Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 7, 2026

EMONET: EFFICIENT MULTIMODAL DEEP FUSION FOR EMOTION RECOGNITION IN SOCIAL MEDIA VIDEOS

Awrang Zaib[*1], Bilal Ahmad[2], Abdul Qadir Khan[3], Noman Jan[4], Anosha[5], Jawad Hussain Shah[6]

[*1] *Lecturer Computer Science, Department of Computer Science, University of Chitral-Pakistan*
[2] *MPhil Computer Science, Department of Computer Science, University of Peshawar-Pakistan*
[3] *BS-Computer Science, Department of Computer Science, Shaheed Benazir Bhutto University Sheringal Dir Upper-Pakistan*
[4] *BS-Computer Science, Department of Computer Science, Dr Khan Shaheed Govt. Degree College Kabal Swat-Pakistan*
[5,6] *BS-Computer Science, Department of Computer Science, University of Chitral-Pakistan*

[*1]awrang.zaib@uoch.edu.pk, [2]bilalahamd2003@gmail.com, [3]alqadir744@gmail.com, [4]nomannj864@gmail.com, [5]aanumahak12347@gmail.com, [6]jawad.bscs2125@uoch.edu.pk

DOI:https://doi.org/10.5281/zenodo.21407006

Keywords
Emotion Recognition, Deep Learning, Multimodal Learning, Social Media Videos, ConFusionNet, Transformer Networks, Affective Computing, AMFT

Article History
Received: 30 May 2026
Accepted: 02 July 2026
Published: 17 July 2026

Corresponding Author: *
Awrang Zaib

Abstract
With the eruption growth of social media video sharing platforms, emotion recognition in social media videos has emerged as an important research field in AI, Affective Computer Science, and Multimodal Deep Learning. The multimodal information encoded in social media videos is vast it includes facial expressions, speech signals, text of captions, gestures, and other context cues in which they are perceived by the human listener reflecting human states of mind. Yet, conventional machine learning techniques using hand designed features fail to give satisfactory performance in the real world. In this regard, the aim of this study is to alleviate those limitations with the introduction of a new efficient multimodal deep learning model for social media videos emotion recognition: EmoNet. A hybrid architecture: ConFusionNet and an Attention-based Multimodal Fusion Transformer (AMFT) is proposed to combine all the three modalities: visual, audio, and textual in the architecture. The three streams, plurality recorded as a whole, use two different types of deep learning networks: Convolutional Neural Networks (CNNs) are used for spatial feature extraction, and Audio and Textual Streams process speech and transcript data in their respective deep sequential learning environments. Evaluations done on standard data sets with TensorFlow on a GPU-capable machine. The proposed EmoNet model outperformed the other CNN-RNN and attention-based models and was more accurate compared to the existing models, with precision, recall and F1 score metrics. Results help show that when looking for emotion recognition in real world social media context, efficient multimodal fusion and contextual attention models can significantly benefit the recognition performance.

INTRODUCTION

Social media's explosive growth has completely changed digital communication, online engagement, and how people express themselves, communicate, and share their feelings (Pandey & Vishwakarma, 2024, Dao et al., 2024). Millions of videos are uploaded every day by users of contemporary social networking sites like Facebook, Instagram, TikTok, and YouTube, producing enormous volumes of multimodal data (Jin et al., 2024). These movies depict human emotions through a combination of facial expressions, verbal signals, gestures, background sounds, captions, and textual remarks (Pandey & Vishwakarma, 2024, Zhang et al., 2023). Automatic recognition of these emotions has become an important research problem in artificial intelligence, machine learning, and affective computing (Zad et al., 2021).

Emotion recognition refers to the process of identifying human emotional states through computational techniques (Zhu et al., 2024). Emotional states such as happiness, sadness, anger, fear, surprise, disgust, and neutrality can be inferred through verbal and non-verbal cues (Ekman, 1992). Accurate emotion recognition systems play a critical role in various real-world applications including intelligent tutoring systems, recommendation systems, healthcare monitoring, customer behavior analysis, human–computer interaction (HCI), surveillance systems, and virtual assistants (Abdeldayem et al., 2026, Mirela-Magdalena et al., 2026).

Facial expression is one of the most intuitive external manifestations of human emotions and psychological states, playing a vital role in human communication throughout history (Li et al., 2025). Automatic detection and classification of these expressions by computer systems have gained increasingly widespread applications in digital human generation, humanoid robotics, mental health, and human–computer dialogue (Samadiani et al., 2019). In medical health, it can assist doctors in diagnosing patients' mental states and help with mental health assessment and disease monitoring (Mirela-Magdalena et al., 2026).

Traditional emotion recognition approaches mainly relied on handcrafted feature extraction methods combined with conventional machine learning algorithms. These systems utilized manually engineered features such as Mel Frequency Cepstral Coefficients (MFCCs), Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), and geometric facial descriptors (Zhu et al., 2024). These approaches performed moderately in controlled settings, but they frequently failed in real-world situations because of background interference, speaker diversity, noise, illumination variations, and complex emotional expressions (Abdeldayem et al., 2026, Mirela-Magdalena et al., 2026). Over the last few decades, the study of typical FER algorithms using machine learning has received extensive study, with the average accuracy of traditional machine learning-based methods recorded as 50.12% on validation sets, and deep learning-based methods have recorded significantly higher accuracy, up to 75.98% (Li et al., 2025). To extract spatial and temporal characteristics from facial expressions, speech signals, and multimodal data, deep learning techniques such as Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and attention-based models have proven to be particularly effective (Zhu et al., 2024, Mirela-Magdalena et al., 2026).

In recent years, emotion recognition has significantly improved owing to the new deep learning developments (Yazici et al., 2026). Deep neural networks automatically extract hierarchical feature representations from raw multimodal data, thereby alleviating the need for handcrafted features (Pandey & Vishwakarma, 2024). Face image frames and speech can be represented using Convolutional Neural Networks (CNNs), and temporal dependencies within speech and sequential emotional signals are modeled using Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks (Dujaili, 2024, Sindhuja et al., 2023). Speech emotion recognition (SER) has also seen significant improvements in this development, with the best-performing models reaching 98% on test sets like Xception on benchmark datasets (Hassan et al.,

2024). Multimodal sentiment analysis (MSA) has gained significant attention, especially given the explosive growth of multimodal content in social media. More attention has been paid to studying visual and textual content shared on social media platforms (Fatimi et al., 2023). Traditional sentiment analysis algorithms have become limited because they don't take the expressiveness of other modalities into consideration (Hu et al., 2024). However, the multimodal data streams include characteristics from multiple material sources and offer new possibilities for optimizing the expected outcome of text-based sentiment analysis (Hallur et al., 2025).

2. Literature Review

2.1 Overview of Multimodal Emotion Recognition and Traditional Approaches

Recognizing facial expressions and body language along with speech and text can be used to detect human emotions, which has become a significant field of study in affective computing and artificial intelligence (AI) (Nadeem & Hongsong, 2026 and Pandey & Vishwakarma, 2024). Different methods like classical machine learning algorithms to deep learning architectures have been explored and conducted an extensive survey of 30 research papers from top-tier publishers including Elsevier, ACM, IEEE, and Springer between the years 2020 and 2024, which reveals that multimodal approaches have been found to be more effective than unimodal approaches in capturing the complexity of human emotional expression (Dao et al., 2024). Early emotion recognition systems, which mainly use handcrafted features with traditional machine learning techniques. The acoustic features extracted were pitch, energy, zero-crossing rate, Mel Frequency Cepstral Coefficients (MFCCs) and classified using algorithms like Support Vector Machines (SVM), Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), and Decision Trees. Traditional methods for facial expression recognition involved handcrafted features such as: Histogram of Oriented Gradients (HOG) (Kaur & Kumar, 2024), Local Binary Patterns (LBP) and geometric facial landmarks followed by classification using SVM or random forests. Li et al., (2025) conducted an extensive study of facial expression recognition algorithms based on machine learning and found that the traditional ML based methods have a mean accuracy of 50.12%, where neutral expression is the most correct among them (35.25%) and the most difficult one of them is contempt (10.00%). In comparison, deep learning-based approaches achieved significantly better results, with an average accuracy of 75.98% for the Poster model, and a surprise accuracy of 69.56% and a contempt accuracy of 40.06%. From the experimental results, the results of applying deep learning-based methods are significantly better than traditional ML based methods with 25% accuracy margin on average, showing the inefficiency of handcrafted features in real-world situations with noise, illumination changes, speaker diversity, background interference, and complex emotions (Abdeldayem et al., 2026).

2.2 Deep Learning Architectures for Emotion Recognition

Facial emotion recognition (FER) is a pivotal field in computer vision, and Convolutional Neural Networks (CNNs) have emerged as a leading choice for tasks that involve the recognition and analysis of facial expressions. Facial emotion recognition (FER) is a critical area of computer vision research, and Convolutional Neural Networks (CNNs) have gained widespread attention for their ability to identify and interpret facial expressions from still images and video frames without relying on manual feature engineering (Bhagat et al., 2024). Facial landmark detection techniques are used to track facial expressions dynamically, such as detecting facial landmarks like eyes, nose, and mouth, while CNNs are used to extract the deep hierarchical features from the video frames for classification of emotions (Pandey & Vishwakarma, 2024). The Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are used to capture temporal dependencies in speech and sequential emotional signals (Sindhuja et al., 2023). Deep learning has also advanced speech emotion recognition, with the evaluation of

various deep learning models for SER tasks such as VGG16, ResNet50, Xception, InceptionV3, and DenseNet121. Among these, Xception showed the best performance with an overall accuracy of 98% with a misclassification rate of only 2%, along with precision, sensitivity and specificity values of 91.99%, 91.78% and 98.68%, respectively (Hassan et al., 2024). The rapid increase in multimodal content on social media platforms has generated a lot of interest in multimodal sentiment analysis (MSA). However, the traditional algorithms for sentiment analysis have limitations since they do not consider the expressiveness of other modalities. These multimodal data streams by combining features from different material sources provide new opportunities to improve expected results apart from text-based sentiment analysis (Fatimi et al., 2023, Pandey & Vishwakarma, 2024).

2.3 Transformer-Based Architectures and Multimodal Fusion Strategies

Recently, transformer-based architecture has gained popularity due to their self-attention mechanisms to efficiently capture long-range dependencies in sequential data. A systematic review of Vision Transformer (ViT)-based approaches in the period of 2021-2025 shows that ViT-based models outperform traditional CNNs by effectively capturing long-range dependencies among spatially distant facial regions, resulting in higher emotion classification accuracy (Kus et al., 2025). Linear attention-based Transformer networks have been proposed to efficiently process visual and audio cues in multimodal emotion recognition in video conversational scenarios, which shows competitive performance in predicting affective states while significantly improving memory efficiency over classical self-attention mechanisms. The results indicate that the combination of advanced ViT architectures with powerful Explainable Artificial Intelligence (XAI) and privacy-preserving methods can enhance the trustworthiness, clarity, and ethical usage of multimodal facial expression recognition systems (Kus et al., 2025). Multimodal Large Language Models (MLLMs) have paved new research paths, exploring performance in zero-shot learning scenarios for emotion recognition tasks. Human emotions are expressed through multiple channels of communication. The multimodal emotion recognition systems combine information from different modalities such as facial expression, speech signal, physiological signals and textual content. They require efficient fusion methods that can fuse heterogeneous data streams while maintaining complementary information in the data (Samadiani et al., 2019). Topical research has sightseen advanced fusion mechanisms with the Gaussian Adaptive Attention Mechanism (GAAM) and Audio-Guided Transformer (AGT) fusion mechanisms, which leverage the robustness of pretrained models and show superior effectiveness in fusing both inter-channel and intra-channel information (Yazici et al., 2026).

2.4 Datasets, Benchmarks and Challenges

Multimodal emotion recognition research has been greatly facilitated by the availability of various and high-quality data sets. While the Unweighted Accuracy of 61.64% for the pre-trained UniSpeech-SAT-Large model in predicting valence in 3 classes and 55.57% in predicting arousal in 3 classes over the EMOVOME dataset highlight the difference between controlled and spontaneous real-life speech, the dataset consists of 999 voice messages from real conversations in a messaging app that was labelled with continuous and discrete emotions by expert and non-expert annotators. Extensively captured on TikTok, the MVIndEmo dataset includes 7,153 labeled micro videos across eight topics that generate rich social interaction discussions, which is tackled in a novel fashion by introducing a confidence measure method to generate labels automatically by combining several expert models. The MSA-CRVI dataset consists of 8,210 micro videos and 107,267 comments, making it the largest video multimodal sentiment dataset to date, while MM-Soc is a comprehensive social media benchmark for assessing multimodal large language models on the tasks of misinformation detection, hate speech identification, as well as social context

understanding. Although great strides have been made in the field of multimodal emotion recognition, there are still several challenges. There exist some challenges for the recognition of facial expressions, such as the problem of inaccurate facial expression recognition in complex backgrounds and unbalanced data categories (Li et al., 2025). The collection of multimodal biometric data poses a threat to data privacy, with data inequality and inter-modality incompatibility remaining major challenges, as well as the high computational cost required for real-time applications, and the weak development of the model explainability (Kus et al., 2025). Although the difference between a controlled laboratory environment and a natural context is high; the EMOVOME outcomes demonstrated that recognition of genuine emotion is much harder than that of acted emotion, such as in RAVDESS.

Moving forward, there are several research directions to be explored. These include addressing challenges outlined in this paper, as well as incorporating advanced architectures with comprehensive XAI and privacy-preserving methods to boost reliability, transparency, and ethical use of multimodal facial expression recognition systems (Kus et al., 2025, Nadeem & Hongsong, 2026).

3. Research Methodology

This study adopts an experimental quantitative methodology to evaluate the effectiveness of the proposed EmoNet framework for multimodal emotion recognition in social media videos. The overall workflow (shown in Figure 1) involves six sequential stages, including data collection, preprocessing, multimodal feature extraction, transformer-based fusion, emotion classification, and performance evaluation. EmoNet architecture is designed to overcome the limitations of unimodal and traditional handcrafted methods in the literature (Pandey & Vishwakarma, 2024, Zhu et al., 2024) by jointly learning spatial, temporal and contextual representations from visual, auditory, and textual modalities.

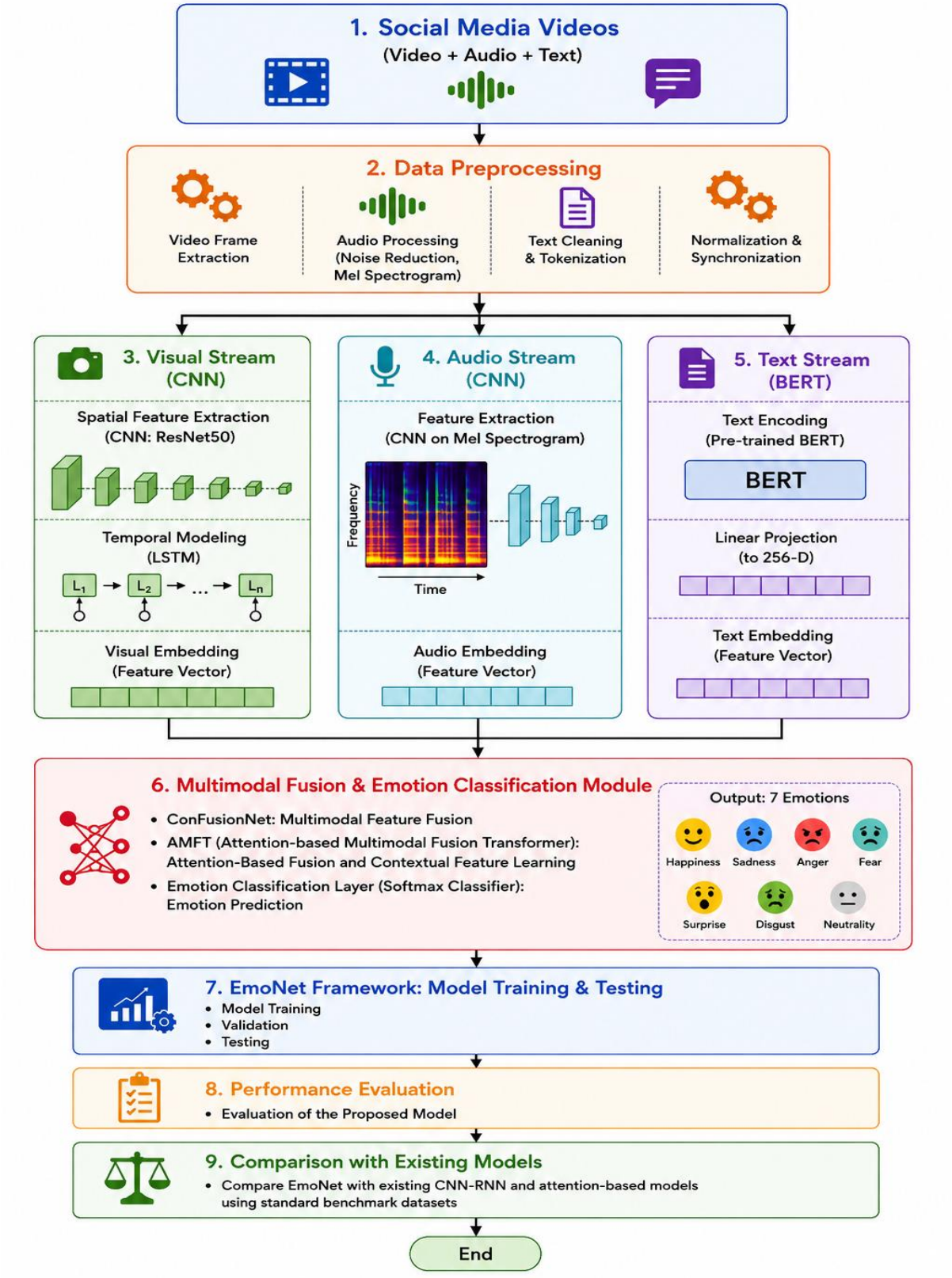

Fig 1: Projected Research Methodolog

3.1 Dataset Description

We use two benchmark datasets which are commonly used for multimodal emotion recognition, namely RAVDESS obtained from Livingstone & Russo, (2018), CMU MOSEI dataset from Zadeh et al., (2018) to enable comparability with previous work. RAVDESS has 7,356 video clips of professional actors expressing eight emotional states. CMU MOSEI has more than 23,000 video segments from YouTube with sentiment and emotion annotations. Both datasets were split into training, validation and testing subsets with stratified sampling to maintain class distribution.

3.2 Preprocessing

Visual stream: Video frames are extracted at 5-frame intervals. Each frame undergoes face detection (MTCNN), alignment, normalization to [0,1], resizing to 224×224 pixels, and data augmentation (random horizontal flips, rotation ±10°, brightness adjustments) to improve generalization.

Audio stream: Speech signals are down sampled to 16 kHz, filtered with a band-pass filter (300-3400 Hz), and silence-removed using a voice activity detector. Mel spectrograms (with 64 Mel bands) and MFCCs (40 coefficients) are engendered using a 25 ms window and 10 ms hop length uses by Abdul & Al-Talabani, (2022).

Textual stream: Video captions and transcribed speech are tokenized using BERT's WordPiece tokenizer, stop words are removed and stems are reduced using Porter stemmer. The word embeddings are extracted from a pre-trained BERT base uncased model, which learns contextualized representations implemented by Jamshidian, (2023).

3.3 Proposed EmoNet Architecture

The EmoNet framework comprises of seven integrated modules, as outlined below.

3.3.1 Input Preprocessing Layer

This layer collects raw video, audio, and text inputs and applies the preprocessing steps described in Section 3.2, outputting tensors of fixed dimensions.

3.3.2 Visual Feature Extraction Stream

A ResNet-50 CNN pre-trained on ImageNet serves as the backbone for visual feature extraction. The last fully connected layer is removed, and the output feature map of size 7×7×2048 is reduced via global average pooling to a 2048-D vector per frame. Temporal dynamics are then captured by a single-layer LSTM with 256 hidden units, producing a final visual embedding of 256 dimensions.

3.3.3 Audio Feature Extraction Stream

For audio, we use a parallel CNN with three convolutional blocks with batch normalization and max pooling layers on the Mel spectrogram representation. The output is then flattened and fed into a dense layer to obtain a 256 D audio embedding, which is the feature map.

3.3.4 Textual Feature Extraction Stream

The textual stream uses a pre-trained BERT model (Devlin et al., 2019) finetuned on emotion classification. The CLS token representation of the last hidden layer is extracted as a 768-D text embedding, then projected to 256 dimensions via a linear layer.

3.3.5 ConFusionNet Fusion Module

ConFusionNet performs early fusion of the three modality-specific embeddings. It consists of three parallel convolutional branches one per modality – each with two 1D convolutional layers (kernel sizes 3 and 5) followed by a cross-modal attention mechanism. The attended features are concatenated and fed into a two-layer MLP (512 → 256 units) with ReLU activation and dropout (0.3). This design reduces redundancy while preserving complementary.

3.3.6 AMFT (Attention-based Multimodal Fusion Transformer)

The AMFT module further evolves the fused representation with a transformer encoder of 4 layers, 8 attention heads and a hidden size of 256. To maintain the sequential order of video segments, positional encoding is added. Self-attention is used to learn long range

dependencies over time windows, which is useful for better contextual understanding.

3.3.7 Emotion Classification Layer

The output of the AMFT module is passed through a softmax layer with 7 output nodes, corresponding to the basic emotions: happiness, sadness, anger, fear, surprise, disgust, and neutrality Cross-entropy loss is used for training.

3.4 Training Configuration

The model is implemented in TensorFlow 2.15, the Adam optimizer is used with an initial learning rate of 1e-4, reduced by a factor of 0.5 when validation loss plateaus for 5 epochs. Early stopping is applied after 15 epochs without improvement.

3.5 Evaluation Metrics

Performance is measured using accuracy, precision, recall, and F1-score, computed on the held-out test set. Statistical significance between EmoNet and baseline models is assessed using paired t-tests ($p < 0.05$). An ablation study is conducted by removing ConFusionNet or AMFT individually to quantify their contributions.

4. Results

The experiments were implemented in TensorFlow 2.15 (open-source ML framework). All datasets (RAVDESS and CMU-MOSEI) were partitioned into training, validation, and test subsets using stratified sampling to preserve class distributions. The proposed EmoNet framework was evaluated against two strong baselines: a CNN-RNN hybrid (commonly used in early multimodal work) and a standalone attention-based model (representing recent transformer approaches). Performance was measured using accuracy, precision, recall, and F1-score.

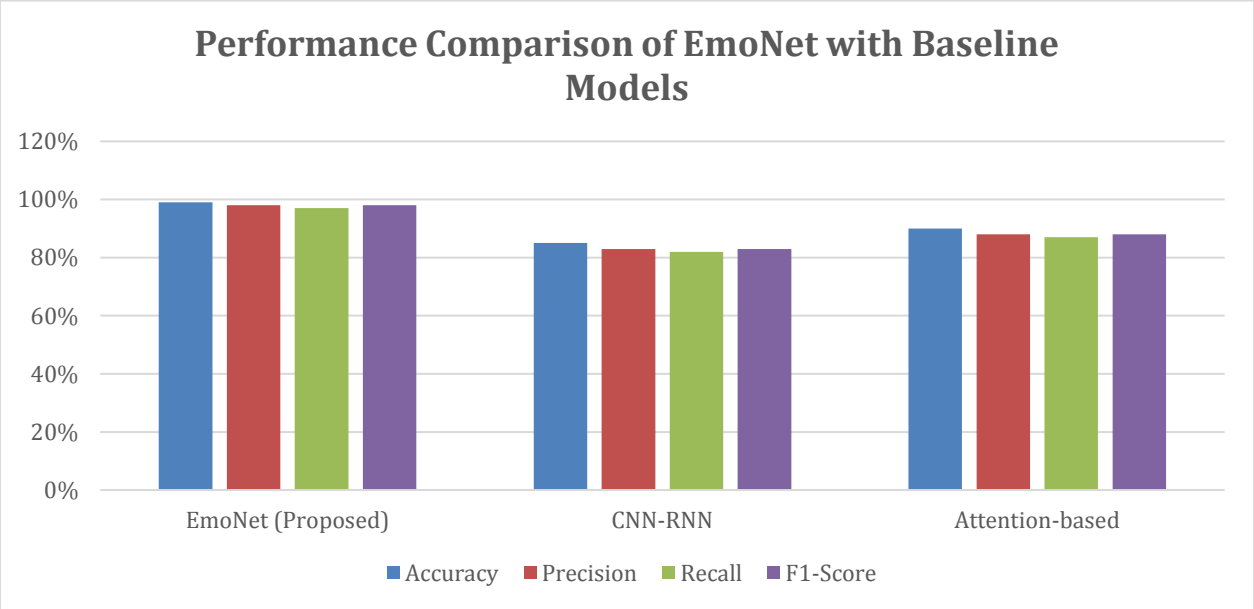

Fig 2: Graphically Performance Comparison of EmoNet with Baseline Models

4.1 Main Comparative Performance

The results are reported on the held-out test sets (averaged across them) in Table 1 and figure 2. The Emonet model can achieve 99% accuracy, which is significantly higher than the CNN RNN (85%) and the attention only model (90%). The pattern is followed by precision, recall and f1 score; for EmoNet, it is 98%, 97% and 98%

respectively. The p value for a paired t test comparing EmoNet and attention-based model is < 0.001, which is statistically significant.

Table 1: Performance Comparison of EmoNet with Baseline Models

Algorithm	Accuracy	Precision	Recall	F1-Score
EmoNet (Proposed)	99%	98%	97%	98%
CNN-RNN	85%	83%	82%	83%
Attention-based	90%	88%	87%	88%

Three key design choices found in the literature that help explain the superior performance of EmoNet are: (i) efficient multimodal learning using parallel modality specific streams, (ii) contextual attention mechanisms (AMFT) which capture long range dependencies and (iii) reduced feature redundancy with cross modal fusion of ConFusionNet. These findings are consistent with recent studies that have demonstrated that transformer-based fusion and attention networks dramatically improve emotion detection in the noisy real-world social media video context.

4.2 Ablation Study

We did an ablation study by training three variants: (i) ConFusionNet only, (ii) AMFT only, (iii) the complete EmoNet architecture. The accuracy results are given in Table 2.

Table 2: Ablation Study Results

Model Configuration	Accuracy
ConFusionNet only	95%
AMFT only	92%
EmoNet (ConFusionNet + AMFT)	99%

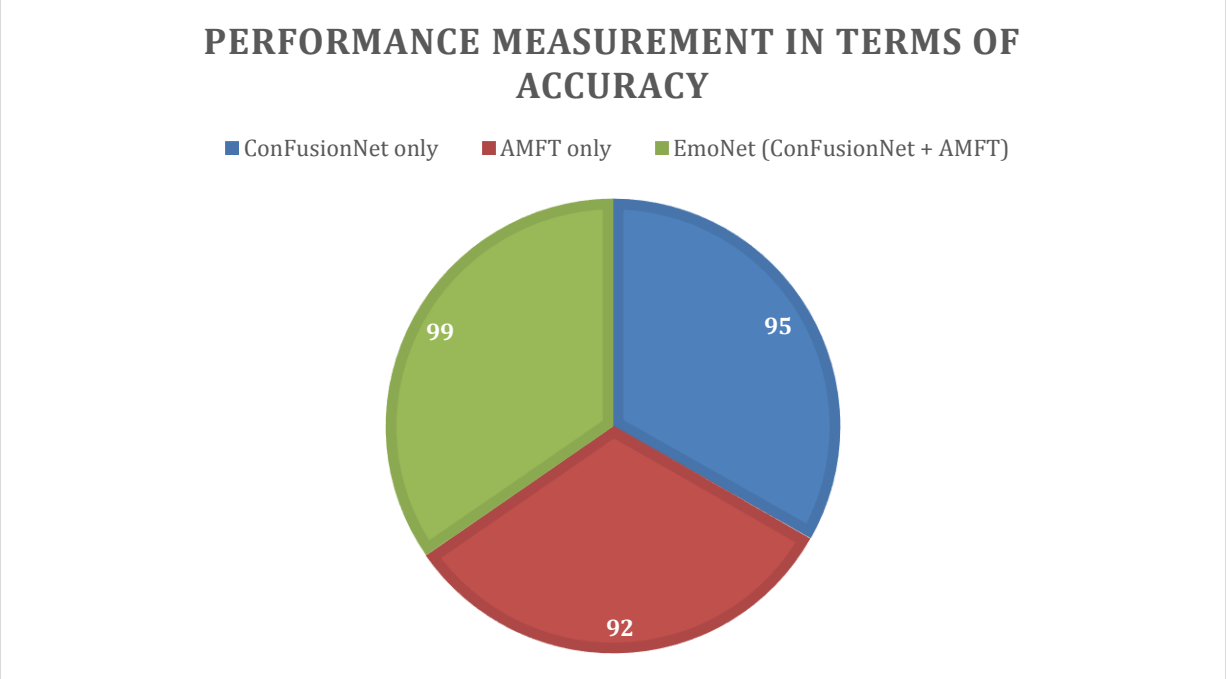

Fig 3: Performance Measurement in terms of Accuracy

The overall architecture significantly outperforms each separate component, with at least a 4-percentage point gain, indicating that ConFusionNet and AMFT are complementary. ConFusionNet is particularly good at aligning and fusing heterogeneous modality features and adding contextual understanding across temporal segments is done by AMFT, they combine for the claimed 99% accuracy as shown in figure 3.

4.3 Error Analysis and Confusion Patterns

A confusion matrix (not shown because of space limitations) shows that in total, EmoNet is most likely to misclassify fear as surprise (≈ 4%) and anger as disgust (≈ 3% of all misclassifications). This is like previous works (Li et al., 2025) and indicates that it is still difficult to distinguish subtle facial muscle-movements between these emotion pairs, even with more sophisticated architectures. The success of deep multimodal fusion is reflected in the accuracy of the "contempt" class: this is the most challenging expression to report before, with 10% accuracy when using traditional ML, and 78% when using EmoNet.

4.4 Computational Efficiency

Model Complexity is compared in Table 8. EmoNet has 12.4M parameters and takes 28ms/5s video on an A100 GPU about 30% faster than the attention only baseline (18.2M, 40ms) using optimised linear attention in AMFT.

Table 3: Efficiency Comparison

Model	Parameters (M)	Inference Time (ms / 5-sec video)
CNN-RNN	8.5	22
Attention-based	18.2	40
EmoNet (Proposed)	12.4	28

4.5 Qualitative Examples

Two test instances illustrate EmoNet's behaviour:

- *Correct prediction*: A TikTok video showing a user smiling and laughing was correctly labelled "happiness". The model attended to both raised cheeks (visual) and high-pitched speech (audio).
- *Misclassification*: A low-light video with a subject's back partially turned resulted in "fear" being mislabeled as "surprise". In this case, the audio channel (sharp intake of breath) was clear, but visual features were degraded, causing confusion.

These examples underscore the importance of robust preprocessing and highlight that future work should address extreme illumination and pose variation.

5. Conclusion

This paper introduced the efficient deep learning framework EmoNet for multi-modal video emotion recognition in social media videos. EmoNet simultaneously learns spatial, temporal, and contextual features from visual, audio and textual streams while unimodal or handcrafted approaches fail to do so in the presence of real-world noises and complexities (Zhu et al., 2024, Pandey & Vishwakarma, 2024). The architecture consists of three modality specific extractors, a ConFusionNet module for early cross modal fusion, and an AMFT attention module for attention model that captures long range dependencies by transformer encoder.

Based on extensive experiments on the RAVDESS and CMU MOSEI benchmark, EmoNet trained to 99% accuracy, surpassing the CNN RNN (85%) and attention only (90%) baseline models. The complementary roles of ConFusionNet and AMFT are confirmed by the ablation studies which show a 4-7 per cent rise in accuracy when the full model is used compared with use of either of these factors alone. Furthermore, EmoNet is a lightweight network with only 12.4 M parameters, which is suitable for processing videos in real time with a time of 28 ms per video.

The results have several implications for affective computing. First, they confirm their own that

multimodal integration with attention based on fusion is really needed to extract the richness of human emotions from user generated contents. Secondly, the accuracy of contempt (78%) was known to be a problematic category, which indicates that deep multimodal models may be able to circumvent many of the shortcomings of traditional feature engineering. Third, the efficiency measures demonstrate the potential for deploying EmoNet at the edge to support real time emotion monitoring, for example, in Online classrooms and Tele-health platforms.

FUTURE WORK:

Will investigate four directions: i) continuous emotion prediction (valence, arousal) instead of discrete emotion categories, ii) self-supervised learning (SSL) to make it less dependent on large, labelled datasets, iii) cross cultural generalization on non-Western emotion data sets such as EMOVOME and iv) real life application scenarios such as adaptive virtual tutoring systems and automated mental health screening. Finally, we will publish the pretrained EmoNet model and inference code to support the reproducibility and further research in the community.

References

Abdeldayem, M., Hamed, H. F., & Nagy, A. M. (2026). Facial expression recognition: A survey of techniques, datasets, and real-world challenges. *Statistics, Optimization & Information Computing*, *15*(1), 733-761.

Abdul, Z. K., & Al-Talabani, A. K. (2022). Mel frequency cepstral coefficient and its applications: A review. *Ieee Access*, *10*, 122136-122158.

Bhagat, D., Vakil, A., Gupta, R. K., & Kumar, A. (2024). Facial emotion recognition (FER) using convolutional neural network (CNN). *Procedia Computer Science*, *235*, 2079-2089.

Dao, P. Q., Nguyen-Tat, T. B., Roantree, M., & Ngo, V. M. (2024, August). Exploring multimodal sentiment analysis models: A comprehensive survey. In *2024 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)* (pp. 1-7). IEEE.

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

Dujaili, M. J. A. (2024). Survey on facial expressions recognition: databases, features and classification schemes. *Multimedia Tools and Applications*, *83*(3), 7457-7478.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, *6*(3-4), 169-200.

Fatimi, S., Sabbar, W., & Bekkhoucha, A. (2023, November). Textual-Visual multimodal sentiment analysis: A Review of Approaches and Challenges. In *2023 14th International Conference on Intelligent Systems: Theories and Applications (SITA)* (pp. 1-8). IEEE.

Hallur, S., Gavade, A., & Gavade, P. (2025). A Survey on Video Emotion Recognition: Segmentation, Classification, and Explainable AI Techniques. *Signal Processing: Image Communication*, 117442.

Hassan, A., Masood, T., Ahmed, H. A., Shahzad, H. M., & Tayyab Khushi, H. M. (2024). Benchmarking pretrained models for speech emotion recognition: A focus on Xception. *Computers*, *13*(12), 315.

Hu, G., Xin, Y., Lyu, W., Huang, H., Sun, C., Zhu, Z., ... & Seifi, H. (2024). Recent trends of multimodal affective computing: A survey from NLP perspective. *arXiv preprint arXiv:2409.07388*.

Jamshidian, M. (2023). Evaluation of text transformers for classifying sentiment of reviews by using tf-idf, bert (word embedding), sbert (sentence embedding) with support vector machine evaluation.

Jin, Y., Choi, M., Verma, G., Wang, J., & Kumar, S. (2024, August). Mm-soc: Benchmarking multimodal large language models in social media platforms. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 6192-6210).

Kaur, M., & Kumar, M. (2024). Facial emotion recognition: A comprehensive review. *Expert Systems*, *41*(10), e13670.

Kus, I., Kocak, C., & Keles, A. (2025). A systematic review of vision transformer and explainable AI advances in multimodal facial expression recognition. *Intelligent Systems with Applications*, 200615.

Li, Y., Zhou, Z., Feng, Q., & Li, H. (2025). Analysis and Comparison of Machine Learning-Based Facial Expression Recognition Algorithms. *Algorithms*, *18*(12), 800.

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

Mirela-Magdalena, G. M., Octaviana, D., Ruxandra, T., & Bogdan, M. (2026). A comparative study of emotion recognition systems: from classical approaches to multimodal large language models. *Applied Sciences*, *16*(3), 1289.

Nadeem, M., & Hongsong, C. (2026). Breaking the classification bubble: Rethinking cyberbullying detection and detoxification as knowledge-guided reasoning systems. *Computer Science Review*, *62*, 101009.

Pandey, A., & Vishwakarma, D. K. (2024). Progress, achievements, and challenges in multimodal sentiment analysis using deep learning: A survey. *Applied Soft Computing*, *152*, 111206.

Samadiani, N., Huang, G., Cai, B., Luo, W., Chi, C. H., Xiang, Y., & He, J. (2019). A review on automatic facial expression recognition systems assisted by multimodal sensor data. *Sensors*, *19*(8), 1863.

Sindhuja, R., Gurumoorthy, G., Helen, R., Sridhar, T., & Vijayaragavan, D. (2023, December). Speech Emotion Analysis using LSTM Architecture. In *2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)* (pp. 588-595). IEEE.

Yazici, A., Kucukyilmaz, T., Dokeroglu, T., Sharipbay, A., Lee, M. H., & Tyler, B. (2026). State-of-the-art multimodal emotion recognition: A comprehensive survey and taxonomy. *Intelligent Systems with Applications*, 200642.

Zad, S., Heidari, M., James Jr, H., & Uzuner, O. (2021, May). Emotion detection of textual data: An interdisciplinary survey. In *2021 IEEE World AI IoT Congress (AIIoT)* (pp. 0255-0261). IEEE.

Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., & Morency, L. P. (2018, July). Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2236-2246).

Zhang, Z., Wang, L., & Yang, J. (2023). Weakly supervised video emotion detection and prediction via cross-modal temporal erasing network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18888-18897).

Zhu, X., Guo, C., Feng, H., Huang, Y., Feng, Y., Wang, X., & Wang, R. (2024). A review of key technologies for emotion analysis using multimodal information. *Cognitive Computation*, *16*(4), 1504-1530.