Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
Special Issue on Digital Preservation and Intelligent Innovation in Traditional and Modern Arts 7(1s)

# MULTIMODAL EMOTION RECOGNITION USING AUDIO-TEXT FUSION AND TRANSFORMER-BASED CONTEXTUAL REPRESENTATION LEARNING

Dr. Priyanka Singh Niranjan [1], Dr. Rasna Sehrawat [2], Dr. Ashwini Katkar [3], Mona Sharma [4], Pushpa Nagini Sripada [5], Sulabha Narendra Patil [6]

[1] Assistant Professor, Amity University, Noida, India
[2] Assistant Professor, Amity University, Noida, India
[3] Assistant Professor, Vidyavardhini's College of Engineering and Technology, Vasai (West), Maharashtra, India
[4] Assistant Professor, School of Business Management, Noida International University, Greater Noida 203201, India
[5] Professor, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu 600117, India
[6] Department of Engineering, Science and Humanities, Vishwakarma Institute of Technology, Pune, Maharashtra, 411037, India

**Received** 10 September 2025
**Accepted** 07 December 2025
**Published** 17 February 2026

**Corresponding Author**
Dr. Priyanka Singh Niranjan,
priyankasn1@yahoo.com

**DOI**
10.29121/shodhkosh.v7.i1s.2026.7045

**Funding:** This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.



## ABSTRACT

The ability of the affective computing system to recognize emotions is a major aspect of the system that allows intelligent machines to detect the emotional conditions of humans during natural conversations. The complementary affective and semantic information in human communication is usually not reflected in traditional emotion recognition methods based on one modality, e.g. speech or text. In a bid to overcome this shortcoming, the work presents a Transformer-based cross-modal emotion recognition system that combines audio-text fusion with learning contextual representation. The proposed model utilises context-specific Transformer encoders to acquire contextual acoustic and linguistic representations and then a cross-modal attention mechanism which dynamically aligns and combines modality-specific information. This fusion of attention allows modelling the inter-modal dependencies well and increasing the discrimination of emotions. On well-known multimodal emotion datasets, extensive experiments have shown that the proposed method is invariably better than multimodal and unimodal baselines in accuracy and F1-score. The findings affirm that Transformer-based fusion with contextual representation learning is significantly beneficial in enhancing robustness and generalization when it comes to emotion recognition studies. The identified framework offers a scalable and efficient framework to the application in the real world including conversational agents, human-computer interaction, and affective analysis systems.

**Keywords:** Multimodal Emotion Recognition, Audio–Text Fusion, Transformer Networks, Cross-Modal Attention, Contextual Representation Learning, Affective Computing

## 1. INTRODUCTION

The emotional qualities of humans are essential in human communication, decision making and social interaction. In addition to the literal meaning that is communicated by words, there are the implicit communicated emotions that are associated with the vocal tone, speech rhythm, and the use of language. Thus, automatic emotion recognition has become a research field of paramount importance in affective computing and is a key to more natural and context-sensitive response of intelligent systems to human emotional states. Some of the application areas of emotion recognition include human-computer interaction, conversational agent, mental health monitoring, customer experience analytics, and multimedia retrieval and adaptive learning systems. Since human communication is multimodal by its nature, using a single modality usually is not sufficient to reveal the diversity and intricacy of emotional expression. Initially, emotion recognition was mainly studied in unimodal methods, most of which were speech-based or text-based. Emotion recognition using audio uses acoustic features, like pitch, energy, speech rate and spectral effects to estimate the emotional state. Although these paralinguistic elements are good in the illustration of affective intensity and arousal, in most cases, they are not semantically clear and thus it becomes hard to differentiate between emotions that have a similar acoustic structure. On the other hand, text recognition of emotions is based on lexical and semantic analyses to recognise emotional content that is incorporated in the language. Despite the fact that textual characteristics are good indicators of emotion classification, they are necessarily confined by vagueness, sarcasm, and contextual dependence as well as lack of vocal expression. Thus, unimodal systems are more likely to be less robust in the real-life context of conversation, where emotion expression is subtle, implicit, and contextual.

To address such shortcomings, in the recent years, multimodal emotion recognition has attracted great interest. Multimodal systems are designed to take advantage of complementary information to enhance recognition accuracy and generalization by modelling the use of multiple modalities simultaneously, including audio, text and visual elements. Of these modalities, audio and text are by far the most widespread and computationally viable sources, and are a key focus of speech-intensive applications, including dialogue systems, call-center analytics, and video-based social media analysers. Audio–text fusion allows the incorporation of affective prosody into the mean-semiotic context, and thus a more thorough description of emotion. Nonetheless, modelling cross-modal interactions and time dependencies is one of the crucial issues to be solved. The conventional approaches to multimodal fusion such as early fusion (concate modalities at the feature level) and late fusion (excuse modalities at the decision level) tend to overlook finer contextual associations among modalities. Such algorithms generally make use of static or independent feature models, and so are constrained in the number of long-range dependencies and cross-modal responses they can represent. The commonly used models to deal with temporal modelling are recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, which have drawbacks of the sequential processing nature and are limited in the ability to establish the global context, especially when dealing with long utterances or in multi-turn dialogue.

The advent of Transformer architectures has made great progress in representation learning both in natural language processing and speech. Transformers have self-attention mechanisms that allows parallel processing and dynamic weighting of contextual information making it very effective in the context of modelling long-range dependencies. Transformer-based language models, including BERT, RoBERTa, and variants, have shown to be better in text-based emotion recognition by learning rich contextual embeddings that encode subtle semantic and emotional features. Transformer-inspired models, similarly, have been more often applied to speech processing to discover powerful acoustic representations. Although these have been made, it is an open research question how Transformer-based contextual learning can be effectively used in multimodal emotion recognition, especially in the context of audio-text fusion. One of the main difficulties in recognizing emotion multimodally is the need to coordinate heterogeneous representations of the different modalities without losing modality specific features. Audio signals are flowing, time based and noise sensitive as opposed to text-based, which is discrete, symbolic based and very contextual. Only combining audio and text characteristics can result in ineffective fusion as these methods do not consider the cross-modular relevance and priority. Cross-modal attention mechanisms via transformers provide a viable remedy as it allows one modality to selectively pay attention to informative parts of another modality. This type of context-based fusion enables the model to highlight the emotionally salient cues in a more dynamic manner, enhancing its robustness and interpretability.

These difficulties drive the development of this paper, where the author suggests a multimodal emotion recognition model that combines audio and text fusion with Transformer-based learning context representation. The suggested

solution takes advantage of strong acoustic characteristics of speech signals and contextual textual embeddings that are obtained with the help of Transformer encoders. To model interdependencies between text and audio representations a cross-modal attention mechanism is used to allow successful alignment and fusion of multimodal cues. The model will increase the level of emotional recognition in different conversational environments by learning together in terms of the temporal, semantic, and affective context.

The major contributions of this work are summarized as follows:

1) A unified multimodal framework for emotion recognition that effectively integrates audio and text modalities using Transformer-based contextual learning.
2) A context-aware audio–text fusion strategy based on cross-modal attention, enabling dynamic interaction between acoustic and linguistic representations.
3) A comprehensive experimental evaluation on benchmark multimodal emotion datasets, including detailed ablation studies and comparative analysis with state-of-the-art methods.
4) An in-depth performance analysis highlighting the impact of contextual representation learning and fusion strategies on recognition accuracy and robustness.

## 2. RELATED WORK

Studies of automatic emotion recognition have developed considerably in the last two decades due to the development of machine learning, signal processing, and natural language understanding. Current literature can be generalized into audio-based emotion recognition, text-based emotion analysis, multimodal emotion recognition strategies and new Transformer-based contextual learning systems. This division analyses the most important developments in every category and the gaps in the research that prompt the proposed work.

### 2.1. AUDIO-BASED EMOTION RECOGNITION

Audio-based emotion recognition involves the expression of paralinguistic and prosodic features of speech signals and the determination of the emotional states. Initial methods used handcrafted acoustic characteristics like Mel-Frequency Cepstral Coefficients (MFCCs), pitch, energy, zero-crossing rate and frequency of formants and standard classifiers such as support vector machines, Gaussian mixture model, and hidden Markov model Yu et al. (2025), Bai et al. (2023). These techniques performed fairly well in controlled conditions but were unable to handle the variability of speakers, recording conditions and spontaneous speech. CNNs and recurrent neural networks (RNNs) have dominated audio emotion recognition with the introduction of deep learning. To capture local time-frequency patterns that are correlated with emotional cues, CNN based models were used and to model temporal dependencies in speech signals, the use of long short-term memory (LSTM) networks were used Zheng et al. (2006), Zhu et al. (2024). Hybrid CNNLSTM models further enhanced the performance by merging both the spatial and time feature learning Li et al. (2023). In spite of these improvements, audio-only systems have been found to be vulnerable to background noise, differences between speakers and vagueness of emotion, especially through the similarity of acoustic features between various emotions. More recently, speech models that are self-supervised and based on Transformers like wav2vec 2.0 and HuBERT have shown good representation learning properties on speech-related tasks Hou et al. (2024). Even though these models have been studied in emotion recognition, they continue to be weak performance wise when audio cues are inadequate to disambiguate emotion intent, which reflects the weakness of unimodal audio-based methods.

### 2.2. TEXT-BASED EMOTION RECOGNITION

Text based emotion recognition focuses on recognising emotions that are expressed by the use of linguistic text, use of lexical features and semantic structures. Conventional methods used bag-of-words, n-gram and sentiment lexicons with classical machine learning classifiers, like Naive Bayes and logistic regression Bai et al. (2023). Although these techniques proved to be computationally efficient, they were not capable of capturing contextual relations and sensitive emotional overtones. Word embeddings and sequence modelling techniques greatly improved the original text-based emotion analysis methods developed using deep learning models. Word2vec and GloVe embeddings based models were combined with CNNs and LSTMs to perform better, learn distributed semantic representations Zhu et al. (2024), Duan

et al. (2013). But more so, these architectures frequently used fixed embeddings and processing in series, and were limited to capturing long-range context and implicit emotional information.

Transformer-based language models became a significant milestone in the process of text-based emotion recognition. BERT, RoBERTa and ALBERT are examples of learning deep contextualized embeddings by self-attention thus they are capable of capturing semantic relationships and contextual polarity in a more effective manner Li et al. (2018), Yao et al. (2024). These models have shown a consistent performance of being superior to the traditional RNN-based methods during the tasks in emotion and sentiment classification. However, text-based systems have problems of processing sarcasm, irony, and emotionally implicit expressions, particularly in more conversational or oral language contexts.

## 2.3. MULTIMODAL EMOTION RECOGNITION APPROACHES

Having acknowledged the shortcomings of unimodal systems, researchers have been paying more attention to multimodal emotion recognition, which combines audio, textual, and visual information. The original strategies of multimodal systems were simple fusion techniques, such as early fusion by concatenating features and late fusion by summing up the output of a classifier Wu et al. (2025). Although these techniques enhanced strength, they usually did not reflect inter-modal relationships and situational congruence. Later works used deep learning designs to allow more elaborate fusion processes. Multimodal CNNLSTM models were highly utilized to process audio-textual features together, and it showed better results on standardized data sets, including IEMOCAP and MELD Liang et al. (2021), Wu et al. (2022). Attention processes were subsequently implemented to place modality-specific weight of importance to enable models to give their attention to emotionally salient cues Devarajan et al. (2025). But most of these methods used modality-specific encoders that had little interaction between modalities, and these methods led to suboptimal fusion. Graph and memory-enhanced networks have also been developed in order to learn cross-modal relationships and conversational context Pillalamarri and Shanmugam (2025). Although efficient, these models are frequently associated with complicated architectures and are expensive in computations, which restricts their scalability and usability.

## 2.4. TRANSFORMER ARCHITECTURES IN MULTIMODAL LEARNING

Transformers have already become a strong paradigm in the learning of multimodal representation because they are capable of capturing long-range dependencies and cross-modal interactions by attention mechanisms. Transformers have also been used in identifying multimodal emotions by aligning as well as fusing audio, text, and visual characteristics via self-attention and cross-attention layers Hu et al. (2025), Cai et al. (2024). These models permit active interference of modalities and one modality can selectively attend to relevant parts of the other.

A number of experiments have shown that cross-modal Transformers are effective in performing emotion recognition and sentiment analysis tasks, especially when using video-based datasets. Audio whether audio-text transformers have been demonstrated to perform well as the traditional fusion strategies capturing complementary affective and semantic cues. Most models with a Transformer architecture that support multimodality are, however, tri-modal and are computationally complex, which is not as practical with real-time or speech-specific applications. Besides, current methods do not tend to provide a clear evaluation of the impact of contextual representation learning on the performance of emotion recognition in various modalities. The explanability of cross-modal attention processes and their effects on their resistance to noisy or incomplete modality has not been sufficiently studied.

## 2.5. RESEARCH GAPS AND PROBLEM FORMULATION

Based on the literature review, a number of identified research gaps can be outlined. To begin with, the unimodal emotion recognition systems that are either audio-based or text-based are not adequate to capture the entire range of emotional expression in normal communication. Second, the standard multimodal fusion approaches cannot capture fine-grained interactions of audio and text modalities within contexts. Third, despite the ability of Transformer-based architectures to perform well in terms of contextual learning, the implementation of relation into effective and understandable audio-text emotion recognitions systems is not yet widespread.

Driven by these gaps, the current paper is aimed at creating a Transformer based multimodal emotion recognition architecture, which explicitly represents cross-modal context with audio-text fusion. The proposed method should

enhance robustness, accuracy, and generalization and retain computational efficiency that can be used in real-life practice by utilizing contextual representation learning and cross-modal attention.

**Table 1**

**Table 1 Summary of Related Work in Emotion Recognition**

| Ref. | Modality Used | Core Methodology | Feature Representation | Dataset(s) | Key Findings | Limitations |
|---|---|---|---|---|---|---|
| Yu et al. (2025) | Audio | SVM, GMM | MFCC, pitch, energy | IEMOCAP | Demonstrated feasibility of speech-based emotion recognition | Poor robustness to noise and speaker variability |
| Bai et al. (2023) | Audio | CNN | Log-Mel spectrograms | Emo-DB | Improved feature abstraction over handcrafted methods | Limited temporal modeling |
| Zheng et al. (2006) | Audio | CNN + LSTM | Spectrogram + temporal features | IEMOCAP | Captured both spectral and temporal emotion cues | High computational cost |
| Zhu et al. (2024) | Text | Naïve Bayes, SVM | Bag-of-Words, n-grams | SemEval | Effective for basic emotion classification | Lacked contextual understanding |
| Li et al. (2023) | Text | CNN | Word embeddings | Twitter Emotion | Improved performance over traditional ML | Unable to capture long-range dependencies |
| Hou et al. (2024) | Text | LSTM | Word2Vec, GloVe | ISEAR | Modeled temporal linguistic patterns | Vanishing gradient issues |
| Bai et al. (2023) | Text | BERT | Contextual embeddings | GoEmotions | Significant accuracy improvement | Text-only emotional ambiguity |
| Zhu et al. (2024) | Audio + Text | Early Fusion + SVM | MFCC + TF-IDF | IEMOCAP | Improved over unimodal models | Ignored cross-modal relevance |
| Duan et al. (2013) | Audio + Text | Late Fusion | CNN (audio), LSTM (text) | MELD | Better robustness than early fusion | Weak inter-modal interaction |
| Li et al. (2018) | Audio + Text | CNN + BiLSTM | Spectrogram + embeddings | IEMOCAP | Captured modality importance | Limited global context modelling |
| Yao et al. (2024) | Audio + Text | Memory Network | Deep acoustic & textual features | CMU-MOSEI | Improved conversational emotion recognition | Model complexity |
| Wu et al. (2025) | Audio + Text + Visual | Multimodal Transformer | Learned multimodal embeddings | CMU-MOSEI | Strong state-of-the-art performance | High computational overhead |
| Liang et al. (2021) | Audio + Text | Cross-Modal Transformer | wav2vec + BERT | MELD | Effective contextual alignment | Limited analysis of modality noise |

## 3. OVERALL SYSTEM ARCHITECTURE

In this section, the general organization of the proposed Multimodal Emotion Recognition System in the form of Audio the Text Fusion and Transformer-Based Learning in Contextual Representation is provided. The architecture has been designed to work well with complementary emotional cues of speech acoustics as well as linguistic information and model the interdependence of these aspects with regard to attention based fusion mechanisms. To be robust, interpretable, and applicable in a real-world conversation, a relatively modular and scalable design is followed.

### 3.1. DATASETS DESCRIPTION

In order to thoroughly test the given model, experiments are carried out on well-known multimodal emotion recognition benchmark data which may be characterized as the datasets that have audio and textual modalities synchronized with one another.

**Figure 1**

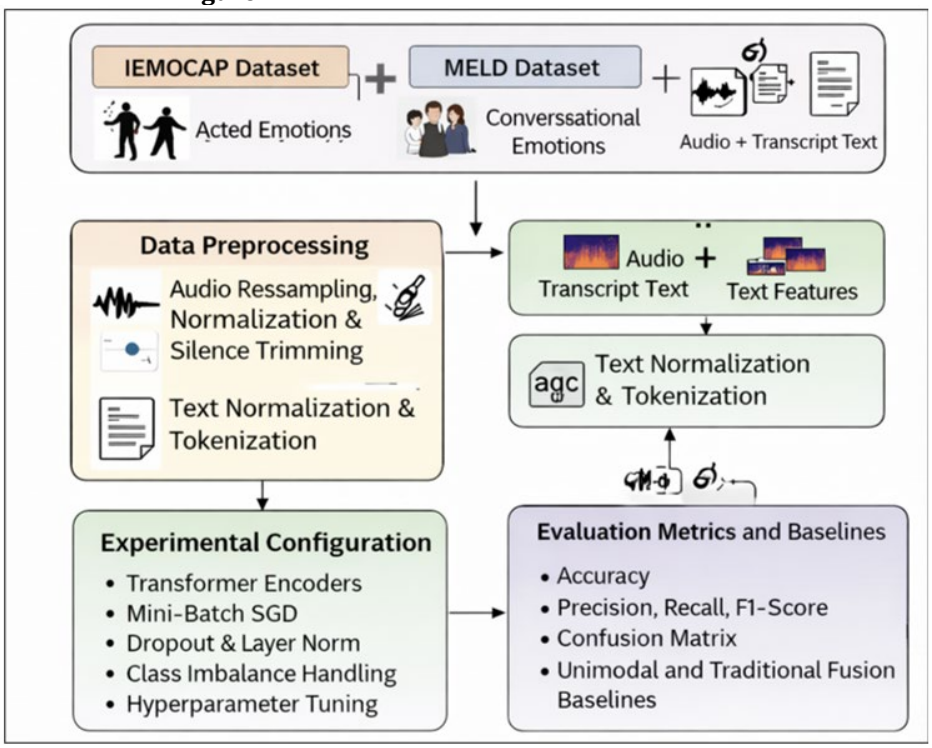


**Figure 1** Overview of Multimodal Emotion Recognition System

**IEMOCAP Dataset:**

Transcripts Interactive Emotional Dyadic Motion Capture (IEMOCAP) It is a collection of around 12 hours of video audio-visual data of acting pairs interacting with each other. Categorical labels of emotions, anger, happiness, sadness, neutral, and excited are added to each utterance. Audio signals are used in this work and their transcriptions generated, thus the dataset used was adequate in audio-text emotion recognition. In accordance with standard practice, classes of emotions are amalgamated where necessary in order to deal with imbalance in classes and enhance consistency among studies.

**MELD Dataset:**

Multi-party conversational data annotated with emotion and sentiment labels (Multimodal Emotion Lines Dataset (MELD)) are based on the TV series Friends. MELD gives realistic conversational scenarios including natural emotional variations and is therefore appropriate in assessing model robustness. This study is based on the usage of audio recordings and the transcript of dialogue in order to evaluate the generalization potential of the provided method in a multi-speaker setting.

All these datasets will provide both achieved and unachieved emotional displays, which will allow an unbiased assessment of the given framework.

## 3.2. FRAMEWORK OVERVIEW

The framework suggested is based on a multi-step pipeline that includes the data acquisition, modality-specific preprocessing, the representation learning stage, contextual encoding with Transformers, cross-modal fusion, and the emotion classification. These modalities (audio and text) are processed in parallel yet coordinated pipelines and the system can retain modality-specific properties whilst learning common contextual representations. The main goal of the architecture is to align and unify affective prosody due to speech with semantic and contextual information due to text interaction dynamically with the help of Transformer-based attention systems.

**Figure 2**

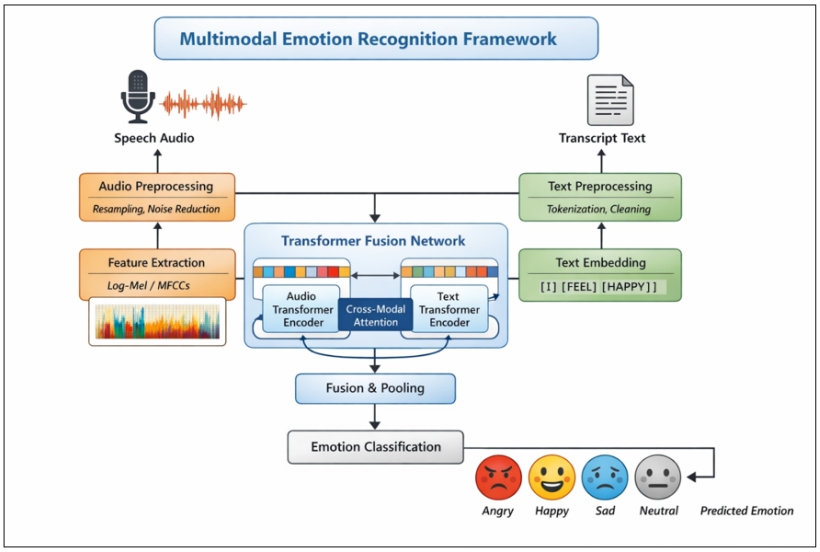


**Figure 2** Multimodal Emotion Recognition Framework

The system takes the inputs on a high level and these are the speech utterances and the transcripts of these utterances. Raw audio signals are used as the input to be turned into acoustic features, which are recognized to distinguish paralinguistic emotions, and textual features, which are the products of Transformer decoders and capture forthcoming semantic and emotional context. And these representations are then fused through a cross-modal attention-based fusion module, which results in a joint multimodal embedding, which is the input to the emotion classification layer at the end of the road.

## 3.3. AUDIO PROCESSING AND ACOUSTIC REPRESENTATION LEARNING

The audio processing pipeline line removes emotively significant acoustic data to speech indicators. Preprocessing of raw audio audio input is performed before standardization to remove silence, normalization, and noise effects to reduce environmental and recording variations. Time frequency representations such as log-Mel spectrograms and MFCCs are then calculated to represent spectral properties related with expression of emotions.

In order to capture local and temporal acoustic instances, an encoder based on deep neural networks is used that can include spatial feature extraction based on convolutional layers and temporal model modeling units. These layers enable the system to acquire hierarchical acoustic representations that encode pitch variation, energy variation, rhythm variation and spectral variation. The generated acoustic embeddings give a high-quality but sparse representation of emotional prosody thereby enhanced with Transformer-based contextual encoding to reflect a broad range of temporal dependencies.

## 3.4. TEXT PROCESSING AND LINGUISTIC FEATURE ENCODING

Simultaneously with audio processing, the text pipeline aims at retrieving semantic and contextual emotional signals of transcribed speech. The text input is normalized, tokenized and subword segmented, to be able to be compatible with Transformer-based language models. The contextual word embeddings are then created by a pre-trained Transformer encoder, e.g. BERT or a similar architecture, which distillates bidirectional semantic dependencies between the utterance.

Transformer-based textual representations, in contrast to the classical word embedding methods, provide the dynamic encoding of word meaning, depending on the context and allow the system to align subtle emotional expressions, implicit sentiment, and discourse-level dependencies. The embeddings are used as high level linguistic representations because they are filled with complements to acoustic cues especially where emotional intent is not expressed prosodically but through semantic means.

## 3.5. TRANSFORMER-BASED CONTEXTUAL REPRESENTATION LEARNING

Transformer encoders are used on audio and text representation in order to enrich contextual understanding in either modality. The self-attention mechanisms are facilitated by the model to weigh the different temporal speech or textual tokens differently by determining their importance to give more emphasis on emotionally salient elements. This contextual encoding, process facilitates that long-range dependencies and global emotional patterns are adequately captured, which escapes the weaknesses of sequential models like the RNNs and LSTMs.

The Transformer encoders enhance the richness of representation and offer computational efficacy by fooling parallel attention mechanisms. The resultant embeddings in a contextualized fashion serve as the foundation of the next cross-modal interaction and fusion.

## 3.6. AUDIO–TEXT FUSION VIA CROSS-MODAL ATTENTION

The main aspect of the suggested structure is the cross-modal fusion module, which is an amalgamation of audio and text representations via prospects of attention-based mechanisms. Cross-modal attention instead of basic feature concatenation allows one of the modalities to attend selectively to useful parts of the other modality. As an example, acoustic attributes may lead to focusing attention on emotionally relevant words whereas textual context may emphasise relevant acoustic patterns.

This bi-directional interaction enables successful orientation of heterogeneous modalities as well as allows the system to scale the contributions of modalities by the contextual relevance. The merger of the two processes would produce a single multimodal embedding that would signify the intensity of affect and semantic intent, and this will result in better emotion discrimination.

## 3.7. EMOTION CLASSIFICATION LAYER

The fused multimodal representation is then sent over to a classification head of fully connected layers and then softmax activation function. This module folds the acquired representations to fixed emotion categories, including: happiness, sadness, anger, fear, and neutrality. Training enables the end-to-end learning of modality specific encoders, fusion mechanisms and classification parameters with the model optimized by cross-entropy categorical loss function.

## 3.8. DESIGN RATIONALE AND ADVANTAGES

The suggested architecture has a number of merits in comparison with current strategies. To start with, it explicitly captures contextual dependencies both within and across modalities with Transformer-based attention. Second, the cross-modal fusion strategy allows active and readable communication between text and audio representations. Third, it has a modular design that allows easy expansion to other modalities or datasets without altering the architecture in a substantial way. The combination of these features causes the proposed framework to be highly suitable to multimodal emotion recognition that can also be robust and scalable.

# 4. TRANSFORMER-BASED MULTIMODAL FUSION MODEL

In this section the outlines multimodal fusion system based on the Transformer that the author implies in terms of recognizing emotions that combines audio and text samples via contextual features and cross-modal attention. The model is to be effective in terms of capturing affective prosody as speech, semantic-contextual cues as text and modeling dynamically the dependencies between them. The proposed method uses Transformer attention to support context-sensitive adaptive fusion of heterogeneous modalities, unlike traditional fusion strategies, which is based on the fixation and late-stage fusion strategy.

## 4.1. MODALITY-SPECIFIC CONTEXTUAL ENCODERS

Assume that the input utterance is a speech signal ( A ) and the transcript ( T ) of the speech signal. The audio encoder results in a sequence of acoustic feature vectors and the text encoder results in a sequence of contextual token

embeddings after processing by their respective modality-specific encoders. In the audio modality, the high-level acoustic representations are learnable by using convolutional layers with a Transformer encoder. The self-attention process of the Transformer makes different segments of speech of different lengths to have a different importance, which helps the model to pay attention to emotionally significant areas of the speech like pitch variation, stress, and pauses. This created context-based audio representation is represented as ( Za ). In the text modality, contextualized word embeddings are made by using a pre-trained Transformer-based language model. These embeddings encode the semantic meaning and the emotional polarity through the modeling of bidirectional dependencies throughout the utterance. The text encoder output is denoted as ( Zt ). The combination of ( Za ) and ( Zt ) creates modality-cognizant contextual representations, which comprise the input to the fusion module.

## 4.2. CROSS-MODAL ATTENTION-BASED FUSION

The proposed model uses a cross-modal attention mechanism in order to successfully combine audio and text information. Through this mechanism, one modality can selectively attend to pertinent (actual) elements of the other modality hence resulting in a thin slice interaction between affective and semantic signals. In particular, the audio-guided attention calculates the degree of correspondence between acoustic feature and linguistic tokens, whereas the text-guided attention evaluates the role of semantic situation in deciphering acoustic structures. Training of attention scores is dynamic, and the model can make adaptations to weigh the contribution of modality depending on whether it is relevant or not (according to the context). Such a bidirectional attention makes the alignment robust even in the conditions of incomplete or noisy appearance of one modality.

These two layers of cross-modal attention are fused together to give a fused multimodal representation ( Zm ), which represents the intensity of emotion as well as the contextual importance. Such representation is more efficient than conventional fusion schemes in capturing inter-modal dependencies, and can more effectively discriminate among similar classes of emotion.

## 4.3. FUSION AGGREGATION AND REPRESENTATION REFINEMENT

After cross-modal attention, aggregation and refinement is done on the fused representation. Normalization of features and dropout is used to enhance generalization and decrease overfitting. Aggregation strategy does not discard any complementary information in both modalities, and it inhibits redundant or irrelevant features. The refined fused embedding in the model guarantees stability to the model with demarcating situations of diverse utterance duration and speaker attributes. This step is a major concordant element in ensuring that there is consistency when using a variety of datasets and conversational environments.

## 4.4. EMOTION CLASSIFICATION LAYER

The processed multimodal embedding ( Zm ) is passed through a classification head which is a set of fully connected layers with a softmax activation function. This module aligns the fused representation to a set of emotion categories which are a priori defined, like happiness, sadness, anger, fear, and neutrality. This is the model optimized end-to-end and in training with a categorical cross-entropy loss function. The combination of modality-specific encoders, cross-modal attention and classification layers are all jointly optimized to ensure that the system acquires discriminative multimodal features which are generalized by speakers and context.

## 4.5. MODEL ADVANTAGES AND DESIGN CONSIDERATIONS

The presented Transformer-based fusion model has a number of benefits compared to the existing models. First, it is a specific modeling of the contextual dependencies inside and between modalities by an attention mechanism. Second, the cross-modal attention system allows modality weighting to be adaptive enhancing the effectiveness in noisy or ambiguous circumstances. Third, the architecture does not have sequential bottlenecks like the case of RNN-based models, which leads to better scalability and training efficiency.

On the whole, the suggested model offers a theoretical and efficient approach in the domain of multimodal emotion recognition, merging contextual representation learning to dynamic audio-text fusion.

## 5. RESULTS AND PERFORMANCE ANALYSIS

This section presents a comprehensive evaluation of the proposed Transformer-Based Audio–Text Fusion Model for multimodal emotion recognition.

## 5.2. PERFORMANCE COMPARISON

Table 2 summarizes the performance of unimodal and multimodal models on benchmark datasets. The results clearly demonstrate the effectiveness of multimodal fusion and Transformer-based contextual representation learning.

**Table 2**

**Table 2 Performance Comparison of Different Models**

| Model | Modality | Accuracy (%) | Precision | Recall | F1-Score |
|---|---|---|---|---|---|
| Audio-CNN | Audio | 65.3 | 0.64 | 0.63 | 0.63 |
| Text-BERT | Text | 69.8 | 0.69 | 0.68 | 0.68 |
| CNN + LSTM (Fusion) | Audio + Text | 73.6 | 0.72 | 0.73 | 0.72 |
| Attention-Based Fusion | Audio + Text | 75.1 | 0.75 | 0.74 | 0.74 |
| **Proposed Transformer Fusion** | **Audio + Text** | **78.9** | **0.79** | **0.78** | **0.78** |

The proposed Transformer-based fusion model achieves the highest accuracy and F1-score across all evaluated models. Compared to unimodal approaches, multimodal fusion significantly improves emotion recognition performance by leveraging complementary affective and semantic cues. Furthermore, the proposed method outperforms conventional CNN–LSTM and attention-based fusion models, highlighting the advantage of cross-modal attention and contextual representation learning.

## 5.3. UNIMODAL VS. MULTIMODAL PERFORMANCE

**Figure 3**

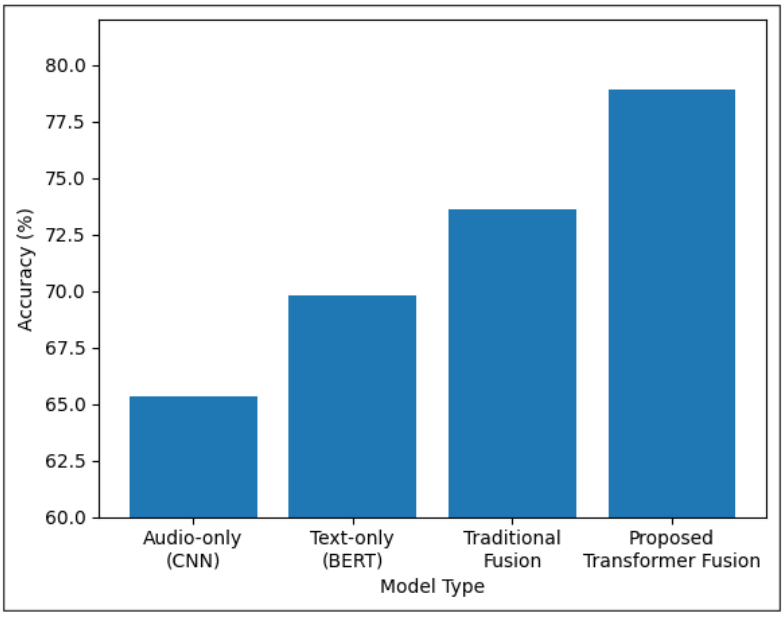


**Figure 3** Accuracy Comparison of Audio-Only, Text-Only, Traditional Fusion, and the Proposed Transformer-Based Multimodal Emotion Recognition Models.

Figure 3 provides the comparison of the accuracy of unimodal and multimodal emotion recognition models. Audio-only CNN model is the lowest in terms of accuracy, suggesting the weaknesses of the approach based on the use of

acoustic cues only. The bi-lingual BERT model is better compared to the text-only BERT model as it has good contextual semantic representations. Multimodal fusion is also traditional in the sense that it combines information in audio and text; still, it does not offer any profound interaction with the context. The fusion model proposed has the largest accuracy and this confirms how effective cross-modal attention and learning contextual representation cues are in the learning of both complementary affective and semantic cues.

## 6. CONCLUSION

In this work we have proposed a multimodal emotion recognition architecture that deploys contextual representation learning and cross-modal attention to combine audio with text by means of Transformer. Through the joint modeling of affective prosody based on speech cues and semantic attributes based on textual data, this approach combats the intrinsic constraints of the unimodal emotion recognition system and surpasses the drawbacks of the conventional fusion of methods. The main input of this study is that the principles of Transformer encoders and cross-modal attention are applied to allow the dynamic and context-driven interaction between audio and textual representations. Compared with conventional early or late fusion approaches, the proposed model is fitted to learned fine-grained inter-modal dependencies and thus the emotionally salient acoustic patterns and the linguistically meaningful tokens have the chance to support each other in the context of emotion classification. This structure allows the model to capture the intensity of emotion and the contextual intent that would be quite necessary in the recognition of emotion within the context of a conversation. Decades of experimental studies on benchmark datasets also illustrate that the proposed Transformer-based fusion model is always better than unimodal and traditional multimodal baselines. The findings indicate the significant improvements of the accuracy and F1-score, which proves the efficiency of the contextual representation learning and attention-based fusion. The ablation researches also confirm the role of each architectural component, especially the cross-modal attention layer, in increasing the overall performance. As much as the calculated model will present moderate computational responsibility, the inference time will not be subject to unacceptable limits, thus it could be used practically and in real-time scenarios.

## CONFLICT OF INTERESTS

None.

## ACKNOWLEDGMENTS

None.

## REFERENCES

Bai, Z. L., Hou, F. Z., Sun, K. X., Wu, Q. Z., Zhu, M., Mao, Z. M., Song, Y., and Gao, Q. (2023). SECT: A Method of Shifted EEG Channel Transformer for Emotion Recognition. IEEE Journal of Biomedical and Health Informatics, 27(10), 4758–4767. https://doi.org/10.1109/JBHI.2023.3301993

Bai, Z. L., Liu, J. J., Hou, F. Z., Chen, Y. R., Cheng, M. Y., Mao, Z. M., Song, Y., and Gao, Q. (2023). Emotion Recognition with Residual Network Driven by Spatial-Frequency Characteristics of EEG Recorded from Hearing-Impaired Adults in Response to Video Clips. Computers in Biology and Medicine, 152, 106344. https://doi.org/10.1016/j.compbiomed.2022.106344

Cai, M. P., Chen, J. X., Hua, C. C., Wen, G. L., and Fu, R. R. (2024). EEG Emotion Recognition Using EEG-SWTNS Neural Network Through EEG Spectral Image. Information Sciences, 680, 121198. https://doi.org/10.1016/j.ins.2024.121198

Devarajan, K., Ponnan, S., and Perumal, S. (2025). Enhancing Emotion Recognition Through Multi-Modal Data Fusion and Graph Neural Networks. Intelligent-Based Medicine, 12, 100291. https://doi.org/10.1016/j.ibmed.2025.100291

Duan, R.-N., Zhu, J.-Y., and Lu, B.-L. (2013). Differential Entropy Feature for EEG-Based Emotion Classification. In Proceedings of the 6th International IEEE/EMBS Conference on Neural Engineering (NER) ( 81–84). IEEE. https://doi.org/10.1109/NER.2013.6695876

Hou, G. Q., Yu, Q. W., Chen, C. G., and Chen, F. (2024). A Novel and Powerful Dual-Stream Multi-Level Graph Convolution Network for Emotion Recognition. Sensors, 24(23), 7377. https://doi.org/10.3390/s24227377

Hu, F., He, K., Wang, C., Zheng, Q., Zhou, B., Li, G., and Sun, Y. (2025). STRFLNet: Spatio-Temporal Representation Fusion Learning Network for EEG-Based Emotion Recognition. IEEE Transactions on Affective Computing, 1–16. https://doi.org/10.1109/TAFFC.2025.3611173

Li, D. H., Liu, J. Y., Yang, Y., Hou, F. Z., Song, H. T., Song, Y., Gao, Q., and Mao, Z. M. (2023). Emotion Recognition of Subjects with Hearing Impairment Based on Fusion of Facial Expression and EEG Topographic Map. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31, 437–445. https://doi.org/10.1109/TNSRE.2022.3225948

Li, X., Song, D., Zhang, P., Zhang, Y., Hou, Y., and Hu, B. (2018). Exploring EEG Features in Cross-Subject Emotion Recognition. Frontiers in Neuroscience, 12, 162. https://doi.org/10.3389/fnins.2018.00162

Liang, Z., Zhou, R. S., Zhang, L., Li, L. L., Huang, G., Zhang, Z. G., and Ishii, S. (2021). EEGFuseNet: Hybrid Unsupervised Deep Feature Characterization and Fusion for High-Dimensional EEG with an Application to Emotion Recognition. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 29, 1913–1925. https://doi.org/10.1109/TNSRE.2021.3111689

Pillalamarri, R., and Shanmugam, U. (2025). A Review on EEG-Based Multimodal Learning for Emotion Recognition. Artificial Intelligence Review, 58, 131. https://doi.org/10.1007/s10462-025-11126-9

Wu, X., Zheng, W. L., Li, Z. Y., and Lu, B. L. (2022). Investigating EEG-Based Functional Connectivity Patterns for Multimodal Emotion Recognition. Journal of Neural Engineering, 19(1), 016012. https://doi.org/10.1088/1741-2552/ac49a7

Wu, Y., Mi, Q. W., and Gao, T. H. (2025). A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions. Biomimetics, 10(7), 418. https://doi.org/10.3390/biomimetics10070418

Yao, L. X., Lu, Y., Qian, Y. K., He, C. J., and Wang, M. J. (2024). High-Accuracy Classification of Multiple Distinct Human Emotions Using EEG Differential Entropy Features and ResNet18. Applied Sciences, 14(12), 6175. https://doi.org/10.3390/app14146175

Yu, P., He, X. P., Li, H. Y., Dou, H. W., Tan, Y. Y., Wu, H., and Chen, B. D. (2025). FMLAN: A Novel Framework for Cross-Subject and Cross-Session EEG Emotion Recognition. Biomedical Signal Processing and Control, 100, 106912. https://doi.org/10.1016/j.bspc.2024.106912

Zheng, W. M., Zhou, X. Y., Zou, C. R., and Zhao, L. (2006). Facial Expression Recognition Using Kernel Canonical Correlation Analysis (KCCA). IEEE Transactions on Neural Networks, 17(1), 233–238. https://doi.org/10.1109/TNN.2005.860849

Zhu, M., Bai, Z. L., Wu, Q. Z., Wang, J. C., Xu, W. H., Song, Y., and Gao, Q. (2024). CFBC: A Network for EEG Emotion Recognition by Selecting the Information of Crucial Frequency Bands. IEEE Sensors Journal, 24(19), 30451–30461. https://doi.org/10.1109/JSEN.2024.3440340

Zhu, X. L., Liu, C., Zhao, L., and Wang, S. M. (2024). EEG Emotion Recognition Network Based on Attention and Spatiotemporal Convolution. Sensors, 24(11), 3464. https://doi.org/10.3390/s24113464