AVaTER: A Multimodal Approach of Recognizing Emotion using Cross-modal Attention Technique

Avishek Das , Moumita Sen Sarma , Mohammed Moshiul Hoque * , Nazmul Siddique * , M. Ali Akber Dewan

Posted Date: 11 July 2024

doi: 10.20944/preprints202407.0917.v1

Keywords: Multimodal Emotion Recognition; Natural Language Processing, Multimodal Dataset; Cross-Modal Attention, Transformers

Article

AVaTER: A Multimodal Approach of Recognizing Emotion Using Cross-Modal Attention Technique

Avishek Das [1], Moumita Sen Sarma [1], Mohammed Moshiul Hoque [1]*, Nazmul Siddique [2]* and M. Ali Akber Dewan [3]

[1] Dept. of Computer Science and Engineering, Chittagong University of Engineering and Technology {avishek, moumita, moshiul_240}@cuet.ac.bd
[2] School of Computing, Engineering and Intelligent Systems, Ulster University; nh.siddique@ulster.ac.uk
[3] School of Computing and Information Systems, Faculty of Science and Technology, Athabasca University; adewan@athabascau.ca
* Correspondence: moshiul_240@cuet.ac.bd; nh.siddique@ulster.ac.uk

Abstract: Multimodal emotion classification (MEC) involves analyzing and identifying human emotions by integrating data from multiple sources, such as audio, video, and text. This approach leverages the complementary strengths of each modality to enhance the accuracy and robustness of emotion recognition systems. However, one significant challenge is effectively integrating these diverse data sources, each with unique characteristics and levels of noise. Additionally, the scarcity of large, annotated multimodal datasets in Bangla limits the training and evaluation of models. In this work, we unveiled a pioneering multimodal Bangla dataset, MAViT-Bangla (Multimodal Audio Video Text Bangla dataset). This dataset, comprising 1002 samples across audio, video, and text modalities, is a unique resource for emotion recognition studies in the Bangla language. It features emotional categories such as anger, fear, joy, and sadness, providing a comprehensive platform for research. Additionally, we developed a framework for emotion recognition that employs a cross-modal attention mechanism among unimodal features. This mechanism fosters the interaction and fusion of features from different modalities, enhancing the model's ability to capture nuanced emotional cues. The effectiveness of this approach was demonstrated by achieving an F1-score of 0.64, a significant improvement over unimodal methods.

Keywords: multimodal emotion recognition; natural language processing; multimodal dataset; cross-modal attention; transformers

1. Introduction

Emotion classification is pivotal in advancing human-computer interaction, making technological applications more intuitive, responsive, and effective. As digital technologies become increasingly integrated into our daily lives, from virtual assistants on smartphones to customer service chatbots and mental health monitoring apps, recognizing and responding to human emotions accurately becomes crucial. It enables these systems to adjust their responses based on emotional cues, creating personalized experiences and facilitating better healthcare, education, and entertainment decision-making. Emotion classification enables systems to bridge the communication gap between humans and machines. By fostering more natural interactions that mirror human-to-human exchanges, technology becomes more accessible and beneficial to diverse global populations. For example, a virtual assistant who understands when a user is frustrated can respond empathetically and helpfully, improving user satisfaction and engagement.

Various techniques have been developed for emotion classification, encompassing both unimodal and multimodal approaches. Unimodal emotion classification refers to analyzing and interpreting human emotions based on data from a single source or modality. This approach uses just one type of input-text, audio, or visual data-to determine emotional states. For instance, a text-based system might analyze the words and phrases used in a social media post to identify the user's emotional state. On the other hand, MEC combines data from multiple sources, such as text, audio, and visual inputs, to analyze and predict emotions. This approach is more robust and can interpret complex emotions more effectively. For example, in a video call, a multimodal system could analyze the spoken words

 doi:10.20944/preprints202407.0917.v1

(text), the tone of voice (audio), and facial expressions (visual) to gain a comprehensive understanding of the user's emotional state. This comprehensive analysis is particularly beneficial in applications that require deep understanding and insight, such as therapeutic settings or advanced customer service platforms.

In this work, we aim to push the boundaries of emotion classification by developing a computational system capable of identifying four principal categories of emotion: joy, sadness, fear, and anger. Our approach utilizes a multimodal methodology, integrating audio, video, and text inputs to achieve a more nuanced and accurate classification of emotions. This multimodal approach allows the system to capture emotional expressions, often conveyed through verbal and non-verbal cues. One of our key motivations is to bridge the resource gap for Bengali-specific datasets used in multimodal emotion analysis. There is a significant need for such datasets, which hampers the development of emotion recognition systems tailored to the Bengali-speaking population. By contributing to creating and expanding these datasets, we aim to support the development of more effective and culturally relevant emotion classification systems for the Bengali language. The specific contributions of this work are:

- To develop MAViTE-Bangla, a multi-modal Bangla emotion dataset containing 1002 multimodal data labeled into four classes: anger, fear, joy, and sadness.
- To develop a pairwise cross-attention-based multimodal framework for effective emotion recognition and exploit several feature extraction and fusion methods to utilize multimodal features for MEC.
- To analyze the classification outcomes of the proposed method with a detailed investigation of the misclassification of samples

This research aims to advance the technical capabilities of emotion classification systems and ensure these advancements are inclusive and applicable to a broader range of languages and cultural contexts. We aim to create technologically advanced and socially impactful systems by integrating multiple modalities and developing Bengali-specific resources, ultimately enhancing human-computer interactions across different cultures and languages.

2. Related Work

Multimodal emotion classification has evolved significantly, driven by advancements in machine learning and an increased understanding of how emotions are conveyed through multiple channels. Early research primarily focused on unimodal approaches, analyzing data types like text, audio, or visual cues. Single modalities complement each other, so combining them offers a richer, more contextual view, enhancing their generalization ability [1]. Most multimodal emotion classification research uses English and other widely known European languages. However, there is a growing desire to broaden this research to include underrepresented languages like Bengali, which presents unique opportunities and problems for furthering this field of study. This section overviews the literature on unimodal and multimodal emotion categorization.

2.1. Unimodal-based Emotion Recognition

Haque et al. [2] proposed a CNN-LSTM deep learning technique to classify four sentiment classes: sexual, religious, political, and acceptable. They used a 42,036 labeled Facebook comments dataset, with a web application developed for real-time sentiment prediction. This model achieved 85.8% accuracy, but the authors did not consider transformer-based models that might perform better. To categorize nine kinds of emotion, Islam et al. [3] introduced EmoNaBa (Bangla language corpus), and a hybrid model was presented using transformers and lexicon features based on this dataset [4]. However, in this instance, the model's effectiveness depends on having an extensive vocabulary of emotions, which calls for constant updating as new words and dialects appear. Three transformer models, m-BERT, BanglaBERT, and XLM-R, were assessed by Das et al. [5] utilizing a variety of DL and

 doi:10.20944/preprints202407.0917.v1

ML models on a corpus BEmoC [6]. This work obtained the highest weighted F1-score of 69.73% with XLM-R. Rahman et al. [7] developed a dynamic strategy leveraging the Word2Vec model, employing both Skip-Gram and Continuous Bag of Words (CBOW) techniques, focusing on three emotion classes: happy, angry, and excited. This work may not capture the full range of human emotions in the text due to the lower number of classes. In contrast, the authors of [8] utilized various ML and neural network models to classify emotions from Bengali song lyrics, achieving moderate accuracy in multi-class classification. Employing an ensemble technique with CNN, GRU, and BiLSTM, Parvin et al. [9] classified six emotion categories from a new corpus containing 9000 annotated texts.

Regarding speech emotion detection, Sultana et al. [10] created a novel architecture called DCTFB, fusing BiLSTM with deep convolutional neural networks (DCNN) that classify emotions from Bengali and English speech along with song corpus. However, the model's generalizability may be impacted by the imbalanced speaker gender and emotion classes in the RAVDESS dataset employed in this work. Similarly, in [11], three kinds of emotions (angry, happy, and neutral) were classified in Bengali speech corpus using ML methods using the MFCC (Mel-Frequency Cepstral Coefficients) and LPC (Linear Prediction Coefficients) features extracted from audio signals. Conversely, five pre-trained deep learning models were used in [12] to categorize hateful and non-hateful emotions in a new dataset of 3000 hand-labeled Bengali memes. However, this work should have considered the multimodality of memes, where images and text jointly convey the whole meaning of the content.

2.2. Multimodal-based Emotion Recognition

Regarding MEC, Ghosh et al. [13] created an emotion corpus called MELD, which includes seven classes and combines visual, textual, and audio modalities. They also evaluated the performance of several baseline models on this dataset. However, misclassification occurred when subtle emotions such as disgust and fear were classified. Hu et al. [14] presented a unified framework by merging multimodal sentiment analysis (MSA) and emotion recognition by fusing syntactic and semantic levels across the modalities on four public benchmark datasets, including MELD. Similarly, a multimodal feature fusion method was proposed in [15] to address unbalanced sample data in social network public opinion analysis. It utilized text and speech emotion features, introduced the MA2PE speech feature retrieval method, and addressed sample disequilibrium through data processing techniques using IEMOCAP and MELD datasets. Additionally, Hossain et al. [16] proposed a model that effectively integrated different specialized architectures for each modality: CNN-LSTM for processing audio features, Inception-ResNet-v2 for extracting video features, and Word2Vec for handling text features. This comprehensive approach enabled the model to recognize four distinct emotion categories from the IEMO-CAP dataset. Similarly, the authors in [17] employed a combination of advanced techniques, using BERT for text features, wav2vec2.0 for audio features, and videoMAE for video features. They used an early fusion strategy and SVM classification to enhance the emotion recognition accuracy. However, the authors of [18] introduced an attention-based multimodal emotion detection system that fuses facial and speech features extracted by independent encoders. The system uses convolutional neural networks (CNNs) to process these features and employs an attention mechanism to focus on the most informative parts evaluated on IEMOCAP and CMU-MOSEI datasets. Apart from multimodal emotion and sentiment recognition, to detect the risk of depression, a study in [19] employed the Audio, Video, and Text Fusion-Three Branch Network (AVTF-TBN), integrating three main networks: an audio feature extractor using a convolutional neural network (CNN), a video feature extractor using a 3D CNN, and a text feature extractor using a BiLSTM network.

Regarding the Bengali language, the authors in [20] applied feature fusion and decision fusion techniques between textual and visual modalities to classify emotions in Bangla social media content despite a significantly imbalanced dataset. Hossain et al. [21] introduced a multimodal method to classify hateful memes using the newly created MUTE dataset, including Bengali and code-mixed captions, leveraging visual and textual modalities. Similarly, the researchers in [22] employed the Multimodal Attentive Fusion (MAF) model to identify five categories of aggression in a Bengali meme

 doi:10.20944/preprints202407.0917.v1

dataset by integrating visual and textual features. In Bengali, no research is available to date combined audio, video, and text modalities for emotion classification using a multimodal approach. This comprehensive integration is the central focus of this work.

3. Dataset Description

In this research, we have created a comprehensive Bengali Multimodal Emotion Recognition Dataset, which we have named **MAViTE-Bangla**. This dataset comprises 1002 samples, each encompassing audio, video, and text modalities, providing a rich, multimodal resource for emotion recognition studies. The MAViTE-Bangla dataset is categorized into four distinct emotional classes: anger, fear, joy, and sadness, ensuring a diverse representation of emotional expressions.

3.1. Data Collection

This work manually collected YouTube videos from multiple domains, such as movies, dramas, and video blogs (vlogs). This was done to ensure that our dataset included a wide range of emotions and scenarios. Figure 1 shows a detailed breakdown of the domains represented in the dataset.

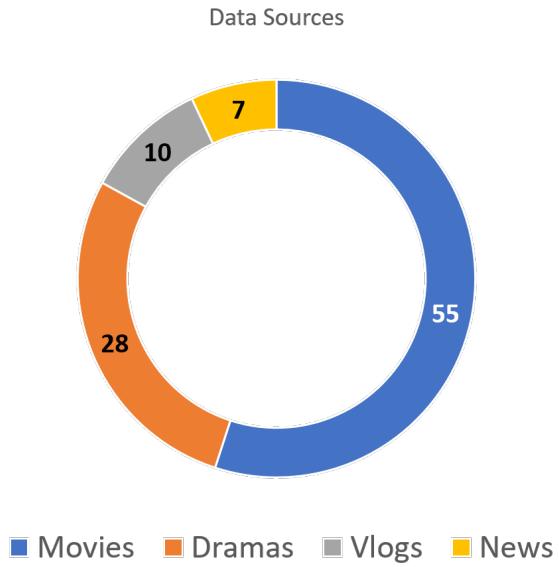


Figure 1. Distribution of data according to sources.

We developed a data annotation tool in Python using the PyQT framework. Figure 2 shows a snapshot of the tool. The main features of this tool are video loading, frame range selection, annotation, and saving. It also shows a total number of files collected in each class. Besides that, forwarding/rewinding by 2/5 seconds is also provided for catching a particular moment. The annotation tool is available at https://github.com/avishek-018/AVaTER.

 doi:10.20944/preprints202407.0917.v1

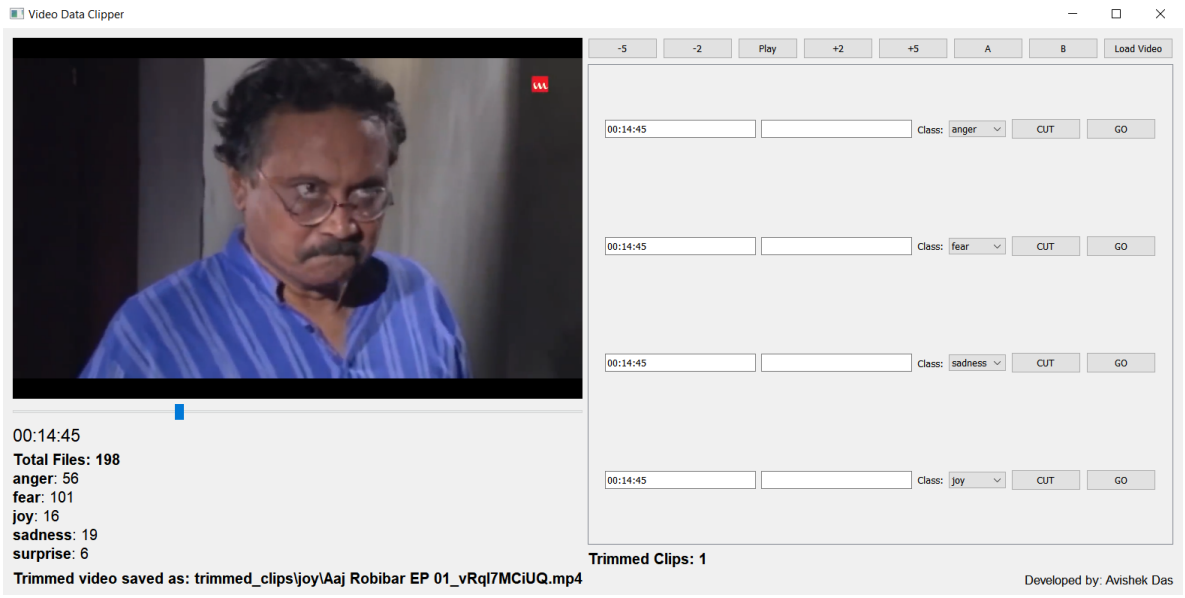


Figure 2. Snapshot of the data annotation tool.

3.2. Data Statistics

The dataset has been systematically divided into training, validation, and test sets to facilitate practical training and evaluation of emotion recognition models. The specific distribution of samples across these sets is detailed in Table 1. By developing MAViTE-Bangla, we address the lack of multimodal emotion recognition resources for Bengali, enhancing research and aiding in creating models that accurately interpret emotions from audio, video, and text.

Table 1. Data distribution among classes.

Class	Train	Validation	Test
Anger	221	48	48
Fear	70	15	15
Joy	191	41	41
Sadness	218	47	47
Total	700	151	151

Table 2 presents the detailed statistics for the audio dataset, focusing on four emotional classes: anger, fear, joy, and sadness. For anger, the minimum duration was 1.022 seconds, while the maximum duration reached 7.002 seconds, with a mean duration of 3.210 seconds. The cumulative duration for anger samples was recorded at 950.071 seconds. Fear exhibited a broader range of durations, with a minimum of 2.001 seconds and a maximum of 14.008 seconds. This resulted in a higher mean duration of 5.491 seconds, and the total duration for fear samples was 549.061 seconds. On the other hand, Joy had the shortest minimum duration among the emotions, at 0.952 seconds, and a maximum duration of 9.015 seconds. The mean duration for joy was 2.937 seconds, contributing to a total duration of 757.709 seconds. Moreover, Sadness had a minimum duration of 1.207 seconds and a maximum of 7.012 seconds, with a mean duration of 3.391 seconds. It also had the highest total duration among the emotional classes, amounting to 986.822 seconds.

 doi:10.20944/preprints202407.0917.v1

Table 2. Statistics for audio data.

Class	Min Duration (sec)	Max Duration (sec)	Mean Duration (sec)	Total Duration (sec)
Anger	1.022	7.002	3.210	950.071
Fear	2.001	14.008	5.491	549.061
Joy	0.952	9.015	2.937	757.709
Sadness	1.207	7.012	3.391	986.822

Table 3. Statistics for video data.

Class	Max Frame Rate (fps)	Min Frame Rate (fps)	Max Resolution	Min Resolution	Max File Size (KB)	Min File Size (KB)
Anger	30.0	24.00	1920x1080	208x210	10,075.15	141.32
Fear	30.0	23.98	2560x1440	450x360	6,845.49	99.54
Joy	30.0	23.98	1920x1080	206x174	9,975.87	102.99
Sadness	30.0	24.00	1920x1080	176x162	11,053.24	91.93

In the analysis of the video dataset, key metrics crucial for video classification tasks were examined and are presented in Table 3. The frame rate across all emotional classes reached a maximum of 30.0 frames per second (fps), a standard for high-quality video playback, ensuring smooth motion representation. The minimum frame rates observed were slightly lower, ranging from 23.98 to 24.00 fps, which still maintained acceptable video quality for classification tasks. The resolution analysis revealed considerable variation within the dataset. A resolution of 2560x1440 pixels, the highest observed, was found in the 'fear' class, indicative of high-definition video quality. High-definition videos up to 1920x1080 pixels were also in the 'anger,' 'joy,' and 'sadness' classes. However, due to resource constraints, some videos had lower resolutions: 208x210 pixels for 'anger,' 450x360 pixels for 'fear,' 206x174 pixels for 'joy,' and 176x162 pixels for 'sadness.' This lower-resolution data was collected due to resource constraints during the data acquisition process, highlighting the balance between resource availability and data quality. Despite these lower resolutions, crucial visual information necessary for effective classification is retained in the dataset.

Table 4. Statistics for text data.

Class	Total Words	Total Sentences	Average Word Length	Average Sentence Length	Lexical Diversity
Anger	2395	88	4.356	27.216	0.509
Fear	996	34	4.070	29.294	0.495
Joy	1819	54	4.318	33.685	0.537
Sadness	1971	31	4.337	63.581	0.464

The textual dataset for this study exhibits significant variations across the emotional classes, as detailed in Table 4. The Joy class stands out with the highest lexical diversity, scoring 0.537 and an average sentence length of 33.685 words. This indicates a varied vocabulary and longer sentences within this class. In contrast, the Sadness class features the most extended average sentence length at 63.581 words but has the lowest lexical diversity, with a score of 0.464. The Anger class is notable for its high total word count, comprising 2395 words, whereas the Fear class has the fewest words, totaling 996, and the least number of sentences, with only 34. These statistics highlight the diverse linguistic patterns within the dataset, which are crucial for developing emotion recognition models. The variations in sentence length and lexical diversity across the emotional classes provide a rich foundation for analyzing how different emotions are expressed in text. This diversity ensures that the

 doi:10.20944/preprints202407.0917.v1

models trained on this dataset can effectively learn to distinguish between the subtle nuances of each emotional class, thereby improving their accuracy and reliability in real-world applications.

4. Methodology

Emotions are expressed across multiple channels, including verbal, vocal, and visual modalities, rather than through a single channel. Utilizing various data modes enables a more precise analysis of emotional states, which may be overlooked by unimodal systems. This study combines features from video, audio, and text modalities to make the final prediction for multimodal emotion classification. Figure 3 illustrates the workflow of multimodal emotion classification using the MAViTE-Bangla dataset.

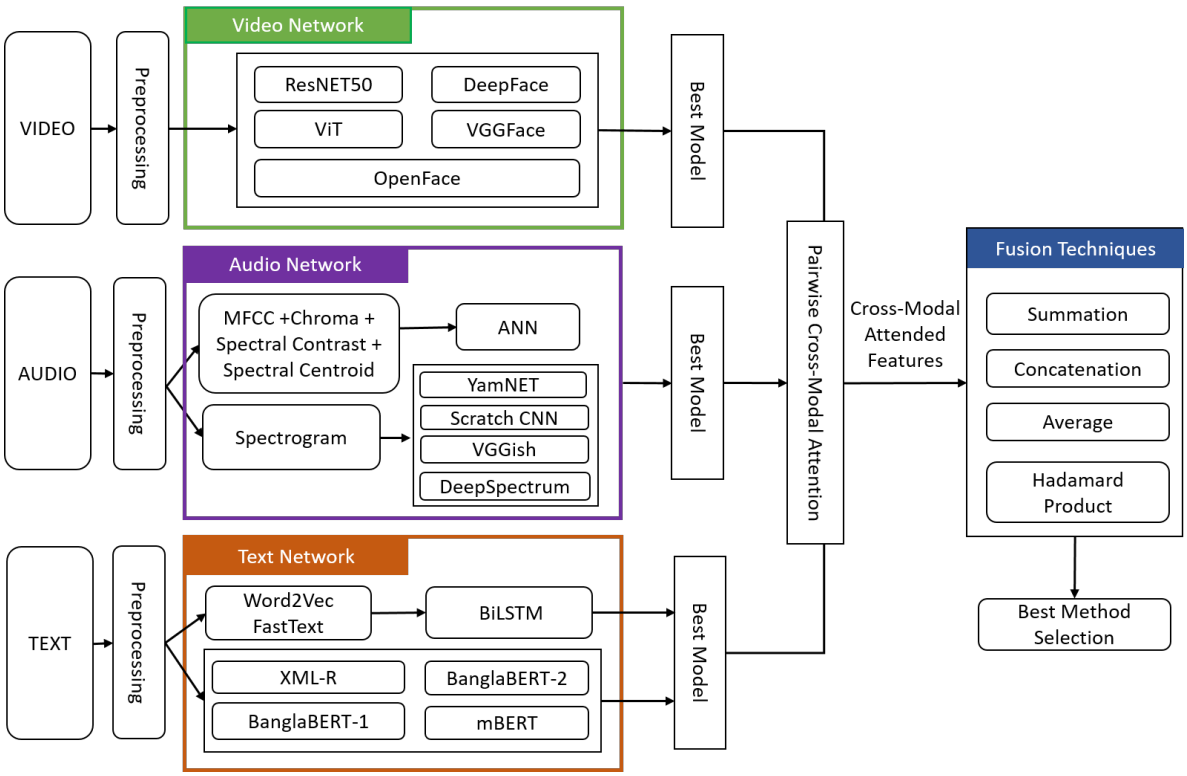


Figure 3. Abstract overview of the multimodal emotion classification system.

4.1. Video Network

The video network is introduced to extract spatial features from the selected frames of the video modality of mp4 input video. The function of this part of the system is depicted in Fig 3. The steps of video modality processing are represented in this section.

4.1.1. Preprocessing

For the video modality, a uniform approach is taken where five frames are extracted from each video, regardless of the video's total length. To achieve this, a skip window, denoted as W, is defined. This window determines the interval at which frames are sampled from the video. The calculation of W is based on the total number of frames N in the video and the desired number of frames L to be extracted for analysis. The formula used to compute the skip window is provided below:

$$W = max(\left\lfloor \frac{N}{L} \right\rfloor, 1) \tag{1}$$

 doi:10.20944/preprints202407.0917.v1

In this equation, the skip window W is determined as the maximum value between the floor division of the total number of frames N by the desired sequence length L and 1. Subsequently, all frames are resized to (224, 224) to maintain uniform dimensions and reduce computational complexity.

4.1.2. Feature Extraction

Feature extraction transforms raw pixel data from video frames into meaningful representations that capture essential patterns and attributes. Pretrained models such as ResNet50, ViT, DeepFace, FaceNet, and OpenFace have extracted the features from the selected video frames.

ResNet50: ResNet50 [23], short for Residual Networks with 50 layers, is a robust convolutional neural network (CNN) known for its deep architecture and ability to handle vanishing gradient problems through residual learning. It is widely used for feature extraction due to its strong performance in image recognition tasks. ResNet50 is pre-trained on the ImageNet dataset [24], a large-scale dataset containing over 14 million images and 1,000 object categories, allowing it to learn rich hierarchical features ranging from edges to complex patterns.

Vision Transformer (ViT): The Vision Transformer (ViT) [25] is a novel model that applies the transformer architecture, initially designed for natural language processing, to image data. ViT divides an image into patches, processes these patches as sequences, and uses self-attention mechanisms to capture long-range dependencies and contextual information. This approach allows ViT to excel at capturing detailed and intricate visual features. ViT is also pre-trained on the ImageNet dataset, leveraging its extensive and diverse image collection to learn effective visual content representations.

DeepFace: DeepFace [26] is a deep learning model designed explicitly for facial recognition and analysis. Developed by Facebook, it uses a combination of convolutional neural networks and deep learning techniques to achieve high accuracy in identifying and verifying faces. DeepFace is pretrained on a large proprietary dataset of over 4 million facial images from more than 4,000 persons. This extensive dataset enables DeepFace to extract detailed facial features that are critical for emotion recognition, such as expressions, landmarks, and subtle variations in appearance. This makes it particularly useful for applications where understanding facial cues is essential.

FaceNet: FaceNet [27] is a deep learning model developed by Google for face recognition and clustering. It learns to map face images to a compact Euclidean space where distances reflect face similarity. Trained on large datasets, including CASIA-WebFace and Google's proprietary datasets with millions of images, FaceNet extracts highly discriminative facial features. This makes face verification, identification, and emotion recognition practical by capturing subtle facial expressions and landmarks variations.

OpenFace: OpenFace [28] is an open-source facial behavior analysis model developed by researchers at Carnegie Mellon University. It employs deep learning techniques for facial landmark detection, head pose estimation, and facial action unit recognition. OpenFace is designed to be highly efficient and versatile, making it suitable for real-time applications. It is trained on a diverse set of facial images from various datasets, including the CMU Multi-PIE dataset and the 300-W dataset. This extensive training allows OpenFace to effectively capture and analyze facial features crucial for emotion recognition and human-computer interaction applications.

4.2. Audio Network

The function of an audio network is to process the audio modality. Here, a WAV file is passed through several steps so that intricate features can be extracted from it.

4.2.1. Preprocessing

For the audio modality, samples are adjusted to a uniform length of 5 seconds through padding and truncation. Normalization is then performed to achieve consistent volume levels, setting the audio to zero mean and unit variance. The noise reduction technique spectral gating is utilized to reduce background noise and enhance clarity.

 doi:10.20944/preprints202407.0917.v1

4.2.2. Feature Extraction

- **Hand-crafted Feature Extraction:**

Various acoustic features, including Mel-Frequency Cepstral Coefficients (MFCCs), Chroma Features, Spectral Centroid, Spectral Contrast, and Spectrograms, are extracted to analyze the audio signal comprehensively. Each feature offers unique insights into different aspects of the audio data.

MFCCs are set to 13 coefficients to capture key characteristics of the audio signal. This is achieved by applying a mel-scale filter bank to the Fourier transform of the signal. The MFCCs effectively capture the perceptual aspects of sound, making them particularly useful in speech and music analysis [29].

Chroma features represent the energy distribution across the 12 pitch classes of the musical octave. These features are valuable for analyzing the harmonic and melodic content of the audio, helping to identify chords and tonal structures.

The spectral centroid is a measure that indicates the "center of mass" of the audio spectrum. It is associated with the balance of frequencies and is often perceived as the brightness of the sound. Higher values of the spectral centroid correspond to brighter sounds.

Spectral contrast measures the difference between peaks and valleys in the spectrum. This feature captures the variations in the harmonic structure and timbre of the audio, providing insights into the textural differences within the sound.

Spectrograms display the spectrum of frequencies as they vary over time, offering a time-frequency representation of the audio signal. This allows for the observation of how different frequency components evolve, which is crucial for understanding the temporal dynamics of the audio.

To ensure uniformity, these features are truncated and padded to a length of 20, resulting in an 80-feature vector for each audio sample. This standardized vector format allows for consistent comparison and analysis across different audio samples, thereby enhancing the robustness of acoustic analysis in applications such as emotion recognition.

- **Deep Feature Extraction:**

Deep feature extraction involves using advanced deep learning techniques to analyze and enhance handcrafted features further. This process leverages the power of deep neural networks to capture intricate patterns and representations that handcrafted features alone might miss. Therefore, we have utilized a scratch ANN to extract the deep features of MFCC, Chroma Features, Spectral Centroid, and Spectral Contrast. On the other hand, for the spectrogram, we have experimented with several pre-trained deep-learning models:

CNN: A 5-layer Convolutional Neural Network (CNN) is designed to extract features from spectrograms. The network begins with an input layer that accepts spectrogram images of size 224x224 pixels rgb channel. It includes three convolutional layers, each followed by a max pooling layer to reduce the spatial dimensions while capturing essential features. The first convolutional layer uses 64 filters with a 3x3 kernel, the second and the third use 128 filters, all activated by the ReLU function. After these convolutional and pooling operations, the data is flattened into a vector. This vector is passed through two fully connected (dense) layers: the first dense layer with 256 neurons activated by ReLU, and the output layer with four neurons activated by softmax, designed for multi-class audio classification. The model is compiled using the Adam optimizer and categorical cross-entropy loss function.

YamNet: YamNet [30] is a pre-trained deep learning model developed by Google for audio event detection and classification. It is based on a MobileNetV1 architecture and is trained on the AudioSet dataset, which contains many audio events. YamNet takes audio waveforms as input and outputs class scores for 521 different audio event classes. It is designed to be lightweight and efficient, making it suitable for real-time audio analysis applications.

VGGish: VGGish is a Google pretrained deep learning model designed for extracting audio features. It is based on the VGG16 architecture, a convolutional neural network originally designed for

 doi:10.20944/preprints202407.0917.v1

image classification. VGGish is adapted for audio by processing log mel spectrograms of audio clips. The model is trained on a large-scale dataset, including millions of YouTube videos, and provides high-level audio features that can be used for various tasks such as audio classification, event detection, and similarity analysis.

DeepSpectrum: DeepSpectrum [31] is a framework that uses pre-trained deep learning models, initially designed for image analysis, to extract features from audio spectrograms. It converts audio signals into visual representations (spectrograms) and then applies image-based deep learning models such as ResNet, Inception, or VGG to these spectrograms. In this work, Inception-Resnet-v2 is chosen for extracting features from spectrograms. DeepSpectrum leverages the power of these image recognition models to capture intricate patterns in the audio data, making it useful for tasks like emotion recognition, sound classification, and other audio analysis applications.

4.3. Text Network

The input text data is utilized and encoded to extract deep features.

4.3.1. Preprocessing

In this study, the textual part of the MAViTE-Bangla dataset underwent comprehensive preprocessing to ensure data quality and consistency. The preprocessing steps included tokenization, normalization, and removing stop words and punctuation. Additionally, special characters and numbers were filtered out to enhance the robustness of the subsequent analysis.

4.3.2. Feature Extraction

- **Word Embedding**

We utilized the following word embedding approach followed by the BiLSTM network as a deep feature extractor.

Word2Vec: Word2Vec is a well-known and widely adopted word embedding technique used to identify semantic similarities between words within a dataset's context [32]. Two variants of the Word2Vec algorithms are skip-gram and a continuous bag of words (CBOW). According to [33], the skip-gram model performs effectively with small training datasets and accurately represents even rare words or phrases. In this study, Word2Vec is trained using the skip-gram model with a window size of 7, an embedding dimension of 100, and a minimum word count of 4.

FastText: The Word2Vec algorithm is limited in handling out-of-vocabulary words, as any word not included in the training set cannot be vectorized with a corresponding embedding. To overcome this limitation, the FastText algorithm was introduced [34]. By leveraging sub-word information, FastText utilizes character n-grams to establish semantic relationships between words within a specific context [35]. This methodology allows for synthesizing embeddings for words not present in the training vocabulary using their constituent n-grams. FastText, like Word2Vec, is available in both Skip-Gram and Continuous-BOW variants. In this research, the FastText algorithm was trained using the skip-gram model with a window size of 5, a character n-gram size of 5, and an embedding dimension of 100.

- **Contextual Embedding**

mBERT: We utilized the 'bert-base-multilingual-cased' model on the MAViTE-Bangla dataset, fine-tuning it by adjusting the batch size, learning rate, and number of epochs. The training of m-BERT [36] included the 104 most widely spoken languages, using the most extensive Wikipedia datasets available, which also covered Bengali. The pre-trained m-BERT model comprises approximately 110 million parameters.

XLM-R: XLM-R [37] was trained using a multilingual masked language model. Various innovative training methods improve BERT's performance, including (1) extending the training duration with more data, (2) using larger batch sizes and longer sequences, and (3) dynamically creating the

 doi:10.20944/preprints202407.0917.v1

masking pattern. The XLM-R model significantly outperforms other multilingual BERT models, particularly in low-resource languages. The 'xlm-Roberta-base' method is applied to the MAViTE-Bangla dataset with a batch size of 12.

Bangla BERT: This study employs two variants of Bangla Bert that are exclusively pre-trained in the Bengali language. The first variant, 'sagorsarker/Bangla-bert-base' (referred to as Bangla-BERT-1) [38], is trained on the Bengali corpus from OSCAR[1] and the Bengali Wikipedia Dump Dataset[2]. The second variant, 'csebuetnlp/banglabert' (Bangla-BERT-2) [39], is also used. Both pre-trained models are based on the masked language modeling approach described in the original BERT paper [40].

4.4. Pairwise Cross-modal Attention

The best models from the Video, Audio, and Text networks are chosen based on their weighted F1 score and prepared for cross-attention between each modality pair. The performance is measured based on the F1-score. Upon extracting features from the three modalities (audio, video, and text), the features of each modality undergo normalization to ensure uniform importance across all features. Following this, a cross-modal attention mechanism is implemented across three modality pairs: audio-video, audio-text, and video-text. In the case of the audio-video pair, this mechanism aids in synchronizing lip movements with speech, enhancing the recognition of spoken emotions. For the audio-text pair, integrating the tone of voice with textual content offers a more nuanced understanding of the sentiment conveyed in the speech. Furthermore, the video-text pair leverages facial expressions and body language to provide additional context to the spoken words, thereby facilitating the disambiguation of textual emotions.

4.4.1. Audio-Video Attention

The audio features are initially projected into a query space using a learned weight matrix. Concurrently, the video features are projected into key and value spaces utilizing distinct learned weight matrices. The critical space is employed for matching with queries, while the value space generates the final attended features. Attention scores are then calculated by taking the dot product of the audio queries and the video keys. Subsequently, these raw attention scores are passed through a softmax function, converting them into probabilities that sum to one. This normalization process aids in determining the relative importance of each video feature concerning each audio feature.

Finally, the normalized attention scores compute a weighted sum of the video values. Thus, an attended audio feature representation is created that emphasizes the most relevant video features. The overall cross-attention mechanism is represented by Eqs. 2-7:

$$Q_A = W_Q^A A \tag{2}$$

$$K_V = W_K^V V \tag{3}$$

$$V_V = W_V^V V \tag{4}$$

$$S_A V = Q_A K_V^T \tag{5}$$

$$\alpha_{ij} = \frac{exp(S_{AV_{ij}})}{\sum_k exp(S_{AV_{ik}})} \tag{6}$$

[1] https://oscar-corpus.com/
[2] https://dumps.wikimedia.org/bnwiki/latest/

 doi:10.20944/preprints202407.0917.v1

$$A' = \sum_j \alpha_{ij} V_V \tag{7}$$

In the above equations, A and V represent audio and video features, respectively. Q_A is the query projection of audio features, K_V is the key projection of video features, and V_V is the value projection of video features. W_Q^A, W_K^V, and W_V^V, are the learned weight matrices for the respective projections. Besides, S_{AV} represents the attention scores, and α_{ij} is the normalized attention score between the i-th audio query and the j-th video key. Also, A' is the attended audio feature representation based on the video features.

4.4.2. Audio-Text Attention

The same approach has been applied to the Audio-Text pair. The audio features are projected into a query space using a learned weight matrix. The text features are projected into key and value spaces using different learned weight matrices. Then, attention scores are calculated by taking the dot product of the audio queries with the text keys. After that, the attention scores are passed through a softmax function to obtain normalized probabilities. At last, the normalized attention scores are used to compute a weighted sum of the text values, resulting in an attended audio feature representation based on text features. Eqs. 8-13 illustrates the mechanism of audio-text cross-attention:

$$Q_A = W_Q^A A \tag{8}$$

$$K_T = W_K^T T \tag{9}$$

$$V_T = W_V^T T \tag{10}$$

$$S_A T = Q_A K_T^T \tag{11}$$

$$\beta_{ij} = \frac{exp(S_{AT_{ij}})}{\sum_k exp(S_{AV_{ik}})} \tag{12}$$

$$A'' = \sum_j \beta_{ij} V_T \tag{13}$$

In the above equations, A and T represent audio and text features, respectively. Q_A is the query projection of audio features, K_T is the key projection, and V_T is the value projection of text features. W_Q^A, W_K^T, and W_V^T, are the learned weight matrices for the respective projections. Besides, S_{AT} represents the attention scores, and βij is the normalized attention score between the i-th audio and j-th text features. Also, A'' is the attended audio feature representation based on the text features.

4.4.3. Video-Text Attention

The video features are projected into a query space using a learned weight matrix. The text features are projected into key and value spaces using different learned weight matrices. Attention scores are calculated by taking the dot product of the video queries with the text keys. The attention scores are passed through a softmax function. The normalized attention scores are used to compute a weighted sum of the text values, resulting in an attended video feature representation based on text features. The mechanism of this cross-attention can be observed in Eqs. 14-16.

$$S_V T = (W_V V)^T W_T^T \tag{14}$$

 doi:10.20944/preprints202407.0917.v1

$$\gamma_{ij} = \frac{exp(S_{VT_{ij}})}{\sum_k exp(S_{VT_{ik}})} \tag{15}$$

$$V' = \sum_j \gamma_{ij} V_T \tag{16}$$

Where W_V and W_T are the learned weight matrices for the respective projections. Besides, S_{VT} represents the attention scores, and βij is the normalized attention score between the i-th audio and j-th text features. Also, V' is the attended audio feature representation based on the text features.

4.5. Fusion Methods

After getting the crossmodal features from each pair, we combined the three pairs using different fusion methods. The best fusion method was selected based on their weighted F1 score, and we proceeded to further experimentation.

4.5.1. Summation Fusion

Summation Fusion combines features from multiple modalities by summing their respective values element-wise. Given feature vectors $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n$ from n modalities, the fused feature vector $\mathbf{f}_{\text{sum}}$ is computed as:

$$\mathbf{f}_{\text{sum}} = \mathbf{f}_1 + \mathbf{f}_2 + \cdots + \mathbf{f}_n$$

This method ensures that contributions from each modality are equally weighted, resulting in a single fused representation.

4.5.2. Concatenation

Concatenation merges features from different modalities by joining them end-to-end to form a single, extended feature vector. Given feature vectors $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n$ from n modalities, the fused feature vector $\mathbf{f}_{\text{concat}}$ is computed as:

$$\mathbf{f}_{\text{concat}} = [\mathbf{f}_1; \mathbf{f}_2; \cdots; \mathbf{f}_n]$$

where $[;]$ denotes the concatenation operation. This approach preserves the original dimensionality of each modality's features.

4.5.3. Average Fusion

Average Fusion combines features by calculating the mean value of the corresponding features across modalities. Given feature vectors $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n$ from n modalities, the fused feature vector $\mathbf{f}_{\text{avg}}$ is computed as:

$$\mathbf{f}_{\text{avg}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{f}_i$$

This method helps balance the contributions from each modality, providing a more stable and generalized representation.

4.5.4. Hadamard Product Fusion

Hadamard Product Fusion integrates features by performing element-wise multiplication of the corresponding features from different modalities. Given feature vectors $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n$ from n modalities, the fused feature vector $\mathbf{f}_{\text{had}}$ is computed as:

$$\mathbf{f}_{\text{had}} = \mathbf{f}_1 \odot \mathbf{f}_2 \odot \cdots \odot \mathbf{f}_n$$

where $\odot$ denotes the element-wise multiplication. This approach emphasizes the interactions and commonalities among the features of different modalities.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 July 2024 doi:10.20944/preprints202407.0917.v1

4.6. Proposed Method

The proposed architecture is illustrated in Figure 4. We employed two distinct forms of cross-attention mechanisms to facilitate the interaction between features from each modality. The first form, which is already mentioned, includes Audio-Video(Att_AV), Audio-Text(Att_AT), and Video-Text attention(Att_VT), and the second form involves Video-Audio(Att_VA), Text-Audio(Att_TA), and Text-Video attention(Att_TV). In the second form, video features act as queries to attend to audio features in the Video-Audio attention mechanism. In contrast, text features serve as queries for audio and video features in the Text-Audio and Text-Video attention mechanisms.

The proposed method combines these two cross-attention mechanisms to enhance the interaction between modalities comprehensively. Doing so ensures that each modality effectively informs and enriches the others, capturing a more holistic view of the multimodal data. Concatenation emerged as the most effective method for the fusion of these cross-attended features. This approach combines the enriched features from each modality into a single, comprehensive feature vector, facilitating robust and more accurate multimodal emotion recognition.

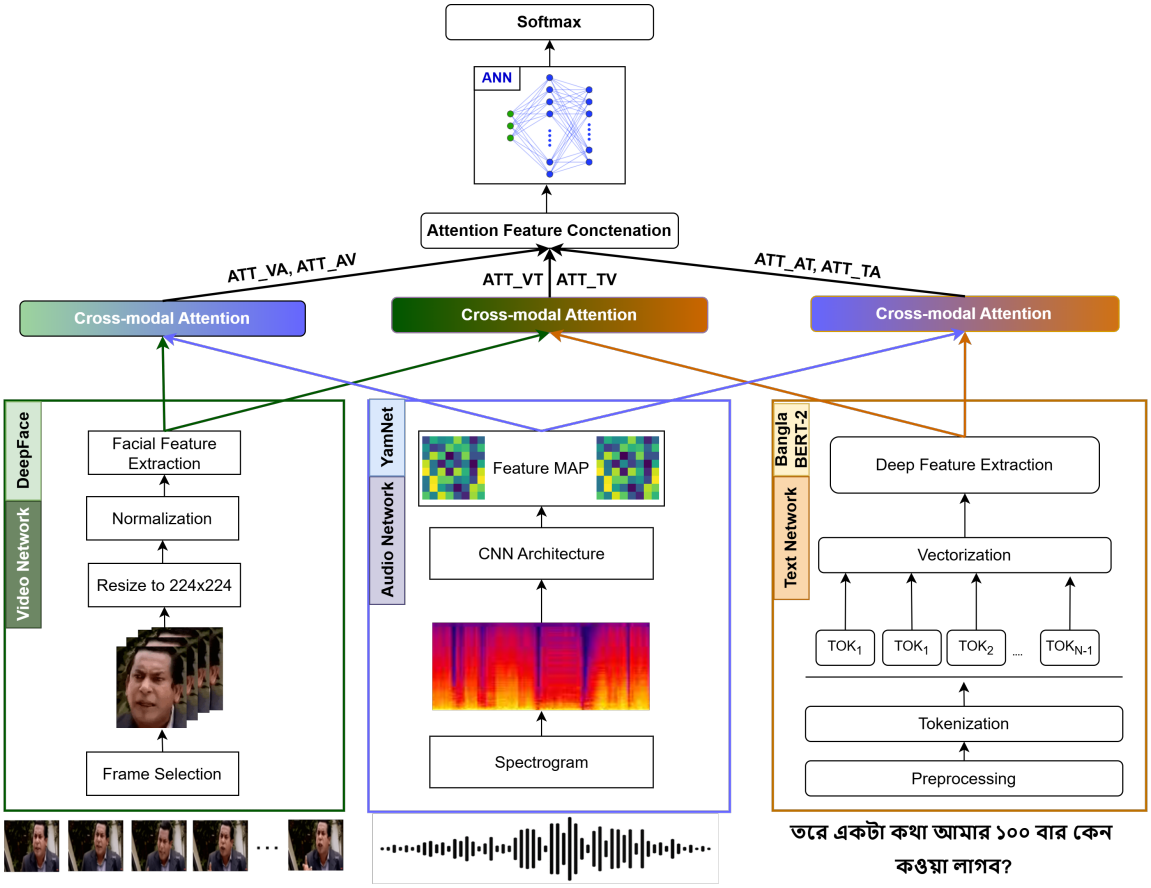


Figure 4. Proposed MEC architecture using cross-modal attention.

5. Experiments

This section provides a concise summary of the experimental results. The metrics evaluated include Precision (Pr.), Recall (Re.), F1 Score (F1.), and Accuracy (Acc.), both for single modalities and multimodality. The model with the highest F1 score was selected as the best-performing model. Table 5 presents the unimodal results for various approaches across the audio, video, and text modalities.

5.1. Unimodal

Table 5 shows the performance of unimodal approaches on audio, video, and text modalities.

 doi:10.20944/preprints202407.0917.v1

5.1.1. Audio Modality

We assessed five methods for the audio modality. A1 uses statistical features: MFCC+Chroma+Spectral Contrast+Spectral Centroid using an Artificial Neural Network (ANN) and achieved an F1 Score of 0.59. The spectrogram is utilized in A2, A3, A4, and A5. Among them, A3 and YamNet performed better, with a 0.62 F1 score. Scratch CNN performed lowest as it is not a pre-trained model.

5.1.2. Video Modality

The video modality was evaluated using inputs of five and ten frames across five models: Resnet50, DeepFace, Vision Transformer (ViT), FaceNet, and OpenFace. The DeepFace model consistently outperformed the others. For the five-frame input (Method V2), DeepFace achieved the highest F1-score of 0.71. Similarly, the ten-frame input (Method V7) maintained a high F1-score of 0.70. In contrast, the Resnet50 and ViT models exhibited moderate performance, with Resnet50 achieving F1-scores of 0.65 (Method V1) for five frames and 0.62 (Method V6) for ten frames, and ViT attaining F1-scores of 0.65 (Method V3) for five frames and 0.64 (Method V8) for ten frames. On the other hand, FaceNet and OpenFace showed comparative performance. Interestingly, DeepFace, with five frames, outperformed its ten-frame counterpart, indicating that five frames are sufficient for more effective emotion recognition.

5.1.3. Text Modality

The text modality evaluated methods based on Word2Vec, FastText, and Transformer-based models. Bangla BERT-2 (Method T6) achieved the highest F1 Score of 0.76, demonstrating its strong ability to capture emotional nuances in text data. XML-R (Method T4) also showed commendable performance with an F1-Score of 0.70. In contrast, BiLSTM models utilizing Word2Vec (Method T1) and FastText (Method T2) displayed lower performance, with F1-scores of 0.51 and 0.52, respectively.

 doi:10.20944/preprints202407.0917.v1

Table 5. Unimodal results for Audio, Video, and Text modalities.

Modality	Approach/ Features	Model	Method No.	Pr	Re	F1	Acc
Audio	MFCC + Chroma+ Spectral Contrast + Spectral Centroid	ANN	A1	0.59	0.59	0.59	0.59
	Spectrogram	CNN	A2	0.57	0.55	0.55	0.56
		YamNet	**A3**	**0.63**	**0.62**	**0.62**	**0.62**
		Vggish	A4	0.59	0.59	0.60	0.59
		DeepSpectrum	A5	0.61	0.60	0.59	0.61
Video	5 Frames	Resnet50	V1	0.65	0.66	0.65	0.65
		DeepFace	**V2**	**0.74**	**0.72**	**0.71**	**0.72**
		Vit	V3	0.67	0.66	0.65	0.67
		FaceNet	V4	0.69	0.67	0.66	0.67
		OpenFace	V5	0.72	0.69	0.67	0.69
	10 Frames	Resnet50	V6	0.64	0.63	0.62	0.63
		DeepFace	V7	0.73	0.71	0.70	0.71
		Vit	V8	0.66	0.65	0.64	0.66
		FaceNet	V9	0.68	0.67	0.65	0.66
		OpenFace	V10	0.71	0.69	0.70	0.70
Text	Word2Vec	BiLSTM	T1	0.52	0.50	0.50	0.52
	FastText	BiLSTM	T2	0.53	0.52	0.52	0.52
	Transformer	MBERT	T3	0.53	0.52	0.53	0.52
		XML-R	T4	0.68	0.67	0.65	0.66
		Bangla BERT-1	T5	0.76	0.76	0.76	0.76
		Bangla BERT-2	**T6**	**0.77**	**0.76**	**0.76**	**0.76**

5.2. Multimodal

The multimodal approach for emotion recognition leverages the best-performing model from each modality: A3 for audio, V2 for video, and T6 for textual feature extraction. We implemented a pair-wise cross-modal attention mechanism encompassing six cross-attention feature pairs, i.e. Audio-Video (Att_AV), Audio-Text (Att_AT), Video-Text (Att_VT), Video-Audio (Att_VA), Text-Audio (Att_TA), and Text-Video (Att_TV). Various fusion techniques were explored to combine the cross-modal attended features, and the performance comparison of these strategies is illustrated in Table 6. It is evident from the table that the concatenation method outperforms the others, achieving an F1 score of 0.64.

This superior performance is attributed to the method's ability to preserve each modality's full diversity of features by merging them into a single, comprehensive feature vector. This approach retains the unique characteristics of each modality, which is essential for multimodal data, as different modalities can provide different information.

Table 6. Fusion techniques of cross-modal attended features.

Fusion Techniques of Cross-Modal Attended Features	Pr	Re	F1	Acc
Summation	0.60	0.59	0.59	0.60
Averaging	0.56	0.57	0.55	0.56
Concatenation	**0.65**	**0.64**	**0.64**	**0.64**
Hadamard Product	0.62	0.61	0.62	0.63

Table 7 illustrates the class-wise performance metrics of the proposed multimodal approach. The Joy class achieved the highest F1-score of 0.71 and recall of 0.73, demonstrating a solid capability to

 doi:10.20944/preprints202407.0917.v1

recognize joyful expressions. The Anger class exhibited the highest precision at 0.70, with a balanced performance and an F1-score of 0.69. Although the Fear class had a precision of 0.42, it showed a recall of 0.53, indicating that while it missed some instances, it accurately identified fear in many cases. Conversely, the Sadness class had relatively lower performance, with an F1-score of 0.58, underscoring the challenges in accurately detecting sadness.

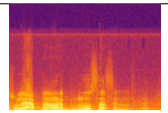
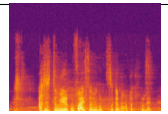
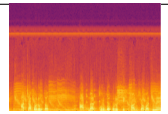
Table 7. Class-wise performance of the proposed approach.

Class	Pr.	Re.	F1.
Anger	0.70	0.69	0.69
Sadness	0.63	0.53	0.58
Joy	0.70	0.73	0.71
Fear	0.42	0.53	0.47

Table 8 shows some test samples and the unimodal and multimodal prediction results comparisons. The video frames and texts can be understood from the table, though the audio content can't be fully understood. We can see that in Example 1, the Video modality is predicted as Joy and Audio and Text are predicted as Sadness, which also matches the multimodal prediction and original label. We can see a man smiling at the frames, which drives the video model to predict it as Joy. But if we see the text, we can realize it represents some pitiful situation, and the tone represents the same. In Example 2, The text is predicted as Anger, whereas Audio and Video are Joy. The content of the text suggests an annoying situation, but the tone includes laughs, and so do the video frames. In Example 3, something different happened: primary modalities, Audio and Video, predict Joy, which is not the original label. The original label matched the Text model's prediction. As all of the above examples are labeled with the overall understanding of the content, the crossmodal features were mapped so that the multimodal prediction can extract the correct emotion level.

 doi:10.20944/preprints202407.0917.v1

Table 8. Example with comparison of unimodal and multimodal approaches.

No.	Modality	Content	Unimodal Prediction	Multimodal Prediction	Original Label
1	Video		Joy	Sadness	Sadness
	Audio		Sadness		
	Text	আসলে আপনি আমাকে ভুল বুঝতেছেন। (Actually you are getting me wrong.)	Sadness		
2	Video		Joy	Joy	Joy
	Audio		Joy		
	Text	আচ্ছা তুমি এরকম ফাজলামি করতেছ কেন হঠাৎ করে! (Why are you doing such nonsense all of a sudden!)	Anger		
3	Video		Joy	Anger	Anger
	Audio		Joy		
	Text	আচ্ছা বলুন তো আপনার পূর্বপুরুষেরা কি খানসামা ছিলেন?(So tell me, were your ancestors housekeepers?)	Anger		

 doi:10.20944/preprints202407.0917.v1

5.3. Ablation Study

To show the effectiveness of our method, we reduced the cross-attention pairs. We made 2 groups: 1) Audio-Video (Att_AV), Audio-Text (Att_AT), and Video-Text (Att_VT) cross attention pairs and 2) another is the reverse, i.e., Video-Audio (Att_VA), Text-Audio (Att_TA), and Text-Video (Att_TV) cross attention pairs. The result is presented in Table 9. We can see that using six pairs of cross-attended features enhances the f1 score by 0.2. This increases the interaction between modalities leading to improved emotion recognition accuracy. This ablation study represents the importance of inter-pair dependencies. For instance, Video-Audio cross-attended features can not alone improve the system, but adding Audio-Video cross-attended features can improve the overall performance.

Table 9. Cross-modal attention strategies and their corresponding performance metrics.

Cross Modal Attention Strategies	Pr.	Re.	F1.	Acc.
Att_AV + Att_VT + Att_AT	0.63	0.63	0.62	0.63
Att_VA + Att_TV + Att_TA	0.62	0.61	0.62	0.62
Att_AV + Att_VT + Att_AT + Att_VA + Att_TV + Att_TA	**0.65**	**0.64**	**0.64**	**0.64**

5.4. Comparison with Existing Work

There are relatively very few works done on 3 modalities of emotion. Table 10 compares the proposed model and two existing multimodal emotion recognition models. Notably, the proposed model achieves a superior F1 score. The model in [16] records the lowest F1 score of 0.57, which can be attributed to the excessive reduction of handcrafted audio features, resulting in the loss of vital information essential for precise emotion recognition.

Table 10. Performance Comparison of Proposed Model with Existing Approaches.

Multimodal Feature Extraction	Classifier	F1
Wav2vec2.0 + videoMAE + BERT [17]	SVM	0.59
(Handcrafted audio features + CNN-LSTM) + Inception-ResNet-v2 + Word2Vec [16]	ANN	0.57
Yamnet + DeepFace + Bangla BERT-2(Proposed)	ANN	0.64

5.5. Error Analysis

The confusion matrix depicted in Figure 5 reveals a significant misclassification pattern between the emotions of Sadness and Anger. Specifically, 6 instances of Sadness were incorrectly classified as Anger. This misclassification can be attributed to the acoustic similarities between these two emotions. Both high-pitched shouting, characteristic of Anger, and the crying associated with Sadness share similar tonal qualities, leading to confusion for the model.

 doi:10.20944/preprints202407.0917.v1

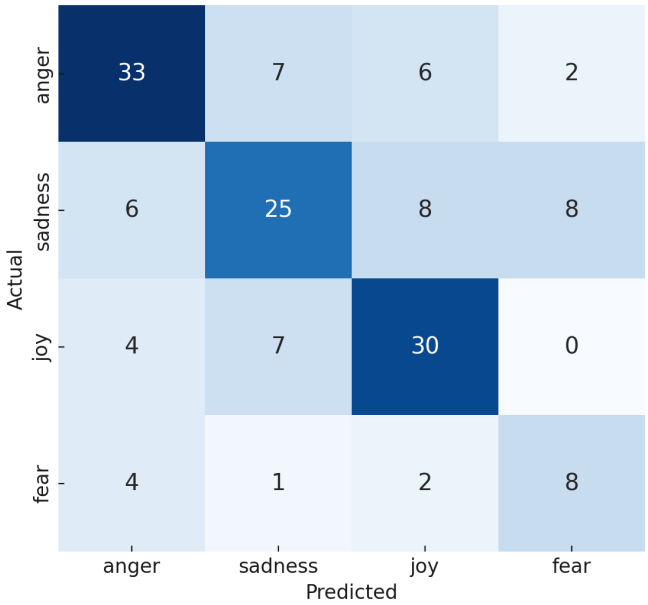


Figure 5. Confusion matrix of the test set prediction.

Additionally, 8 instances of Sadness were misclassified as Fear. This misclassification might stem from the overlapping vocal expressions and physiological responses between Sadness and Fear, such as quivering voices and hesitations, making it challenging for the model to differentiate between these emotions accurately.

Furthermore, the Fear class exhibits a high rate of misclassification, with four samples erroneously identified as Anger, another two as Joy, and one as Sadness. This issue is primarily due to the relatively small size of the Fear class within the dataset, which hinders the model's capacity to learn distinguishing features effectively. The insufficient data for Fear limits the model's exposure to diverse expressions of this emotion, impairing its performance. Enhancing the dataset size for underrepresented classes and improving the model's capability to discern subtle emotional expressions can significantly boost the overall performance of the emotion recognition system.

6. Conclusion

This study introduced a crossmodal fusion framework to classify multimodal emotions into four categories: Anger, Fear, Joy, and Sadness. The framework addresses the complexities of emotion recognition across different modalities by integrating Bengali acoustic, visual, and textual features. To facilitate the evaluation of the proposed framework, a novel Bengali multimodal dataset named MAViTE-Bangla has been created, comprising 1002 samples. Experimental outcomes demonstrated that the combination of different feature extraction techniques, such as Yamnet (acoustic), DeepFace (visual), and Bangla BERT-2 (textual) with ANN classifier outperformed other exploited techniques, achieving the highest F1 score (0.64)for MEC. Despite the current multimodal approach yielding a lower F1 score than unimodal results, the advantages of multimodality, such as providing a more comprehensive, robust, and context-aware understanding of emotions, underscores its value. These benefits justify the ongoing use and further development of multimodal systems. By emphasizing the broader perspective and the potential for future enhancements, we can bolster the adoption of multimodal approaches in emotion recognition tasks, paving the way for more accurate and nuanced emotion detection systems.

Author Contributions: Conceptualization, A.D., M.S.S. and M.M.H.; Methodology, A.D., M.S.S. and M.M.H.; Software, A.D. and M.S.S.; Data curation, A.D. and M.S.S.; Writing—original draft, A.D. and M.S.S.; Writing—review & editing, M.M.H., N.S., M.A.A.D; Supervision, M.M.H. All authors have read and agreed to the published version of the manuscript.

 doi:10.20944/preprints202407.0917.v1

Funding: This research received no external funding.
Institutional Review Board Statement: Not applicable
Informed Consent Statement: Not applicable
Data Availability Statement: Data are contained within the article. **Conflicts of Interest:** The authors declare no conflict of interest.
Acknowledgement: This work is supported by the Directorate of Research and Extension (DRE), Chittagong University of Engineering & Technology (CUET).

References

1. Beard, R.; Das, R.; Ng, R.W.; Gopalakrishnan, P.K.; Eerens, L.; Swietojanski, P.; Miksik, O. Multi-modal sequence fusion via recursive attention for emotion recognition. Proceedings of the 22nd conference on computational natural language learning, 2018, pp. 251–259.
2. Haque, R.; Islam, N.; Tasneem, M.; Das, A.K. Multi-class sentiment classification on Bengali social media comments using machine learning. *International journal of cognitive computing in engineering* **2023**, *4*, 21–35.
3. Islam, K.I.; Yuvraz, T.; Islam, M.S.; Hassan, E. Emonoba: A dataset for analyzing fine-grained emotions on noisy bangla texts. Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2022, pp. 128–134.
4. Kabir, A.; Roy, A.; Taheri, Z. BEmoLexBERT: A Hybrid Model for Multilabel Textual Emotion Classification in Bangla by Combining Transformers with Lexicon Features. Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), 2023, pp. 56–61.
5. Das, A.; Sharif, O.; Hoque, M.M.; Sarker, I.H. Emotion classification in a resource constrained language using transformer-based approach. *arXiv preprint arXiv:2104.08613* **2021**.
6. Iqbal, M.A.; Das, A.; Sharif, O.; Hoque, M.M.; Sarker, I.H. Bemoc: A corpus for identifying emotion in bengali texts. *SN Computer Science* **2022**, *3*, 135.
7. Rahman, M.; Talukder, M.R.A.; Setu, L.A.; Das, A.K. A dynamic strategy for classifying sentiment from Bengali text by utilizing Word2vector model. *Journal of Information Technology Research (JITR)* **2022**, *15*, 1–17.
8. Mia, M.; Das, P.; Habib, A. Verse-Based Emotion Analysis of Bengali Music from Lyrics Using Machine Learning and Neural Network Classifiers. *International Journal of Computing and Digital Systems* **2024**, *15*, 359–370.
9. Parvin, T.; Sharif, O.; Hoque, M.M. Multi-class textual emotion categorization using ensemble of convolutional and recurrent neural network. *SN Computer Science* **2022**, *3*, 62.
10. Sultana, S.; Iqbal, M.Z.; Selim, M.R.; Rashid, M.M.; Rahman, M.S. Bangla speech emotion recognition and cross-lingual study using deep CNN and BLSTM networks. *IEEE Access* **2021**, *10*, 564–578.
11. Dhar, P.; Guha, S. A system to predict emotion from Bengali speech. *Int. J. Math. Sci. Comput* **2021**, *7*, 26–35.
12. Nahin, A.S.M.; Roza, I.I.; Nishat, T.T.; Sumya, A.; Bhuiyan, H.; Hoque, M.M. Bengali Hateful Memes Detection: A Comprehensive Dataset and Deep Learning Approach. 2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems (iCACCESS). IEEE, 2024, pp. 01–06.
13. Ghosh, S.; Ramaneswaran, S.; Tyagi, U.; Srivastava, H.; Lepcha, S.; Sakshi, S.; Manocha, D. M-MELD: A Multilingual Multi-Party Dataset for Emotion Recognition in Conversations. *arXiv preprint arXiv:2203.16799* **2022**.
14. Hu, G.; Lin, T.E.; Zhao, Y.; Lu, G.; Wu, Y.; Li, Y. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv preprint arXiv:2211.11256* **2022**.
15. Zhao, J.; Dong, W.; Shi, L.; Qiang, W.; Kuang, Z.; Xu, D.; An, T. Multimodal Feature Fusion Method for Unbalanced Sample Data in Social Network Public Opinion. *Sensors* **2022**, *22*. doi:10.3390/s22155528.
16. Hosseini, S.S.; Yamaghani, M.R.; Poorzaker Arabani, S. Multimodal modelling of human emotion using sound, image and text fusion. *Signal, Image and Video Processing* **2024**, *18*, 71–79.
17. Shayaninasab, M.; Babaali, B. Multi-Modal Emotion Recognition by Text, Speech and Video Using Pretrained Transformers. *arXiv preprint arXiv:2402.07327* **2024**.
18. Mamieva, D.; Abdusalomov, A.B.; Kutlimuratov, A.; Muminov, B.; Whangbo, T.K. Multimodal Emotion Detection via Attention-Based Fusion of Extracted Facial and Speech Features. *Sensors* **2023**, *23*. doi:10.3390/s23125475.

19. Zhang, Z.; Zhang, S.; Ni, D.; Wei, Z.; Yang, K.; Jin, S.; Huang, G.; Liang, Z.; Zhang, L.; Li, L.; Ding, H.; Zhang, Z.; Wang, J. Multimodal Sensing for Depression Risk Detection: Integrating Audio, Video, and Text Data. *Sensors* **2024**, *24*. doi:10.3390/s24123714.
20. Taheri, Z.S.; Roy, A.C.; Kabir, A. BEmoFusionNet: A Deep Learning Approach For Multimodal Emotion Classification in Bangla Social Media Posts. 2023 26th International Conference on Computer and Information Technology (ICCIT). IEEE, 2023, pp. 1–6.
21. Hossain, E.; Sharif, O.; Hoque, M.M. Mute: A multimodal dataset for detecting hateful memes. Proceedings of the 2nd conference of the asia-pacific chapter of the association for computational linguistics and the 12th international joint conference on natural language processing: student research workshop, 2022, pp. 32–39.
22. Ahsan, S.; Hossain, E.; Sharif, O.; Das, A.; Hoque, M.M.; Dewan, M. A Multimodal Framework to Detect Target Aware Aggression in Memes. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 2487–2500.
23. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
24. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A.C.; Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* **2015**, *115*, 211–252. doi:10.1007/s11263-015-0816-y.
25. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; others. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
26. Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1701–1708. doi:10.1109/CVPR.2014.220.
27. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A unified embedding for face recognition and clustering. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015. doi:10.1109/cvpr.2015.7298682.
28. Baltrušaitis, T.; Robinson, P.; Morency, L.P. OpenFace: An open source facial behavior analysis toolkit. 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), 2016, pp. 1–10. doi:10.1109/WACV.2016.7477553.
29. Sen Sarma, M.; Das, A. BMGC: A deep learning approach to classify Bengali music genres. Proceedings of the 4th International Conference on Networking, Information Systems & Security, 2021, pp. 1–6.
30. Google. YamNet: Pretrained model for audio event detection, 2020. Accessed: 2024-06-22.
31. Amiriparian, S.; Gerczuk, M.; Ottl, S.; Cummins, N.; Freitag, M.; Pugachevskiy, S.; Baird, A.; Schuller, B. Snore Sound Classification Using Image-Based Deep Spectrum Features. Interspeech 2017. ISCA, 2017, pp. 3512–3516.
32. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* **2013**.
33. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2; Curran Associates Inc.: Red Hook, NY, USA, 2013; NIPS'13, p. 3111–3119.
34. Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jégou, H.; Mikolov, T. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651* **2016**.
35. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* **2017**, *5*, 135–146.
36. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019, [arXiv:cs.CL/1810.04805].
37. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* **2019**.
38. Sarker, S. BanglaBERT: Bengali Mask Language Model for Bengali Language Understading, 2020.
39. Bhattacharjee, A.; Hasan, T.; Samin, K.; Islam, M.S.; Rahman, M.S.; Iqbal, A.; Shahriyar, R. BanglaBERT: Combating Embedding Barrier in Multilingual Models for Low-Resource Language Understanding. *CoRR* **2021**, *abs/2101.00204*, [2101.00204].
40. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **2018**.

 doi:10.20944/preprints202407.0917.v1