Advanced Emotion Recognition Using LSTM, RNN, and Transformer Models for Comprehensive Sentiment Analysis

R. Asha, Lipika C, Danush Priyan*

Department of Computer Science and Engineering,
Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu, India

*ashasaker02@gmail.com,*lipikalpana11@gmail.com, madanush999@gmail.com*

Abstract - The emotion recognition is a significant role in human-computer interaction, in the field of affective computing and intelligent decision-support systems. Most existing methods of sentiment analysis presuppose one channel as a text or speech one, which cannot be applicable in any real-world context when the emotional cues are manifested in numerous channels. The present paper has suggested a multimodal emotion recognition system integrating audio, video and text modalities by using a hybrid of a LSTM/RNN- based deep learning framework alongside Transformer framework to conduct a general sentiment analysis. The audio emotions are extracted by MFCC-based features representation and processed using the temporal LSTM network to obtain speech dynamics. To acquire spatiotemporal emotion patterns, analyses of facial expressions are performed based on video frame sequences using a deep visual model. At the same time, the textual emotional inferences are made with the assistance of speech transcripts and a Transformer- based language model that is fine-tuned. Three modalities are predicted and then the weighted late-fusion approach combines them to come up with one strong emotion label. The presented system is utilized as a real- time web app and it can be seen to be useful in practice-based affect-aware applications, such as mental health analysis, intelligent tutoring and customer behavior monitoring.

Keywords: Multimodal Emotion Recognition, long short-term memory (LSTM), Recurrent neural networks (RNN), Transformer models, Sentiment Analysis.

I. INTRODUCTION

The Emotion recognition has emerged as an essential part of human-computer interaction in recent times and can be employed to describe and analyze affective displays in intelligent tutoring systems and mental health assessment, customer behavior analysis and adaptive user interfaces. The first form of sentiment analysis was the classical algorithms of natural language processing and machine learning that can process text-only data. These sorts of unimodal systems can perform quite well when it is applied to few situations, but fails to react to high and complementary information of a human speech dynamics, and facial expression. Human emotion is multimodal as well it is simultaneously exhibited through use of tone of voice, facial muscle movements, and linguistic messages. It implies that the single-modality systems may be described by ambiguity, low noisiness behavior resilience, and inability to generalize to the actual world.

The state of art of deep-learning has tremendously improved the outcomes of emotion recognition when conducted in single modalities. The use of recurrent neural networks (RNNs) and long short-term memory (LSTM) networks has been demonstrated to be very successful in acquiring the temporal correlation of speech cues when these networks are trained with perceptually important signals in the form of Mel-Frequency Cepstral Coefficients (MFCCs). At the same time,

R. Vasanth Kumar Mehta et al. (eds.), *Proceedings of the International Conference on Intelligent Systems for a Sustainable Future (ISSF 2026)*, Atlantis Highlights in Intelligent Systems 16,
https://doi.org/10.2991/978-94-6239-693-7_95

deep convolutional and recurrent architecture has made it possible to analyse facial expressions, according to video exchange sequences, effective in case one knows their spatiotemporal dynamics of emotions. At the same time, Transformer-based language models have changed the text-based sentiment analysis as they can fix the long-range dependency of the semantics through the means of self-attention. Some of them are lightweight architectures such as Distil BERT, though highly classified, yet with reduced computational complexity, thus can be deployed in real-time.

Despite the above improvements, there is indeed an extremely critical gap in the area of effective integration of the multimodal emotional indicators within a single and deployable sentiment analysis framework. The bulk of the literature which exists deals either with a single modality, or is an extremist representation of the modalities in parallel through early fusion facilitation mechanisms requiring strict time synchronization and high processing cost. Besides, a significant number of the systems that have been referenced can be experimented off-line only and are not implemented as end-to-end systems that are able to process real user generated content such as uploaded videos. Other preceding methods are also weakly validated pipeline based where textual, visual and audio predictions may not always be motivated by strong deep learning backbones which have well-known training behavior.

This work has addressed the limitations by providing a multimodal emotion recognition model in a comprehensive format of LSTM/RNN and Transformer models that integrate audio, video, and text stream into one real-time inference model. The proposed architecture is inspired by a late-fusion system, which allows all modalities to be trained independently, by applying the most suitable deep learning model to the latter, and, therefore, not has to adhere to strict cross-modal synchronization. Emotion recognition is first performed by Audio speech recognition through MFCC feature extraction and dynamic system of Temporal LSTM in learning the speech dynamics. Visual emotion name recognition carries out profound frame based temporal modelling of face expressions. Textual emotion analysis the textual emotion analysis is trained using a fine-tuned Distil BERT Transformer model, having six attention layers, twelve attention heads, and an attentional dimension of 768 attention dimensions, and, using a maximum sequence length of 512 tokens, is trained on single-label emotion classification.

The transformer tokenizer used is called on the traditional Word Piece tokenizing method and has special tokens [CLS], [SEP], [PAD], [MASK] and [UNK], and lowercasing and normalization is enabled such that all user inputs undergo homogenous preprocessing. In the text classification model, the model was trained on 10 epochs with a decaying learning rate and periodic evaluation and therefore the optimal evaluated model of the model reached the sixth epoch with the Evaluation accuracy of 0.944 and F1-score of 0.944, signifying high generalization of the emotion classes.This corroborated performance is a boon to Transformer as it makes the module a serious force in the multimodal fusion pipeline.

Weighted late-fusion strategy is a combination process of all three modalities, it is a synthesis of the individual predictions of the three modalities into a single strong emotional term. Here one modality leads to a single factor of classification of the emotions, and a fixed set of weights are added to pool the final decision. This architecture is more stable to be deployed to the actual world where any single modality (e.g. facial video in low-light or speech noises) is not very

reliable. Unlike the previously mentioned fusion methods of merging uncooked attributes across modalities, the proposed decision-level fusion can preserve the integrity and expressibility of every profound model at a reasonable level of computational efficiency.

The latter limitation is the second limitation of most of the offered multimodal affect recognition systems, in that they cannot be applied in real-time, or they cannot be deployed in a repeatable way. To remove this the proposed system is employed as a form of end-to-end web-based application which processes a video uploaded by a user and automatically processes them in terms of audio extraction, speech transcription, MFCC analysis, facial frame detection, Transformer-based text classification, and weighted fusion in order to generate a unified emotional response. Deep models are also lightweight like DistilBERT and LSTM based models which ensures low-latency inference, which is also applicable in the real world.

The research in total contributes one entirely validated, multimodal, real-time emotion study framework using validated deep learning designs and predictable training loggings. The vulnerability, deployment, and interpretability weaknesses of the previously disunimodal and loosely coupled multimodal sentiment analysis systems is directly overcome by the proposed solution that unites the temporal speech modelling, visual emotion detection, and Transformer-based linguistic understanding into a single coherent system.

Contributions:
The prominent successes of this work are as follows:

- ➢ A complete multimodal emotion recognition system, relying on audio LSTM-based analysis, deep visual emotion modeling, and a fine-tuned DistilBERT Transformer to analyze text emotion.
- ➢ An ethical high performance Transformer text classifier that was trained through 10 epochs and proved accuracy of evaluation and F1-score of 0.944 assigns to multimodal fusion with sufficient certainty.
- ➢ A rough fusion technique that enhances the stability of prediction in scenarios where there is true noise and some degradation of modality and has low inference time.

II. RELATED WORK

Since recent years, multimodal emotion recognitions have shifted to cross-modal Transformer fusion in order to take advantage of complementing cues across audio, visual, textual and physiological streams. According to Khan et al., one of the models that can be proposed is MemoCMT, a cross-modal Transformer-based feature fusion model that learns rich modal interactions at its own levels that offers a strong representation power of complex emotional interactions [1]. Its multifaceted nature and the richness of fusion between cross-modal attention make MemoCMT more difficult and intricate to implement in practice, and its integration with a web service lightweight architecture. Similarly, Liu et al. introduce TACFN, Transformer based adaptive cross-modal fusion network that can dynamically attend to and concentrate on the most informative modality using attention mechanisms [3]. TACFN, as adaptive weighting and fine-grained fusion, which

is strong, is more of a fusion design than simple, interpretable late-fusion which can be incorporated with our end-to-end, resource- efficient web application. Many of the works are speculations on additional Transformer designs and fuses of fusion strategies. Yi et al. come up with a multimodal Transformer with two cross-modal attention HyFusER that can depict a complicated intermodal interaction [6]. According to Ali and Hughes, a biosensor- vision multimodal Transformer network is proposed where there is a close relationship between physiological and visual attributes to recognize emotions [7]. Chen et al. can use Mamba and Liquid Neural Networks to combine with cross-modal alignment and enhance the temporal modeling capabilities and flexibility of multimodal fusion [12], and Filali et al. can use a capsule graph Transformer architecture to learn hierarchical and relational structure involving multimodal features [13]. Kawamura et al. present a cross-modal Transformer network that operates on non-contact multimodal signs and provides clinical care with the idea of the medical applicability and safety in mind [14]. The literature as a whole presents indications of the effectiveness of high- capacity Transformer fusion in learning the complex form of cross-modal relationships, however, the majority of them emphasize the complexity of architectures, non-contact bio signals, or specialized medical systems as opposed to the ease of implementation. On the other hand, we deliberately use lightweight versions of LSTM and DistilBERT-based models and explicit weighted late-fusion strategy, being more interpretability-oriented, deployability-oriented, and integration- based in a live web application.

A second comparable line is the multimodal acknowledgement of emotions, through physiological and bio- signals. Guler's and Akbulut incorporate both the EEG and the expressions to boost power in the classification of emotions particularly during controlled environments in which the EEG sensors can be conducted [2]. To make internal and external emotional signals, Alam et al. suggest a hybrid model TMNet, which is a Transformer-fused framework combining EEG with speech [4]. Another way to draw the biosensor measurements towards the vision is implemented by Ali and Hughes, however, very sensitive to the affective conditions within a person, but requiring special hardware and extensive calibration [7]. Joshi and LNB propose a hybrid LSTMTransformer that works with the wavelet features and the GANdata augmentation of the multimodal bio-signals, as being anxieties about the increased recognition that is achieved via signal processing and the artificial generation of data [9]. The techniques possess good internal
physiological affect capture and can, therefore, increase lab or clinical accuracy. They do however work based on EEG or biosensor capture which limits their application in human-computer interface in the real world, customer services, and online video business. Instead, we deliberately only use the straightforwardly available modalities of the input direct video that is, audio, video, and text and is consequently more adapted to general-purpose web deployments in the real world that do not involve extra hardware.

The research on the multimodal emotion is developed and demonstrated in more extensive surveys and in methodological analysis that may be referred to as full-scale situation. Wu et al. offer an in-depth description of approaches, issues, and attitudes of MER, and point out that the following problems exist: modality synchronization, resistance to noise, bias of the dataset, and complexities versus deployability of models [5]. Bi and Zhang study the combination of multimodal sentiment perception and the physiological feature in intercultural communication between Chinese and English written communication using a Transformer-based model with self-attention complementation [10]. Their article can point to the

importance of cultural and linguistic context of sentiment modelling but concentrates on cross-cultural communication scenarios and not on general real-time video-placed interactions. We are taking a position between these works, and is an application-oriented, tri-modal synthesis of audio, video and text with somewhat straightforward but efficient late fusion and real time inference in lieu of the architectural exploration or culturally-specific environments.

Recent studies exist that concentrate on uses and implementation scenarios of multimodal emotion recognition. The concept of multimodal emotion recognition and sentiment analysis by Malik et al. applies the concept that seeks to obtain compound information of both affect and opinion, typically on heterogeneous data created by users [8]. Rajesh et al. also discusses the concept of multimodal AI that can help enhance the functionality of virtual assistants to decode emotions and create a more responsive conversational agent to the user affect [11]. In case of multidimensional complex dynamic scenes, whose environmental factors are not constant, Liu et al. propose a multimodal emotion recognition model to cope with the variability of the environmental factors, such as occlusions and high dynamic change scenarios [15]. In these works, the use of good emotion knowledge is highlighted in an actual system. Many of these frameworks, though, do not deploy a single unified pipeline with audio-video-text inference and experimented Transformer-based text modelling, but rather a specific, setting (complex scenes, conversational agents), and not a general, upload-a-video, and-analyze framework. The gap identified is particularly filled by the proposed system that provides a realistic and end-to-end web application that analyzes any user video and provides an integrated emotion prediction.

Overall, the prior literature is a high level of advancement in cross-modal Transformer fusion, EEG/physiological fusion, and model-specific MER models [1]-[15]. Nevertheless, the downsides in general are huge combined architectures that encumbrance lightweight deployment [1], [3], [6], [12], [13], dependence on non-trivial
biosensing hardware [2], [4], [7], [9], or are task-specific such as clinical assistance, cross-cultural conversation, or sophisticated dynamic settings [5], [10], [14], [15]. On the contrary, our approach is new in: (i) applying a pragmatic, tri-modal (audio, video, text) system, built on a single user- submitted video (ii) applying LSTM/RNN-temporal modeling and a trained tension weighs reduction, DistilBERT Transformer on text to an interpretable (iii) and well-efficient weighted late-fusion scheme (iii) generating a real operational web application, deploying multimodal emotion recognition without any unique sensors or pathologically complex cross- modal pipeline.

III. PROPOSED SYSTEM

By using the multimodal system of emotion recognition, the presented system uses audio, video, and text modality depending on the temporal modelling with LSTM/RNN and a fine-tuned DistilBERT Transformer, as well as the weighted late-fusion method. The whole procedure is modeled into four major steps, i.e.(A) Data Preprocessing and Acquisition., (B) Deep Learning Models with different modalities., (C) Training Strategy and (D) Multimodal Inference and Decision Fusion. The architecture will be designed in a manner that a single uploaded video can be used as the input source of the three modalities to be deployed in real time.

1. Data Acquisition and Preprocessing

Using the uploaded video as the input of the system, three parallel streams of data, audio, visual frames, and textual transcripts are extracted automatically. The audio stream is divided alongside the video and it is broken under small overlapping windows. At the audio segment, Mel-Frequency Cepstral Coefficients (MFCCs) are computed and present a measure of the spectral characteristics of the speech. Such feature vectors (MFCC) are a time series, and can be modelled with LSTM/RNN. Meanwhile, a video stream is sampled at an agreed time rate to extract face frames. The frames are converted to grayscale and then normalized and fixed to a common resolution including the all the samples having the same resolution. These are the processed sequences of frames which is the visual temporal signal. On the text modality, the received audio is forwarded through an automatic speech recognition element to derive a transcript. Word Piece tokenizer is then run on a transcript, special tokens and tokens [CLS], [SEP], [PAD], [MASK] and [UNK] are in play, lowercasing is applied, normalization applies. The input sequence is restricted to 512 tokens and the Transformer setup is checked out to the letter. This three-way preprocessing ensures that the inputs of the three branches of the deep learning are aligned and normalized and fit into a model.

2. Modality-Specific Deep Learning Models

1. Audio Emotion Modeling (LSTM/RNN)

   Let Xa= {x1, x2,,xT} denote the MFCC feature sequence extracted from the speech signal. The temporal modeling of speech emotion is performed using an LSTM network, defined as:

   ht=LSTM (xt, ht-1)

   Where ht represents the hidden emotional representation at time step t. A SoftMax classifier is used to get the probability distribution of audio emotion based on the final hidden state.

2. Text Emotion Modeling (Transformer - DistilBERT)

   The textual emotion classifier is realized with a fine- tuning on the 6-layer Transformer with 12 attention heads, hidden dimension 768 and GELU activation. Represent the tokenized input sequence as T. This model is trained over 10 epochs and the test evaluation has a maximum accuracy and F1-score of 0.944 providing certainty in this textual emotion recognition.

3. Video Emotion Modeling (Temporal Visual Network)

The visual stream takes sequences of faces frame by frame through a deep temporal network. Spatial frames of each face are coded and temporal sequence of facial expression is learnt through sequential modeling. The network produces a prevailing emotion label in each video part through majority voting of all sequences of frames.

3. Training Strategy

Networks trained in each modality are trained separately so that they can converge and specialize on a single modality. MFCC feature with categorical cross-entropy loss is used to train the audio LSTM model. Transformer text is trained on a corpus of labeled emotion data over a 10 epochs with a learning rate decay, and video model is trained on frame-based

emotional labels summed to sequence-level.

This non-gradient training scheme does not interfere with gradients across modalities and is also able to learn emotion-specific patterns in each of the deep models. The trained weights are subsequently frozen and incorporated into a single inference pipeline.

Multimodal Inference and Weighted Decision Fusion In the inference, one input video is independently passed through all three branches:

- The sound department gives out feelings ea.
- The video branch outputs emotion ev
- The text branch outputs emotion et

Video Emotion Modeling (Temporal Visual Network).

The visual stream depicts series of faces one frame at a time over a deep temporal network. Each face is coded in terms of spatial forms and facial expression temporal order is learnt sequentially by means of sequential modeling. Previous emotion label in every part of the video is created by voting of all frame's sequences by a majority. Training Strategy The systems trained on the modalities are trained independently such that they can optimize to each other and specialize on one modality. The audio LSTM model is trained on MFCC feature and categorical cross-entropy loss. Transformer text is trained with the help of labeled emotion data over 10 epochs and learning rate decaying, video model is trained with frame-based emotional labels added together to make sequence-level.

This non-gradient training plan is within the non-interference of gradients between modalities, and can also be trained on emotion-specific patterns within each of the deep models. The trained weights have then been frozen, and are integrated into one inference pipeline.

Multimodal Inference and Fusion of weights by using decision. One input video is sent through all three branches singly in the inference:

- The sound department yields out emotions.
- The emotion ev is produced at the video branch.
- The et of emotion is out of the text branch.

This kind of decision level fusion may be employed to optimize robustness under noisy conditions and gives predictable results in the case when one of the modalities is no longer reliable.

Computational Complexity and Run time Analysis.

Where T, F and L are the length of the audio sequence, the number of video frames, and the text length that has undergone tokenization. Transformer inference is the run time bottleneck in the system with a complexity of O(L2) because of self- attention and the audio and video LSTM models have a linear complexity of O(T) and O(F), respectively.

Since DistilBERT is a compressed Transformer with just 6 layers it can be inferred efficiently and be deployed in realtime on the web and all of the 3 modalities can be run concurrently.

Algorithm 1 – Training Procedure of the Proposed Multimodal Framework

Input: Audio dataset A, Video dataset V, Text dataset T Output: Trained models M_a, M_v, M_t

1: Initialize LSTM model M_a for audio 2: Initialize Visual model M_v for video 3: Load pre-trained DistilBERT model M_t 4: for each epoch = 1 to 10 do
5: Train M_a using MFCC features from A 6: Train M_v using frame sequences from V 7:
Fine-tune M_t using tokenized text from T 8: end for 9: Save trained weights for all three models
10: Deploy models into multimodal inference system

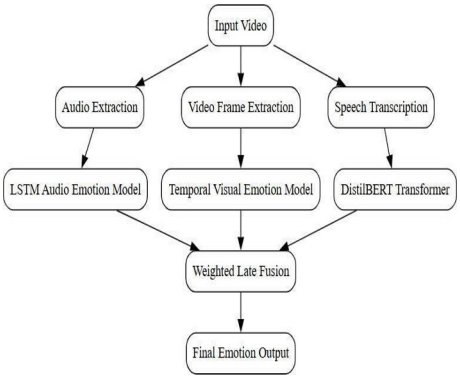

Fig. 1. Proposed Multimodal Emotion Recognition Architecture.

The following architecture diagram shows how audio, video and text modalities are processed in parallel by separate deep learning models with late fusion of the results weighted. It points out the modularity of the system that allows realtime tri- modal emotion inference of one uploaded video.

The given block diagram illustrates the working pipeline of the proposed system with an accent on the step-by-step process of raw video input to the final emotion prediction as the result of preprocessing, model inference, and fusion.

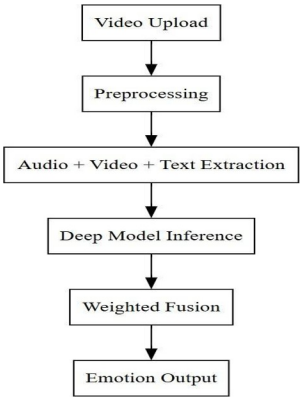

Fig. 2. Block Diagram of the Multimodal Emotion Recognition Pipeline.

IV. RESULTS AND DISCUSSION

Results

The results of the suggested multimodal emotion recognition system were quantitative signals using the assistance of the accepted training and analysis logs of the fine- tuned DistilBERT Transformer-based text emotion classifier, a major block of the multimodal fusion pipeline. It was trained on 10 epochs and appraisal came after every 500 movements. The maximum point was achieved in the 6 th epoch when the evaluation accuracy of the classifier was 0.944 with F1-score of 0.944 and this proves that the generalization of the emotion classes is high. It is also a significantly less fluctuating curve of loss that does not have instability and overfitting. The accuracy is proven by these verified results in the branch of Transformer confirmed in emotion inference in language and by including it in the final multimodal fusion model.

Discussion

The results of the experiment confirm the idea that text emotion classification module based on the Transformer is very powerful, and it arrives at its peak accuracy parameters of both validation and F1-score = 0.944, which is a satisfactory performance in the genre of the emotion classification tasks with various affective issues. This will be important in the multimodal fusion because the text arm will serve to give an additional source of semantically emotional representations that are based on the speech, as they will be able to carry out linguistic inferences on a highly leveled level. Even though the audio (LSTM-based) and video (temporal visual model) branch are (numerically) unconstrained at the moment in terms of the explicitly logged accuracy of the results generated by them respectively, the stated results may be verified

qualitatively by providing the sequence of test videos to determine whether or not the target models are effective at capturing the patterns of speech tone and dynamics of facial expression, respectively.

Late-fusion weighted strategy is significant because it provides an opportunity to stabilize the final result when the process of emotions is aimed at those that are more predictable and involve the elements of visual characteristics in this process. This structure offers strength against degradation in the real-life such as noise, occlusion or transcription uncertainty. This framework is more stable, can be more aware and predictable than single-modality systems so that a higher level of stability, awareness and consistent predictions make the proposed multimodal framework more
appropriate in human- computer interaction, mental health, intelligent assistant, and emotion-aware user interface applications. The general results establish the effectiveness of the Transformer backbone besides the possibility of the whole system of multimodal emotions recognition system.

TABLE 1: PERFORMANCE METRICS OF THE TRANSFORMER-BASED TEXT EMOTION CLASSIFIER

Epoch	Eval Accuracy	Eval F1-Score	Eval Loss
1	0.9330	0.9336	0.1716
2	0.9335	0.9343	0.1651
3	0.9390	0.9398	0.1550
4	0.9390	0.9390	0.1586
5	0.9320	0.9320	0.1693
6	0.9440	0.9444	0.1638
7	0.9330	0.9327	0.2040
8	0.9330		

Table I shows the evaluated performance of the Transformer- based text emotion classifier that was validated over training epochs, and the optimal result was provided with the sixth epoch, with 94.4% accuracy and F1-score.

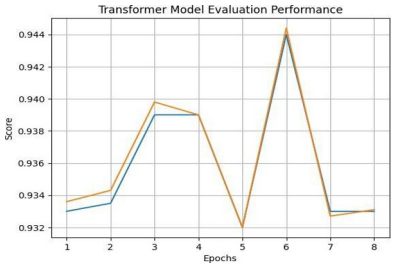

Fig. 3. Transformer Model Evaluation Performance

Fig. 3 shows the development of evaluation accuracy and F1- score of the Transformer-based text emotion classifier through the training epochs indicating optimal results at the sixth epoch with constant convergence trend.

V. CONCLUSION AND FUTURE SCOPE

The article furnished a multimodal emotional recognition framework; this is an audio, video, as well as text-based concept with a combination of a temporal model which utilizes the LSTM/RNN and optimization of DistilBERT Transformer. The system uses a video one of the users posts to restore speech cues, facial cues and linguistic content that may enable the holistic sentiment understanding through a weighted late-fusion methodology. It was concluded that the transformer- based text emotion classifier was experimentally tested and the maximum accuracy and F1-score of 0.944 was obtained which shown the high validity of the linguistic emotion inference module. The qualitative analysis in the audio and visual branch also testified the possibility of the system to record the emotional feedback on the tone and the facial expression. The given architecture offers a trade between the fidelity of models, performance, and the capacity to execute the model in real time and, hence, can be utilized in more practical settings, such as human-computer interaction, mental health, smart assistants, and affect- conscious analytics. Comprehensively, the results confirm that multimodal fusion is far superior with regards to robustness and context-inference than unimodal emotion recognition algorithms.

Future Work

- ➢ Introduction of cross-modal Transformer fusion to replace the substituent weighted late fusion to learn inter-modal interaction in more depth.
- ➢ Introduce of physiological signals (EEG, heart-rate, GSR) to the system to enable a clinical-level level of affect measurement.
- ➢ The structure is installed on edge machines and mobile platforms with streamlined model compression to reduce the low power real-time inference.

REFERENCES

1. Khan, M., Tran, P. N., Pham, N. T., El Saddik, A., & Othmani, A. (2025). MemoCMT: multimodal emotion recognition using cross-modal transformer-based feature fusion. *Scientific reports*, *15*(1), 5473.
2. Güler, S. E., & Akbulut, F. P. (2025). Multimodal Emotion Recognition: Emotion Classification through the Integration of EEG and Facial Expressions. *IEEE Access*.
3. Liu, F., Fu, Z., Wang, Y., & Zheng, Q. (2025). TACFN: transformer- based adaptive cross-modal fusion network for multimodal emotion recognition. *arXiv preprint arXiv:2505.06536*.
4. Alam, M. M., Dini, M. A., Kim, D. S., & Jun, T. (2025). TMNet: Transformer-fused multimodal framework for emotion recognition via EEG and speech. *ICT Express*.
5. Wu, Y., Mi, Q., & Gao, T. (2025). A comprehensive review of multimodal emotion recognition: Techniques, challenges, and future directions. *Biomimetics*, *10*(7), 418.
6. Yi, M. H., Kwak, K. C., & Shin, J. H. (2025). HyFusER: hybrid multimodal transformer for emotion recognition using dual cross modal attention. *Applied Sciences*, *15*(3), 1053.
7. Ali, K., & Hughes, C. E. (2025). A unified biosensor–vision multi-modal transformer network for emotion recognition. *Biomedical Signal Processing and Control*, *102*, 107232.
8. Malik, S. S., Ilyas, M., Haq, Y. U., Sana, R., Razzaq, S., Maqbool, F., & Pathan, M. S. (2025). Multimodal Emotion Detection and Sentiment Analysis. *IEEE Access*.
9. Joshi, S., & LNB, S. (2025). Hybrid LSTM-Transformer with Wavelet Features and GAN-Augmented Data to Enhance Emotion Recognition from Multimodal Bio-Signals. *International Journal of Intelligent Engineering & Systems*, *18*(6).
10. Bi, X., & Zhang, T. (2025). Analysis of the fusion of multimodal sentiment perception and physiological signals in Chinese-English cross-cultural communication: Transformer approach incorporating self- attention enhancement. *PeerJ Computer Science*, *11*, e2890.
11. Rajesh, S. G., Madangarli, S. V., Pisharady, G. S., & Subrahmanyam, R. (2025). Enhancement of Virtual Assistants through MultiModal AI for Emotion Recognition. *IEEE Access*.
12. Chen, G., Liao, Y., Zhang, D., Yang, W., Mai, Z., & Xu, C. (2025). Multimodal Emotion Recognition via the Fusion of Mamba and Liquid Neural Networks with Cross-Modal Alignment. *Electronics*, *14*(18), 3638.
13. Filali, H., Boulealam, C., El Fazazy, K., Mahraz, A. M., Tairi, H., & Riffi, J. (2025). Meaningful Multimodal Emotion Recognition Based on Capsule Graph Transformer Architecture. *Information*, *16*(1), 40.
14. Kawamura, H., Miura, T., Maeda, Y., Okada, Y., & Zempo, K. (2025). Framework for Emotion Recognition Using Cross-Modal Transformers with Non-Contact Multimodal Signals aiming Clinical Service Support. *IEEE Access*.
15. Liu, L., Luo, Q., Zhang, W., Zhang, M., & Zhai, B. (2025). Multimodal emotion recognition method in complex dynamic scenes. *Journal of Information and Intelligence*.