph: To find if the model is overfitting as the number of epochs increases.

Confusion Matrix: To find the proportions of what the model is classifying each instance of emotion as.

Classification Report: Acts as a condensed report for the different evaluation metrics.

The error rate graph (Fig. 11) found in the validation set displays a continuous decrease in the misclassification of emotions, which signifies that the model isn't overfitting. The results, including classification reports (Tables 9 and 10) and confusion matrices (Fig. 12 and 13), provide valuable insights into the model's strengths and weaknesses.

The model achieves impressive and consistent overall accuracy on both datasets, reaching 93.18% for RAVDESS and 93.69% for BAUM-1 after only running it for 18 epochs. However, a closer look reveals dataset-specific challenges. The RAVDESS dataset presents difficulties in accurately detecting the neutral category. This could be attributed to inherent ambiguity in the expression of neutrality or potential imbalances in the dataset. Conversely, emotions such as

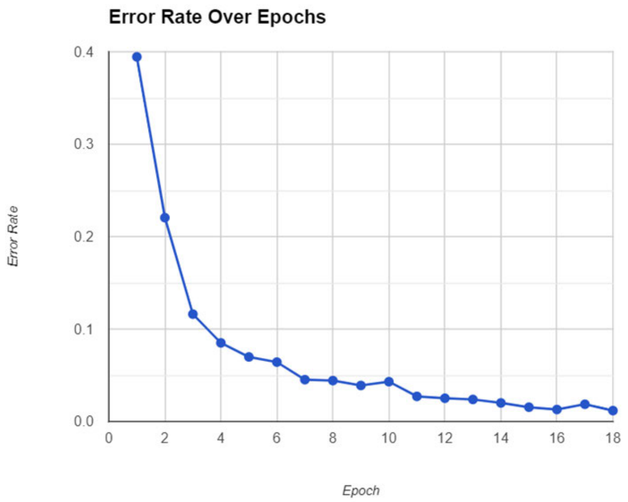


FIGURE 11. Error rate of the proposed model over 18 epochs on the validation data on the proposed LLM-based approach.

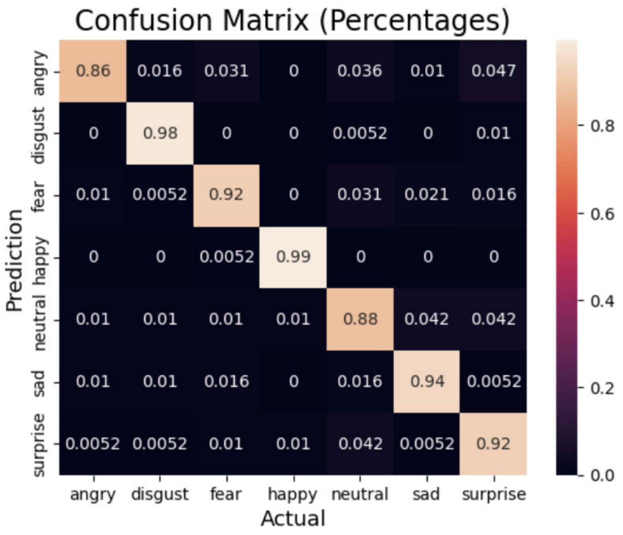


FIGURE 12. Confusion matrix on using RAVDESS on the proposed LLM-based approach.

disgust, happiness, and sadness are classified with near-perfect precision, recall, and an F1-score exceeding 94%. For emotions like anger, fear, and surprise, the results deviate from the high scores observed for other categories. This suggests potential ambiguity within the extracted features themselves. These emotions often involve exclamatory expressions, which might share some characteristics, leading to classification confusion. This observation highlights the need for further investigation into data augmentation techniques or alternative feature representations that could enhance the model's ability to recognise neutral emotions and differentiate between emotions with potentially overlapping expressions. However, a closer look at the ‘‘neutral’’ category reveals a dip in scores, hinting at potential ambiguity in its expression or possible imbalances within the dataset itself. This observation sparks questions for further investigation, prompting exploration into data augmentation techniques or alternative representations that could bolster the recognition of neutral emotions. There could be a potential avenue with the introduction of RENS [1], which could enhance the latency of subtle nuances in emotional states, thereby refining them.

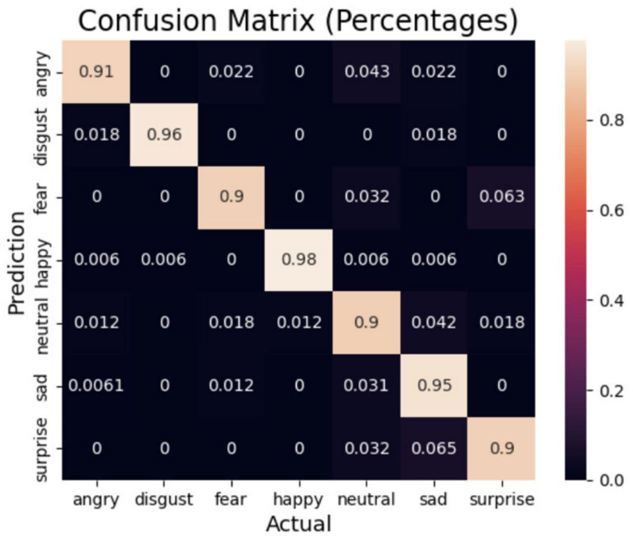


FIGURE 13. Confusion matrix on using BAUM-1 on the proposed LLM-based approach.

The analysis of the BAUM-1 dataset yielded similar findings. While overall scores have improved when compared to RAVDESS, particularly for disgust detection, anger recognition shows a slight decline. The neutral emotion category remains a challenge, suggesting that the model, while generally robust, might benefit from further refinement. Moving beyond raw accuracy, a deeper understanding of the model's behaviour can be understood by analysing the confusion matrices. These matrices reveal specific patterns of misclassifications, providing valuable insights into the complexities of human emotional expression. Interestingly, both datasets exhibit occasional misclassifications of ‘‘anger’’ as ‘‘surprise.’’ This confusion could be due to a potential overlap in the exclamative nature of these emotions. Additionally, misclassifications between ‘‘sad’’ and ‘‘fear’’ are also observed. This could be attributed to the fleeting nature of surprise expressions or cultural variations in how these emotions are manifested. These observations highlight the need for further investigation into data augmentation techniques or alternative feature representations that could enhance the model's ability to recognise emotional undercurrents and differentiate between emotions with potentially overlapping expressions.

C. COMPARISON WITH STATE-OF-THE-ART SYSTEMS

The proposed multimodal LLM-based emotion recognition model demonstrates promising results on both the RAVDESS and BAUM-1 datasets, achieving high accuracy scores. Accuracy is chosen as the metric in this scenario as it

TABLE 9. Evaluation metrics on using RAVDESS on the proposed LLM-Based approach.

	Precision	Recall	F1-Score	Support
Angry	0.96	0.86	0.91	192
Disgust	0.96	0.98	0.87	192
Fear	0.93	0.92	0.92	192
Happy	0.98	0.99	0.99	192
Neutral	0.77	0.88	0.82	96
Sad	0.94	0.94	0.94	192
Surprise	0.90	0.92	0.91	192
Accuracy	0.93			1248
Macro Average	0.92	0.93	0.92	1248
Weighted Average	0.93	0.93	0.93	1248

TABLE 10. Evaluation metrics on using BAUM-1 on the proposed LLM-Based approach.

	Precision	Recall	F1-Score	Support
Angry	0.93	0.91	0.92	92
Disgust	0.99	0.96	0.98	110
Fear	0.89	0.90	0.90	63
Happy	0.99	0.98	0.98	167
Neutral	0.92	0.90	0.91	167
Sad	0.92	0.95	0.93	163
Surprise	0.80	0.90	0.85	31
Accuracy	0.94			793
Macro Average	0.93	0.91	0.92	92
Weighted Average	0.99	0.96	0.98	110

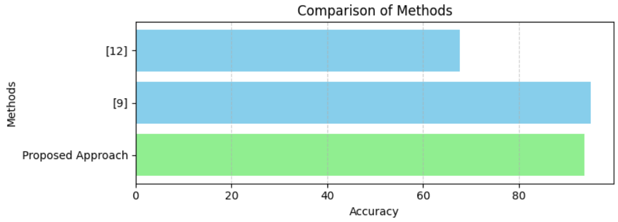


FIGURE 14. Graphical comparison with SOTA on BAUM-1.

signifies the overall performance of the model, and most of the SOTA tools display accuracy as the benchmark on the datasets. However, a thorough evaluation necessitates a deeper examination that acknowledges the nuances of various comparisons, as shown in Figs. 14 and 15, and Tables 11 and 12, and explores potential avenues for further improvement.

When compared to existing literature, the proposed model exhibits clear advantages. Reference [19] presents a similar approach, but the proposed model surpasses their performance in terms of F1 score while providing more audio features. However, a direct comparison is limited due to the lack of identical datasets in both studies.

Reference [9] employs a 2-input model utilising Mel-frequency cepstral coefficient features alongside facial features (the best model). While their model achieves an accuracy of 75.58% at the 10th epoch (training iteration), the proposed model reaches a significantly higher accuracy of 93.69% on the BAUM-1 dataset after only 18 epochs. Additionally, the proposed model performed competitively with the model of [9] (94.16%) at the 20th epoch, despite requiring fewer training iterations. Similarly, on the RAVDESS dataset, the proposed model achieves an accuracy of 93.18%, comparable to the model's performance at the 20th epoch (93.44%) in [9]. Reference [12] presents a simpler CNN-based approach that achieves an accuracy of 97.57% on RAVDESS but only 67.59% on BAUM-1. This significant disparity suggests potential overfitting towards the RAVDESS dataset. While the proposed model surpasses the approach of [12] on BAUM-1, it exhibits slightly lower performance on RAVDESS. This highlights the importance of addressing dataset-specific challenges and potential biases. Reference [5] proposes a model utilising an average probability ensemble with a 1D CNN architecture, achieving an accuracy of 92% on a dataset with disruptive and non-disruptive emotion labels. Notably, the proposed model outperforms this approach (92%), despite incorporating separate emotion classes and leveraging the video modality within the RAVDESS dataset. Finally, [15] presents a tensor fusion network approach that achieved an accuracy of 82.99% on RAVDESS. The proposed model demonstrates superior performance compared to this approach, further solidifying its effectiveness.

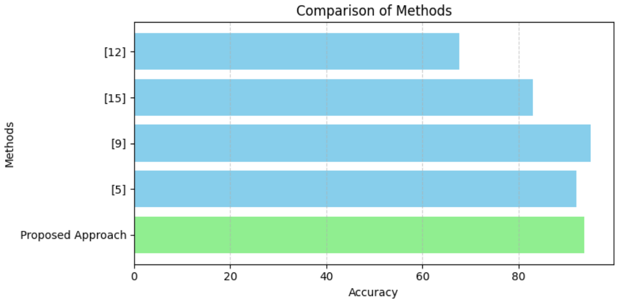


FIGURE 15. Graphical comparison with SOTA on RAVDESS.

TABLE 11. Tabular comparison with sota on BAUM-1.

Methods	Year	Accuracy	No. of Labels	Modals
[12]	2020	67.59	6	Audio-Visual
[9]	2023	95.00	5	Audio-Visual
Proposed Approach	**2024**	**93.69**	**7**	**Audio-Visual**

TABLE 12. Comparison with sota on ravdess.

Methods	Year	Accuracy	No. of Labels	Modals
[12]	2020	67.59	6	Audio-Visual
[15]	2023	82.99	7	Audio-Visual
[9]	2023	95.00	5	Audio-Visual
[5]	2024	92.00	2	Audio
Proposed Approach	**2024**	**93.69**	**7**	**Audio-Visual**

Moving forward, several key considerations will guide further development and refinement of the proposed model. First, normalising accuracy values across various models is crucial due to potential discrepancies in the employed label sets. This normalisation will enable more meaningful comparisons across different studies. Second, exploring various fusion approaches for integrating audio and visual features offers an opportunity to gain a broader perspective on potential trade-offs between accuracy and model complexity. Third, enhancing the model's interpretability is essential. Techniques that elucidate how the model integrates and interprets individual modalities (audio and visual features) will provide valuable insights and facilitate further optimization. Finally, rigorous evaluation of diverse datasets and cultural contexts is paramount for establishing the model's generalizability and real-world applicability. By addressing these aspects, the proposed model can solidify its position as a leader in the field of multimodal sentiment analysis, not only achieving high accuracy but also offering interpretability and the potential for real-world applications.

V. CONCLUSION AND FUTURE WORK

The proposed LLM-based model leverages early fusion, combining facial features (Action Units) and audio features (loudness, pitch variance, spectral flux, and MFCCs) after a rule-based conversion to a human-understandable textual representation. This approach achieves state-of-the-art performance on the BAUM-1 (93.69%) and RAVDESS (93.18%) datasets, demonstrating its effectiveness in emotion recognition. Beyond accuracy, the model offers advantages over prior methods that rely solely on facial frames or Mel spectrograms. Although this approach falls behind [9] in terms of accuracy, it presents a whole new innovative approach to solving emotion recognition while maintaining cross-dataset performance and having scope to add more modalities, thereby proving that:

- An increase in the number of modalities mostly increases performance while interconnecting multiple cues that complement each other to form the rich nature of emotions.
- By converting it to textual representations using rule-based systems, it presents a unified approach to managing modalities while providing scope to add more modalities.
- The proposed system is on par with the current SOTA systems after benchmarking on publicly available datasets such as RAVDESS and BAUM-1.

However, a key challenge remains: variations in performance, as signified by the evaluation of one dataset when trained on another, which might be due to class imbalances and cultural variances. Future work will explore strategies to enhance generalizability. One promising avenue involves incorporating additional modalities, such as body language analysis or physiological signals, which can be readily converted into text using the proposed rule-based system. This would provide a more comprehensive picture of human emotion, potentially improving recognition accuracy. Supervised learning or reinforcement learning approaches to deal with the generation and optimization of rules could be utilized to simplify the creation of rules and thresholds for each of the possible modalities. Deploying a cloud-based solution and using continuous integration and deployment (CI/CD) pipelines would make it much more scalable and easier to maintain. Additionally, techniques like domain adaptation or data augmentation could be employed to enable the model to navigate diverse data landscapes, ensuring robustness in real-world applications. Furthermore, investigating alternative fusion strategies presents an opportunity for improvement. While the current early-fusion approach yields impressive results, a comparative analysis with late- or intermediate-fusion techniques could reveal potential benefits. Research into the strengths and weaknesses of each approach, coupled with the exploration of novel fusion methods that effectively exploit inter-modal relationships, could lead to significant advancements. Future research will focus on enhancing generalizability, exploring various emotional models, exploring alternative fusion strategies, and potentially incorporating larger and more diverse pre-trained language models (LLMs) for feature extraction. The ultimate goal lies in developing robust, interpretable, and adaptable models capable of deciphering human emotions across diverse contexts. This pursuit holds immense potential for fostering deeper human-machine interaction and a future where technology can not only understand but also respond to human emotions with empathy and understanding.

ACKNOWLEDGMENT

The authors extend their appreciation to King Saud University for funding the publication of this research through the Researchers Supporting Project number (RSPD2024R809), King Saud University, Riyadh, Saudi Arabia.

REFERENCES

[1] M. Sharma, I. Kandasamy, and W. B. Vasantha, "Emotion quantification and classification using the neutrosophic approach to deep learning," *Appl. Soft Comput.*, vol. 148, Nov. 2023, Art. no. 110896, doi: 10.1016/j.asoc.2023.110896.

[2] A. Chowanda, I. A. Iswanto, and E. W. Andangsari, "Exploring deep learning algorithm to model emotions recognition from speech," *Proc. Comput. Sci.*, vol. 216, pp. 706–713, Jan. 2023, doi: 10.1016/j.procs.2022.12.187.

[3] V. Singh and S. Prasad, "Speech emotion recognition system using gender dependent convolution neural network," *Proc. Comput. Sci.*, vol. 218, pp. 2533–2540, Jan. 2023, doi: 10.1016/j.procs.2023.01.227.

[4] A. Chakhtouna, S. Sekkate, and A. Adib, "Unveiling embedded features in Wav2vec2 and HuBERT msodels for speech emotion recognition," *Proc. Comput. Sci.*, vol. 232, pp. 2560–2569, Jan. 2024, doi: 10.1016/j.procs.2024.02.074.

[5] E. Mancini, A. Galassi, F. Ruggeri, and P. Torroni, "Disruptive situation detection on public transport through speech emotion recognition," *Intell. Syst. With Appl.*, vol. 21, Mar. 2024, Art. no. 200305, doi: 10.1016/j.iswa.2023.200305.

[6] Y. Khaireddin and Z. Chen, "Facial emotion recognition: State of the art performance on FER2013," 2021, *arXiv:2105.03588*.

[7] J. H. Cheong, E. Jolly, T. Xie, S. Byrne, M. Kenney, and L. J. Chang, "Py-feat: Python facial expression analysis toolbox," *Affect. Sci.*, vol. 4, no. 4, pp. 781–796, Aug. 2023, doi: 10.1007/s42761-023-00191-4.

[8] C. Gautam and K. R. Seeja, "Facial emotion recognition using handcrafted features and CNN," *Proc. Comput. Sci.*, vol. 218, pp. 1295–1303, Jan. 2023, doi: 10.1016/j.procs.2023.01.108.

[9] U. Bilotti, C. Bisogni, M. De Marsico, and S. Tramonte, "Multimodal emotion recognition via convolutional neural networks: Comparison of different strategies on two multimodal datasets," *Eng. Appl. Artif. Intell.*, vol. 130, Apr. 2024, Art. no. 107708, doi: 10.1016/j.engappai.2023.107708.

[10] R. A. Jaswal and S. Dhingra, "Empirical analysis of multiple modalities for emotion recognition using convolutional neural network," *Meas., Sensors*, vol. 26, Apr. 2023, Art. no. 100716, doi: 10.1016/j.measen.2023.100716.

[11] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human emotion recognition: Review of sensors and methods," *Sensors*, vol. 20, no. 3, p. 592, Jan. 2020, doi: 10.3390/s20030592.

[12] F. Ma, W. Zhang, Y. Li, S.-L. Huang, and L. Zhang, "Learning better representations for audio-visual emotion recognition with common information," *Appl. Sci.*, vol. 10, no. 20, p. 7239, Oct. 2020, doi: 10.3390/app10207239.

[13] T. H. Farook, F. H. Saad, S. Ahmed, and J. Dudley, "Dental loop SnP: Speech and phonetic pattern recognition," *SoftwareX*, vol. 24, Dec. 2023, Art. no. 101604, doi: 10.1016/j.softx.2023.101604.

[14] S. Gupta, P. Kumar, and R. Tekchandani, "An optimized deep convolutional neural network for adaptive learning using feature fusion in multimodal data," *Decis. Anal. J.*, vol. 8, Sep. 2023, Art. no. 100277, doi: 10.1016/j.dajour.2023.100277.

[15] M. Wozniak, M. Sakowicz, K. Ledwosinski, J. Rzepkowski, P. Czapla, and S. Zaporowski, "Bimodal emotion recognition based on vocal and facial features," *Proc. Comput. Sci.*, vol. 225, pp. 2556–2566, Jan. 2023, doi: 10.1016/j.procs.2023.10.247.

[16] J. Yang, X. Dong, and X. Du, "SMFNM: Semi-supervised multimodal fusion network with main-modal for real-time emotion recognition in conversations," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 35, no. 9, Oct. 2023, Art. no. 101791, doi: 10.1016/j.jksuci.2023.101791.

[17] P. Kozlov, A. Akram, and P. Shamoi, "Fuzzy approach for audio-video emotion recognition in computer games for children," *Proc. Comput. Sci.*, vol. 231, pp. 771–778, Jan. 2024, doi: 10.1016/j.procs.2023.12.139.

[18] W. Liu, S. Cao, and S. Zhang, "Multimodal consistency-specificity fusion based on information bottleneck for sentiment analysis," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 36, no. 2, Feb. 2024, Art. no. 101943, doi: 10.1016/j.jksuci.2024.101943.

[19] S. Li and S. Okada, "Interpretable multimodal sentiment analysis based on textual modality descriptions by using large-scale language models," 2023, *arXiv:2305.06162*.

[20] C. Gan, J. Zheng, Q. Zhu, D. K. Jain, and V. Štruc, "A graph neural network with context filtering and feature correction for conversational emotion recognition," *Inf. Sci.*, vol. 658, Feb. 2024, Art. no. 120017, doi: 10.1016/j.ins.2023.120017.

[21] P. Ekman and W. V. Friesen, *Facial Action Coding System*, Environmental Psychology & Nonverbal Behavior, Jan. 1978.

[22] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.

[23] F. Eyben, M. Wöllmer, and B. Schuller, "OpenEAR—Introducing the Munich open-source emotion and affect recognition toolkit," in *Proc. 3rd Int. Conf. Affect. Comput. Intell. Interact. Workshops*, Dec. 2009, pp. 1–6, doi: 10.1109/ACII.2009.5349350.

[24] S. R. Livingstone and F. A. Russo, "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*, vol. 13, no. 5, May 2018, Art. no. e0196391, doi: 10.1371/journal.pone.0196391.

[25] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, "BAUM-1: A spontaneous audio-visual face database of affective and mental states," *IEEE Trans. Affect. Comput.*, vol. 8, no. 3, pp. 300–313, Jul. 2017, doi: 10.1109/TAFFC.2016.2553038.

[26] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The Munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.

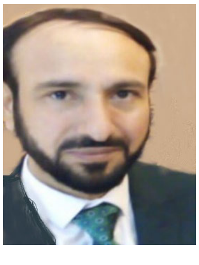
[27] A. Mazarei, R. Sousa, J. Mendes-Moreira, S. Molchanov, and H. M. Ferreira, "Online boxplot derived outlier detection," *Int. J. Data Sci. Anal.*, pp. 1–8, May 2024.

[28] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*, 2020, pp. 38–45.

OMKUMAR CHANDRAUMAKANTHAM received the B.Tech. and M.Tech. degrees (Hons.) in CSE from JNTU-Anantapur, in 2010 and 2013, respectively, and the Ph.D. degree in cyber security from Anna University, Chennai, in 2020. He is currently an Assistant Professor (Sr. G) with the School of Computer Science Engineering, Vellore Institute of Technology–Chennai Campus, has a distinguished academic and professional journey. He embarked on his teaching career at SRM-Easwari Engineering College, Chennai, from 2013 to 2016, where he discovered his passion for teaching and honed his pedagogical skills. His research interests include the IoT and deep learning, leading to the filing of an international patent and the publication of 25 research articles in reputable journals, establishing an H-index value of seven.

N. GOWTHAM was born in Chennai, Tamil Nadu, India, in 2003. He is currently pursuing the B.Tech. degree in computer science with the School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology–Chennai Campus, Chennai. Since 2023, he has been involved in research work in the domain of artificial intelligence. His research interests include AI-based systems and generative AI.

MOHAMMED ZAKARIAH (Member, IEEE) received the B.Sc. degree in computer science and engineering from Visvesvaraya Technological University, India, in 2005, the master's degree in computer engineering from Jawaharlal Nehru Technological University, India, in 2007, and the Ph.D. degree in informatics. As a researcher, he has published more than 45 articles in reputed journals indexed in *ISI Thomson Reuters* in various topics ranging from bioinformatics, image processing, speech processing, and audio forensics in reputed ISI indexed journals, such as *Molecules*, *Sensors*, *Applied Sciences*, *Electronics*, *Multimedia Tools and Applications*, IEEE Access, and *Applied Soft Computing*. He is experienced in machine learning, artificial intelligence, and image and speech Processing. He has worked on five government-funded (KACST) projects and has experience in writing research grant proposals.



ABDULAZIZ ALMAZYAD received the Ph.D. degree in computer engineering from Syracuse University, Syracuse, NY, USA. He is currently a Professor with the College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. His research interests include the Internet of Things, cloud computing, artificial intelligence, mobile and wireless networks, and information security.