Enhanced Trimodal Emotion Recognition Using Multibranch Fusion Attention with Epistemic Neural Networks and Fire Hawk Optimization

Bangar Raju Cherukuri
Senior Web Developer, Germantown MD, USA.
rajucherukuri5@gmail.com

Correspondence should be addressed to Bangar Raju Cherukuri: rajucherukuri5@gmail.com

Article Info
Journal of Machine and Computing (https://anapub.co.ke/journals/jmc/jmc.html)
Doi: https://doi.org/10.53759/7669/jmc202505005
Received 20 March 2024; Revised from 28 July 2024; Accepted 22 September 2024.
Available online 05 January 2025.

Abstract – Emotions are very crucial for humans as they determine our ways of thinking, our actions, and even how we interrelate with other persons. Recognition of emotions plays a critical role in areas such as interaction between humans and computers, mental disorder detection, and social robotics. Nevertheless, the current emotion recognition systems have issues like noise interference, inadequate feature extraction, and integration of data for the multimodal context that embraces audio, video, and text. To address these issues, this research proposes an "Enhanced Trimodal Emotion Recognition Using Multibranch Fusion Attention with Epistemic Neural Networks and Fire Hawk Optimization." The proposed method begins with modality-specific preprocessing: Natural Language Processing (NLP) for text to address linguistic variations, Relaxed instance Frequency-wise Normalization (RFN) for the audio to minimize distortion of noise's importance and iterative self-Guided Image Filter (isGIF) for the videos to enhance the image quality and minimize the artifacts. This preprocessing facilitates and optimizes data for feature extracting; an Inception Transformer for capturing the textual contexts; Differentiable Adaptive Short-Time Fourier transform (DA-STFT) to extract the audio's spectral and temporal features; and class attention mechanisms to emphasize important features in the videos. Following that, these features are combined through a Multi-Branch Fusion Attention Network to harmonize all the multifarious modalities into one. The last sanity check occurs through an Epistemic Neural Network (ENN), which tackles issues of uncertainty involved in the last classification, and the Fire Hawk algorithm is used to enhance the emotion recognition capabilities of the framework. Finally the proposed approach attains 99.5% accuracy with low computational time. Thus, the proposed method addresses important shortcomings of the systems developed previously and can be regarded as a contribution to the development of the multimodal emotion recognition field.

Keywords – Trimodel Emotion Recognition, Audio, Video, Text, NLP, Epistemic Multibranch Fusion Attention Neural Network, Fire Hawk Optimizer.

I. INTRODUCTION

Among the most important life skills is the ability to understand other people's feelings. As soon as others detect that someone is feeling something, it affects the way they react and behave [1]. Human emotions are basic characteristics that are crucial to social communication. Humans custom a variety of expressions to convey their emotions, including body language, word choice, tone of voice, and facial emotions [2-4]. Correctly recognizing the feelings of various individuals is essential to effective communication. If someone is furious, one should approach them carefully.

Emotion recognition and understanding in sensor data is critical for many industries and enterprises, including Artificial Intelligence (AI), robots, gaming, Human-Computer Interaction (HCI), entertainment, and surveillance [5-8]. Focus on identifying emotions as they are supplied in the assignment rather than attempting to recognize them in real life. Gathering information about people's emotions is essential for building effective AI systems that are capable of identifying them [9-10].

Textual information generated during emotional communication is frequently used to infer people's emotional states because it is the most fundamental and direct carrier. Additionally, with the quick growth of social media and smart terminals in recent decades, online social networks like Facebook, Weibo, Line, and Twitter have become an unparalleled worldwide phenomenon. Individuals are used to exchanging ideas and communicating with one another on these platforms

[11-14]. There is an increasing need for fine-grained emotion detection due to the development of large amounts of emotionally rich public and personal data, which has driven research in mental health and emotional management.

Natural language processing and speech have experienced a sharp rise in interest in audio-based emotion recognition in recent years. The challenge of automatically identifying the various human emotional conditions conveyed in natural speech, such as happiness, sadness, rage, and neutrality, is referred to as emotion recognition [15-16]. In numerous domains, including customer analysis and service call review, human-machine interaction, mental health surveillance, etc., it has been an essential subtask in the development of intelligent systems. The fact that the interaction between language and audio frequently modifies the expressed emotional states is a significant barrier in speech emotion detection [17].

Researchers' continued interest in emotion identification technology is keeping it developing at a rapid pace. In order to better adapt to real-world application scenarios like car-hailing services, education support, mental health care, and smart homes, a number of recent research are dissatisfied with just identifying a sentence, single phrase, or article using traditional emotion classification methods. [18-19]. The analysis of conversations, particularly those in videos, is becoming more popular.

The public can now access data created by emotion recognition researchers that identify different emotions in both text and images. However, in contrast to datasets limited to one or two modalities, a notable deficiency exists in datasets that simultaneously incorporate all three modalities: text, video, and audio. Such extensive, trimodel datasets combining text, audio, and video are difficult and expensive to create [20]. The main contributions of this research are;

- Comprehensive Multimodal Preprocessing: It uses NLP for text, RFN for audio, and isGIF for video, along with noise reduction, dimension reduction, and normalization.
- Innovative Feature Extraction Techniques: For text input, it applies Inception Transformer for the audio (DA-STFT) and video (class attention mechanisms); it improves the details of the features and their definition for all the inputs.
- The proposed method integrates the Multi-Branch Fusion Attention Network with an Epistemic Neural Network (ENN), which makes a novel network called the Epistemic Multi-Branch Fusion Attention Neural Network, or EMFANN, for enhanced trimodel emotion recognition.
- The Fire Hawk Optimization methodology enhances the classification process and increases the speed and accuracy of the models dealing with multifaceted multimodal combinations.
- The proposed method aims to overcome the gaps in current systems by significantly improving the assessments made about the accuracy of recognizing human emotion in textual, audio, and video sources.

The manuscript is organized as follows: The sections are as follows: Section 1 contains the introduction; Section 2 focuses on the literature review; Section 3 describes the methodologies that are proposed; Section 4 is dedicated to the results and discussions; and Section 5 contains the conclusion of the manuscript.

II. LITERATURE SURVEY

The trimodel emotion recognition system has been the focus of numerous researches. This section discusses a few of them.

Peña D et al. (2023) [21] have suggested a structure for assessing fusion techniques for multimodal emotion identification. The IEMOCAP dataset provided the source of the input images. Equivalent conditions are provided by the suggested approach to enable a fair evaluation of fusion techniques. Based on this study, the suggested approach assesses several fusion techniques for multimodal emotion identification. Based on the chosen architecture and dataset, the study finds that self-attention and weighted techniques work best. Multilayer Perceptron and Self-Attention models are the most effective since they require the fewest operations. The imbalance in the dataset caused the suggested approach to perform poorly.

In 2023 Chen S et al. [22] have suggested a dynamical fusion network with multiple stages for multimodal emotion recognition. This study used the multimodal emotion dataset DEAP. With the multi-stage heterogeneous dynamical fusion network, the combined representation utilizing cross-modal correlations is obtained. After investigating latent interactions between variables from various modalities, the study creates a fusion network with several stages to take advantage of fine-grained inter-correlations that are unimodal, bimodal, and trimodal. The model's processing requirements may rise as a result of combining bimodal data with a more intricate self-attention technique.

Liu X et al. (2023) [23] have demonstrated the use of transmitted multichannel and multilayer fusion for multidimensional emotion recognition. Images for the input were gathered from CMU-MOSI and IEMOCAP. The suggested approach uses text, visual data, and voice as inputs in several modes simultaneously. It presents a unique framework for multimodal emotion detection dubbed using transmitted multichannel and layered fusion, multidimensional emotion recognition (CMC-HF). To enhance recognition performance, the CMC-HF model leverages hierarchical fusion and deep learning method for feature extraction. The model's presentation is verified by experiments conducted on the CMU-MOSI and IEMOCAP benchmark datasets. The suggested method's higher computational requirements and complexity are a limitation.

Mai S et al. (2022) [24] have suggested the hybrid contrastive learning for multimodal sentiment analysis with tri-modal representation. CMU-MOSI was the source of the input image collection in this study. By employing the suggested approach, which simultaneously carries out semi-contrastive learning and inter-/intra-modal contrasting learning, the model may fully examine cross-modal encounters, preserve inter-class linkages, and close the modality gap.

Lian Z et al. (2022) [25] have presented the SMIN: Emotion Recognition in Conversations via Multipurpose Interaction Network with Semi-Supervision. The input images were from CMU-MOSI, CMU-MOSEI, MELD, and IEMOCAP. SMIN offers two excellent semi-supervised modules for learning interactions between modes and within modes: the "Intra-modal Interaction Module (IIM)" along with the "Cross-modal Interaction Module (CIM)." These two sections extract prominent emotional representations through the use of additional unlabeled data. One limitation of this work is that its multimodal analysis does not incorporate visual information.

In 2022 Zhang F et al. [26] have suggested utilizing a deep emotional response network for heterogeneous sentiment analysis and emotion detection. The input image sets were from the CMU-MOSEI, IEMOCAP, and CMU-MOSI datasets. The network of deep emotional arousal (DEAN) model that has been suggested consists of three parts: a multimodal BiLSTM system that mimics the intellectual comparator; the human emotion algorithm's activation mechanism is simulated by a multimodal gating block, and the human perception analysis system's operations are replicated by a cross-modal transformer.

Wang N et al. (2022) [27] have presented M2R2: Mode of Missing a strong foundation for recognizing emotions with iterative data enhancement. MELD and IEMOCAP were the sources of the input images. The study suggested the Robust Missing-Modality Emotion Identification (M2R2) framework, an iterative data enhancement model for recognizing emotions is trained using this learned common representation. The suggested model still has to be tested using additional structures and methods for common representation learning, which is a disadvantage. **Table 1** displays the comparison of the existing methods.

Table 1. Comparison Of the Existing Methods

References	Dataset	Method	Advantages	Disadvantages
[21]	IEMOCAP	Self-attention Weighted methods Embrace-net Multilayer Perceptron	Fast computation and resistant to missing modalities.	Imbalanced data
[22]	Multimodal emotion dataset DEAP	Multi-stage multimodal dynamical fusion network (MSMDFN)	Utilizes inter-correlations that are well tuned across several modalities.	Requiring a lot of processing power and being difficult to implement.
[23]	CMU-MOSI and IEMOCAP	Cascaded multichannel and hierarchical fusion (CMC-HF)	Improving feature extraction within each modality and encouraging inter-modality interactions lead to improved recognition performance.	Complexity and Higher computational requirements
[24]	CMU-MOSI	HyCon for tri-modal representation's hybrid contrastive learning.	Improves generalization and cross-modal interactions.	Requires intricate tuning and implementation.
[25]	CMU-MOSI, CMU-MOSEI, MELD, and IEMOCAP	Multipurpose Interaction Network with Semi-Supervision (SMIN)	Incorporates intra-modal and cross-modal interactions. Uses unlabeled data to improve the understanding of emotions.	Does not assist with visual data.
[26]	CMU-MOSI, IEMOCAP , and CMU-MOSEI datasets	Deep Emotional Arousal Network (DEAN)	Combines attention-based theories with temporal coherence.	High requirements on computational resources.
[27]	MELD and IEMOCAP	Robust Missing-Modality Emotion Identification (M2R2)	Use data augmentation to handle missing modality.	Challenging to manage and implement.

Problem Statement

Trimodal emotion identification techniques previously in use confront many difficulties because of data imbalances, growing complexity, greater computational needs, and low accuracy. These shortcomings reduce the efficacy and efficiency of existing methods. Using cutting-edge data balancing techniques, streamlining computational procedures, and boosting accuracy using creative algorithms are the ways in which the proposed method aims to overcome these challenges.

In particular, to address data imbalance and simplify, the research method presents a novel Epistemic Multibranch Fusion Attention Neural Network (EMFANN) framework that combines data fusion and classification techniques. The proposed method incorporates cutting-edge algorithms and domain-specific knowledge to increase computational efficiency and significantly enhance emotion identification accuracy. This system offers a considerable improvement over conventional methods and opens up the possibility of more accurate and efficient emotion recognition in a variety of applications. It is designed to be more durable and scalable.

III. PROPOSED METHODOLOGY

The proposed method involves preprocessing the input images for noise reduction, cleaning, and normalization. Next, the text, audio, and video that have been preprocessed are passed through feature extraction to extract the features. Finally, the data are fused using the Multi-Branch Fusion Attention Network, and the trimodel emotion recognition is classified using the Epistemic Neural Network (ENN). Finally, the optimization is applied to the Fire Hawk optimizer. Accurate and effective trimodel emotional recognition is aided by the Epistemic Multibranch Fusion Attention Neural Network (EMFANN). EMFANN, a proposed method, assists in accurately and efficiently identifying human emotions using IEMOCAP dataset labels like frustrated, angry, neutral, sad, happy, and excited, and the CMU-MOSI dataset labels are positive, negative, and neutral. **Fig 1** shows the proposed methodology block diagram.

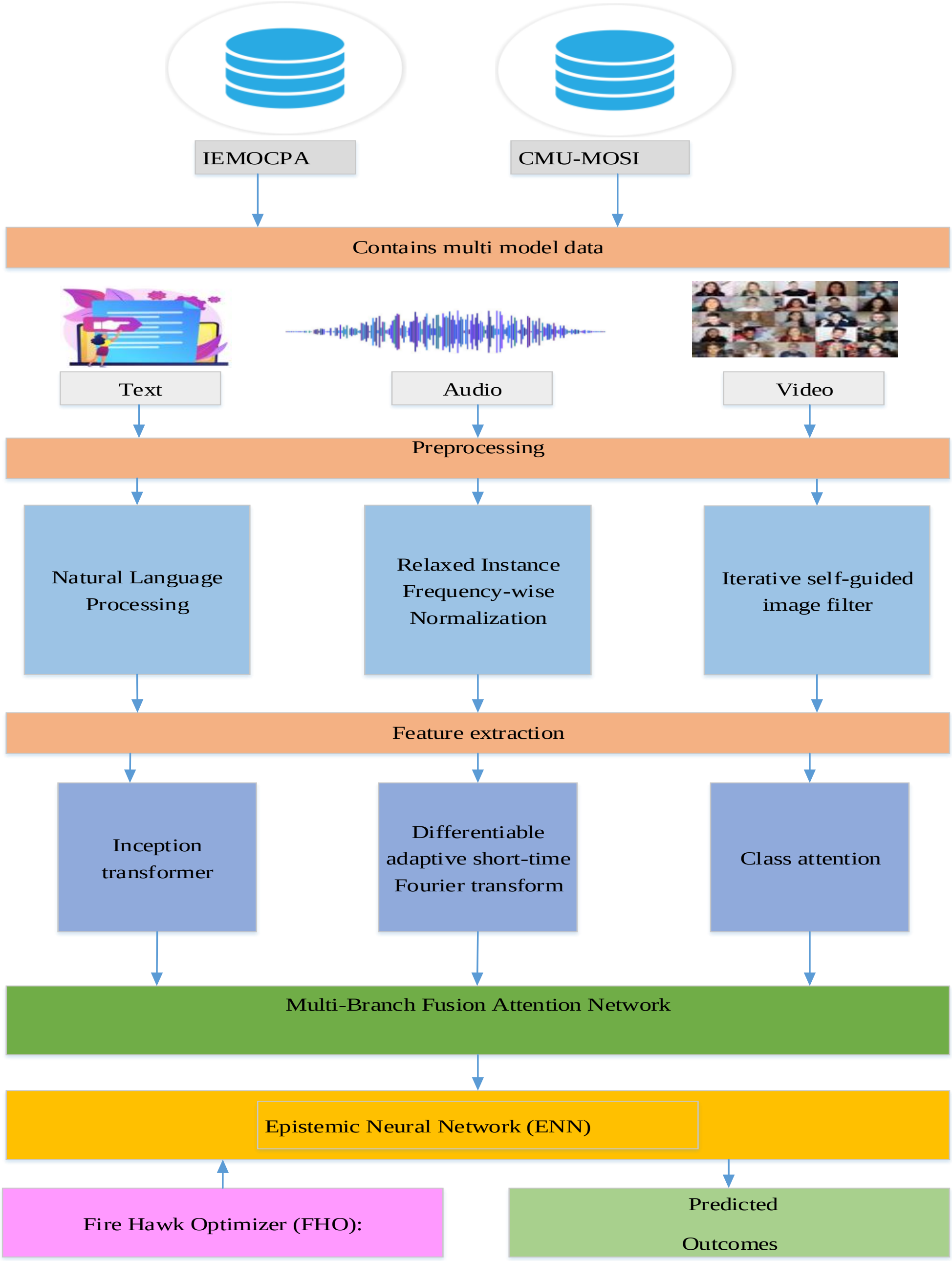
IEMOCPA
CMU-MOSI
Contains multi model data
Text
Audio
Video
Preprocessing
Natural Language Processing
Relaxed Instance Frequency-wise Normalization
Iterative self-guided image filter
Feature extraction
Inception transformer
Differentiable adaptive short-time Fourier transform
Class attention
Multi-Branch Fusion Attention Network
Epistemic Neural Network (ENN)
Fire Hawk Optimizer (FHO):
Predicted Outcomes

Fig 1. Proposed Methodology Block Diagram.

Dataset

The proposed method uses Interactive Emotional Dyadic Motion Capture (IEMOCAP) and Carnegie Mellon University Multimodal Opinion-level Sentiment Intensity (CMU-MOSI) datasets. Researchers examining the nuances of human emotions can find a wealth of information in the extensive IEMOCAP dataset. It was developed in USC's SAIL lab. One of the best examples of multimodal sentiment analysis's efficacy in the domain of emotion recognition is the CMU-MOSI dataset. The dataset is preprocessed for textual, audio, and video data normalization, standardization, noise reduction, and dimensionality reduction using the method described in the next section.

Preprocessing For Trimodel Emotion Recognition

The input images are fed into preprocessing for noise reduction, dimensionality reduction, and normalization and standardization. The proposed method uses Natural Language Processing (NLP) for text preprocessing [28], Relaxed Instance Frequency-wise Normalization (RFN) for audio preprocessing [29] and an iterative self-Guided Image Filter (isGIF) for video preprocessing [30].

NLP Based Text Preprocessing

Tokenization

Tokenization is the act of splitting up sentences into discrete tokens, which are punctuation, words, and letters. The dividing criteria are the primary area where a space or punctuation appears.

Part of Speech Tagging (POS)

POS tagging is the method of grouping words in a phrase according to the basic grammatical classes they belong in, which include adjective, verb, noun, preposition, and pronoun. To assign speech parts, the POS tagger mimics human speech by using phrase context. It can recognize speech segments more accurately with improved sentence segmentation, which reduces mistakes.

Lemmatization

Lemmatization is the process of taking a word and replacing or eliminating its suffix to reduce it to its basis, or lemma. Lemma words, in contrast to stemmed words, always have a meaning. Lemmatization is a well-liked text preparation method in NLP that yields good results.

Relaxed Instance Frequency-Wise Normalization (RFN) Based Audio Preprocessing

The method seeks to improve the consistency and relevance of the audio features for the classification model by preprocessing them using a relaxed instance frequency-wise normalization. The method presents a domain generalization module called RFN using Instance Frequency-wise Normalization (IFN) and Layer Normalization (LN) in the following ways:

$$RFN(y) = \lambda \cdot LN(y) + (1-\lambda) \cdot IFN(y) \tag{1}$$

Where y is the input to RFN, $\lambda \in [0,1]$ denotes the degree of relaxation. The suggested approach does not employ affine transformation for IFN and LN, thus the mean and standard deviation of RFN(x) that results are $\hat{\mu}_j^{(f)} = \frac{\lambda}{\sigma_j}\left(\mu_j^{(f)} - \mu_j\right)$ and $\hat{\sigma}_j^{(f)} = \lambda \cdot \frac{\sigma_j^{(f)}}{\sigma_j} + (1-\lambda),$ respectively, where $\mu_j^{(f)}$ and $\sigma_j^{(f)}$ are calculated over $S_j = \{h | h_n = j_n, \quad h_f = j_f\}$, and μ_j and σ_j are statistics over $S_j = \{h | h_n = j_n\}$. Improved generalization and increased accuracy in audio classification tasks can result from this.

Iterative Self-Guided Image Filter (isGIF) Based Video Preprocessing

For video preprocessing, initially the input videos are converted into frames. Then the frames are preprocessed using the isGIF. In image processing, the isGIF method is frequently used for tasks like detail enhancement, edge-preserving smoothing, and noise reduction. The isGIF can use the input image as the guidance image and generate excellent edge-preserving filtering results.

$$S^{h+1} = F^{-1}\left[\frac{F(J) + \lambda F(\nabla\_R_y^{h+1} + \nabla\_R_y^{h+1})}{F(1) + \lambda \sum_{d \in (y,x)} \overline{F(\nabla_d) \cdot F(\nabla_d)}}\right] \tag{2}$$

Where J denotes the input image, λ represents the smoothing weight, and $\in$ represents the parameter. Where $\nabla\_R_y^{h+1}$ represents the inverse first order derivative of R_y^{h+1} along the x-axis. The isGIF is a very useful tool for preparing images for additional analysis or processing since it continuously modifies the guide image to preserve significant image features while lowering noise. Generally, the pre-processing step is accomplished successfully as it handles issues relating to noise removal, dimensionality, convergence, normalization, and standardization for text, audio, and video information. These steps are very important in improving the competency of the input data, which, in extension, results in the improvement of the quality of the next stages of analysis and the overall performance of the model. After pre-processing, the data passes to the feature extraction stage, where relevant features are efficiently captured and extracted.

Feature extraction

To extract the features from the trimodel emotion recognition, the pre-processed text, audio, and video are fed into the feature extraction. The proposed method uses text feature extraction using an inception transformer [31], audio feature extraction using a Differentiable Adaptive Short-Time Fourier transform (DA-STFT) [32], and video feature extraction using class attention.

Inception Transformer Based Text Feature Extraction

Revisit Vision Transformer

Initially, the method uses the Vision Transformer. Transformers divide the input image into a series of tokens for vision tasks, and then each patch token is denoted as $\{x1, x2, ...xn\}$ and projected into a leaner-layered hidden representation vector. Next, a positional embedding is applied to every token, and the resulting Feed-Forward Network (FFN) and Multi-head Self-Attention (MSA) are input into the Transformer layers.

Inception Token Mixer

The proposed method combines transformers with the potent capacity of CNNs to extract high-frequency representation using an Inception mixer. The "Inception" method, which takes its cues from the Inception module, uses parallel convolution operations and max-pooling to divide input features along channel dimensions into low-frequency and high-frequency mixers.

High Frequency Mixer

The method provides a parallel framework to learn the high-frequency elements by taking into account the maximum filter's sharp sensitivity and the convolution operation's detail perception. The method divides the input x_k into $x_{k1} \in Q^{n\times\frac{c_k}{2}}$ and $x_{k2} \in Q^{n\times\frac{c_k}{2}}$ along the channel. x_{k2} is fed into a depth-wise and linear convolution layer, whereas x_{k1} is embedded with a linear and max-pooling layer.

$$y_{k1} = FC(\max pool(x_{k1})), \tag{3}$$

$$y_{k2} = DwConv(FC(x_{k2})), \tag{4}$$

Where y_{k1} and y_{k2} indicate the high-frequency mixers' outputs.

Finally, concatenation of the low- and high-frequency mixers' outputs along the channel dimension occurs:

$$y_c = concat(y_w, y_{k1}, y_{k2}) \tag{5}$$

Low Frequency Mixer

The vanilla multi-head self-attention is used for low-frequency mixer communication, but its large feature map resolution leads to high computation costs. To reduce computational overhead, an average pooling layer and up-sample layer are used, focusing on embedding global information.

$$y_w = upsample(MSA(avepooling(x_w))) \tag{6}$$

Where y_w indicates the output of low-frequency mixer.

Frequency Ramp Structure

The method presents a frequency ramp structure for visual frameworks, allowing lower layers to capture high-frequency details and gather local information. The high-frequency mixer is given fewer dimensions by the structure, which divides channel dimensions from lower to higher layers. The four stages of the backbone define a channel ratio that balances high-frequency and low-frequency components. These stages have distinct channel and spatial dimensions. The flexible structure allows inception transformer to trade-off components across all layers.

Differentiable Adaptive Short-Time Fourier Transform (DA-STFT) Based Audio Feature Extraction

The Short-Time Fourier Transform (STFT) has been for a while regarded as a fundamental method for examining time-frequency representations of audio signals in the context of audio feature extraction. However, the efficacy of standard STFT addresses in learning-based models may be limited due to their frequent shortcomings in terms of distinction and adaptation. In order to overcome these difficulties, the DA-STFT presents a framework that enables feature extraction flexibility and end-to-end optimization. By incorporating learnable parameters into the STFT process, this method allows the model to dynamically modify its time-frequency resolution in response to the demands of the task and the audio content.

The term-by-term differentiation is used since $S[j,g]$ is complex. Complex numbers are viewed as vectors that have two real components, specifically $\exp(jy) = [\cos(y), \sin(y)]$.

$$\frac{\partial S[j,g]}{\partial \theta_{jg}} = \sum_{h=0}^{n-1} \exp\left(-2i\pi \frac{hg}{n}\right) \frac{\partial f(h)}{\partial \theta_{jg}} s[t_j + h] \tag{7}$$

Where the numerical audio support, denoted by n, can be interpreted as the maximum value of the continuous temporal resolution parameter θ_{jg}. with the trimodel emotional recognition, the DA-STFT facilitates the precise and effective extraction of data from audio.

Class Attention Based Video Feature Extraction

To locate features based on categories, this model employs a class attention learning layer, and the projected layer is divided into two stages: (1) using a 1x1 convolutional layer with stride 1 to produce class attention maps and (2) obtaining class-based characteristics by vectorizing every class attention map. Next, the feature maps that were applied were retrieved from the feature extraction component. They had a $W \times W \times H$ size. Assume that w_c represents the c-th convolutional filters in the class attention learning layer. The following notion is used to complete the attention map M_c for class c:

$$M_c = Y * w_c \tag{8}$$

Where, the convolution work is displayed by c grading from 1 to several classes. Assuming that convolutional filters have a size of 1×1, a class attention map M_c can be defined as the linear integration of Y's channels. During the deployment, discriminative class focus maps are learned by the projected class focus learning layer. In order to create a class-based feature that uses all classes, the quantity of filters that are applied is equal to the quantity of classes. It is observed that absent classes are highlighted on class attention maps, which also emphasize the discriminative areas for varied courses. Class attention mappings M_c are thus vectorized to provide class-wise feature vectors $V1$ of W^2 dimensions. Class-wise connections between class attention maps and relevant concealed units are nevertheless developed by layers of FC of the class attention map to all concealed layers.

The process of extracting features from videos is aided by the class attention-based video feature extraction. The characteristics from the text, audio, and video are extracted using three distinct feature extraction algorithms in the proposed method. Following the process of extracting features from the text, audio, and video data, the features are subsequently subjected to data fusion. Then the combined data are classified using the Epistemic Neural Network (ENN). It is crucial to the trimodel emotion recognition process. It makes the outcomes of the proposed method more accurate and efficient.

Epistemic Multibranch Fusion Attention Neural Network (EMFANN)

Data fusion is the process by which the extracted features are combined. Next, for the purpose of final categorization, the combined data is fed into the classification step. The proposed method combines the multi-branch fusion attention network [33] with an epistemic neural network [34] to give a novel epistemic multibranch fusion attention neural network (EMFANN). In the proposed method the data fusion is accomplished by the Multi-Branch Fusion Attention Network, and trimodel emotion recognition is classified using the Epistemic Neural Network. The proposed EMFANN, when it comes to classifying emotions, it's critical to use a comprehensive approach that incorporates knowledge from several modalities. This method works particularly well for expressing the complex range of people emotions. In particular, the classification model separates emotions into happy, angry, neutral, and sad, frustrated, and excited negative, neutral, and positive emotions. **Fig 2** displays the architecture of EMFANN.

Multi-Branch Fusion Attention Network

In the proposed method, data are fused using the multi-branch fusion attention network. The three input feature maps should first be subjected to positional encoding before local and global features are fully integrated. Positional encoding is frequently employed in transformer systems to manage the flattened feature maps, giving each pixel a positional value. The method will solely concentrate on the $\mathrm{Re}\,s_4, Con_4$ and $Trans_4$ procedures in order to prevent monotonous repetition. To reduce the dimensionality, the flatten operation must be completed first. The following are the exact dimensions adjustments:

$$\mathrm{Re}\,s_4 \in R^{c_4^1 \times \frac{hw}{64}} \leftarrow Flatten\left(\mathrm{Re}\,s_4 \in R^{c_4^1 \times \frac{h}{8} \times \frac{w}{8}}\right) \tag{9}$$

$$Con_4 \in R^{c_4^1 \times \frac{hw}{64}} \leftarrow Flatten\left(Con_4 \in R^{c_4^1 \times \frac{h}{8} \times \frac{w}{8}}\right) \tag{10}$$

$$Trans_4 \in R^{c_4^2 \times \frac{hw}{64}} \leftarrow Flatten\left(Trans_4 \in R^{c_4^2 \times \frac{h}{8} \times \frac{w}{8}}\right) \tag{11}$$

Where, the method uses '×' to indicate changes in feature map size. It is important to remember that positional encoding does not change a feature map's size. In order to revert to the dimensions of their individual initial feature maps, the encoded Res4, Con4, and Trans4 reshape. In particular, the following size variations occur:

$$X1 \in R^{c_4^1 \times \frac{h}{8} \times \frac{w}{8}} \leftarrow \mathrm{Re}\,shape\left(\mathrm{Re}\,s_4 \in R^{c_4^1 \times \frac{hw}{64}}\right) \tag{12}$$

$$X2 \in R^{c_4^1 \times \frac{h}{8} \times \frac{w}{8}} \leftarrow \mathrm{Re}\,shape\left(Con_4 \in R^{c_4^1 \times \frac{hw}{64}}\right) \tag{13}$$

$$X3 \in R^{c_4^1 \times \frac{h}{8} \times \frac{w}{8}} \leftarrow \mathrm{Re}\,shape\left(Trans_4 \in R^{c_4^2 \times \frac{hw}{64}}\right) \tag{14}$$

Fully integrating local and global features will be accomplished through the usage of a multi-branch aggregation attention (MAA), using the acquired $X1, X2,$ and $X3$.

Attention To Multi-Branch Aggregation

Inspired by the self-attention mechanism in ViT, the proposed method built MAA within the multi-branch attention fusion module for trimodal data fusion in order to enhance the accurateness of emotion recognition and boost the detection of subtle emotional cues. The process generates a query (T), key (A), and value (V) by applying a depthwise separable convolution with a 3 × 3 kernel size on $X1, X2,$ and $X3$.

$$MT = \mathrm{Re}\,shape(T) \tag{15}$$

$$MA = \mathrm{Re}\,shape(A) \tag{16}$$

$$MV = \mathrm{Re}\,shape(V) \tag{17}$$

$$F_W = Soft\max\left(MA \otimes MA^T\right) \otimes MV \tag{18}$$

Where $\operatorname{Re}shape(\cdot)$ denotes the arrange operation, $MT \in R^{\varphi\times\frac{hw}{64}\times\frac{c_4^1}{\varphi}}, MA, MV \in R^{\varphi\times\frac{hw}{64}\times\frac{c_4^2}{\varphi}}$ and φ indicates the number of heads in the multi-head focus and $\otimes$ denotes matrix multiplication. After integration, the data is given into the classification phase. Text, audio and video data fusion improves classification by utilizing the complementing features of each modality, resulting in more reliable and accurate emotion recognition.

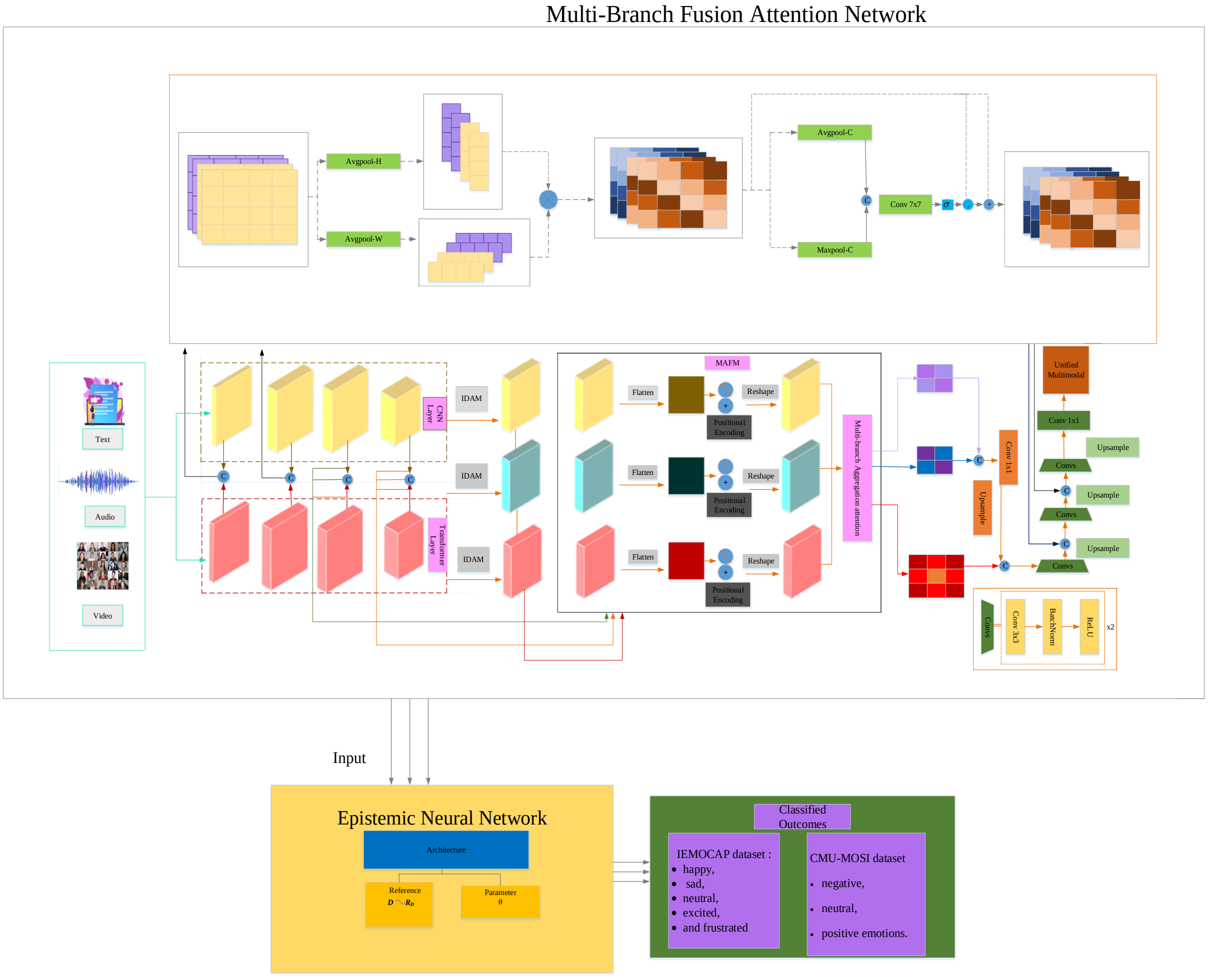

Fig 2. Architecture of EMFANN.

Epistemic Neural Network (ENN)

The integrated data is classified using the ENN. The ENN architecture is defined by two: a reference distribution R_D and a parameterized function class f_c. An epistemic index d, which denotes epistemic uncertainty, determines the output $f_c\theta(x,d)$. Changes in the network's output with d provide valuable joint forecasts by indicating future data-resolved uncertainty.

A joint prediction, given inputs $x1,...., xn$ gives each class combination $y1,...., yn$ a probability $\hat{R}_{1:n}(y1:n)$. Although joint predictions are not intended by typical neural networks, joint estimates can be generated by multiplying marginal predictions:

$$\hat{R}_{1:n}^{NN}(y1:n) = \prod_{t=1}^{n} soft\max(f_c\theta(xt))_{yt} \tag{19}$$

Unfortunately, this representation cannot discriminate between ambiguity and insufficient data because it treats each outcome $y1:n$ as independent. In order to overcome this, ENNs integrate across epistemic indices to enable more expressive joint predictions:

$$\hat{R}^{ENN}_{1:n}(y1:n) = \int_{d} R_D(zd) \prod_{t=1}^{n} soft\max(f_c\theta(xt,d))_{yt} \tag{20}$$

In order to ensure that joint predictions are more than merely the product of marginal, dependencies are introduced through this integration. The complex range of human emotions is particularly well-captured using this method. The ENN is used in the classification stage to efficiently classify the combined data. The next stage after classification is optimization, which involves adjusting the neural network's parameters to improve accuracy and performance. By ensuring that the model converges to the optimal solution, optimization raises the model's overall effectiveness and capacity for prediction.

Fire Hawk Optimizer (FHO)

Native Australians rely on fire to preserve both cultural practices and the balance of the ecology. Fires can be ignited on purpose or happen spontaneously as a result of lightning, making local wildlife more vulnerable. Flames Hawks purposefully propagate fire by transporting flaming sticks, a practice that causes havoc. These little fires frighten the animal, which facilitates easier hawk capture. The propagation of fire is also aided by other elements such as black kites, whistling kites, and brown falcons. The FHO is used to optimize the loss function in relation to the ENN's weight parameters [35].

$$Fitness\ \ function = \min\,[L(\theta)] \tag{21}$$

Where, θ denotes the weight parameter of the ENN. **Table 2** shows the pseudocode of FHO

Table 2. Pseudocode of FHO

```
Procedure FHO
  Determine the starting locations of solution candidates $(xi)$ in the search space containing $n$ candidates
  Evaluate the fitness values of the first potential solutions.
  Choose the Global best (Gb) option to be the primary fire
  While Iteration < Maximum Iteration Count
     In order to find the number of Fire Hawks, generate n as a random integer.
     Determine the prey (Pr) and fire hawks (Fh) in the search area.
     Determine the total distance that the Fire Hawks must go to reach their target.
     By distributing the prey, determine the Fire Hawks' territory.
    for $l = 1:N$
      Utilizing Equation 22, ascertain the Fire Hawks' new location.

        $Fh_1^{new} = Fh_1 + (R_1 \times Gb - R_2 \times Fh_{near}),\ \ l = 1,2,..., N,$        (22)

       for $q = 1:R$
         Using Equation 23, determine the safe area beneath Fire Hawk territory.

           $Sp_1 = \frac{\sum_{q=1}^{R} \Pr_q}{R},\quad \begin{cases} q = 1,2,..., R. \\ l = 1,2,...., N. \end{cases}$        (23)

         Using Equation 24, ascertain the preys' new location.

           $\Pr_q^{new} = \Pr_q + (R_3 \times Fh_1 - R_4 \times Sp_1),\quad \begin{cases} l = 1,2,..., N. \\ q = 1,2,..., R. \end{cases}$        (24)

         Using Equation 25, determine the safe location outside of the Ith Fire Hawk's domain.
```

```
                $Sp = \frac{\sum_{k=1}^{M} \Pr_k}{M}, \quad k = 1,2,...., M$        (25)

        Utilizing Equation 26, determine the preys' new location.

                $\Pr_q^{new} = \Pr_q + \left(R_5 \times Fh_{alter} - R_6 \times Sp\right), \quad \begin{cases} l = 1,2,...., N. \\ q = 1,2,..., R. \end{cases}$        (26)

      end
  end
Examine the newly developed Fire Hawks' and preys' fitness values.
Choose the Global best (Gb) option to be the primary fire
end while
return Gb
end procedure
```

The proposed three algorithms also bring novelty to trimodal emotion recognition by providing end-to-end solutions for pre-processing, feature extraction, data fusion, classification, and optimization. A clean and appropriate data set can be obtained through the pre-processing stage, which provides an excellent basis for feature extraction. Fine-grained emotional cues can be extracted from audio, video, and text using the most advanced feature extraction methodology. These elements are effectively integrated into the data fusion stage, and the efficiency and accuracy of emotion recognition are enhanced in the classification stage by the Fire Hawk optimization and the ENN. In addition to that, it overcomes the limitations of current work and advances a novel paradigm for multimodal emotion analysis.

IV. RESULTS AND DISCUSSION

This section presents the F1-score, accuracy, recall, and precision metrics for the EMFANN model on the current system using the IEMOCAP and CMU-MOSI datasets. The efficacy and precision of EMFANN in identifying emotions through text, video, and audio modalities is also assessed and contrasted with alternative methods. Throughout the analysis, the model's advantages and possible areas for improvement are recognized. The proposed method is implemented in Python.

Dataset Description

The field of emotional evaluation and partiality in internet opinion videos uses a large dataset to understand complex emotions, enabling practitioners and scholars to develop multimodal sentiment analysis tools. The proposed method used IEMOCAP and CMU-MOSI datasets.

IEMOCAP Dataset

A comprehensive view of human emotions is provided by the IEMOCAP dataset, which comes from the SAIL lab at USC. It includes text transcriptions, audio, video, and face motion capture. The data, which includes text, audio, videos, and physical signals, spans twelve hours. A sophisticated sensitivity recognition system may be developed using this dataset, which is also useful for in-depth study. The classes are categorized as happy, angry neutral, sad, frustrated and excited using the IEMOCAP dataset.

CMU-MOSI Dataset

Combining visual and aural data, the CMU-MOSI is an effective tool for sentiment recognition that can identify a wide range of emotions from articulated words. It is useful for emotion detection models in real-world communication contexts because of its thorough categorization of emotions. The classes are categorized as positive, negative, and neutral using the CMU-MOSI dataset.

Performance Evaluation

Accuracy

The percentage of accurately anticipated instances in the dataset is quantified by this metric. It is mathematically expressed in equation (27):

$$accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{27}$$

Where T_p denotes the True Positive, T_N denotes the True Negative, F_p represents the False Positive and F_N represents the False Negative.

Precision

The precision of a model indicates its ability to prevent false positives. It is calculated as the ratio of correctly predicted positive cases to all positive predictions in equation (28):

$$precision = \frac{T_P}{T_P + F_P} \tag{28}$$

Recall

The percentage of real positives that are successfully identified is measured by recall, also known as sensitivity, and is given by (29):

$$recall = \frac{T_P}{T_P + F_N} \tag{29}$$

F1-score

The harmonic mean of recall and precision, or the F1 score, provides a balance among the two, which is important when there are differences in the class, as shown in (30):

$$f1 - score = \frac{2 * precision * recall}{precision + recall} \tag{30}$$

Performance Comparison with Existing Methods

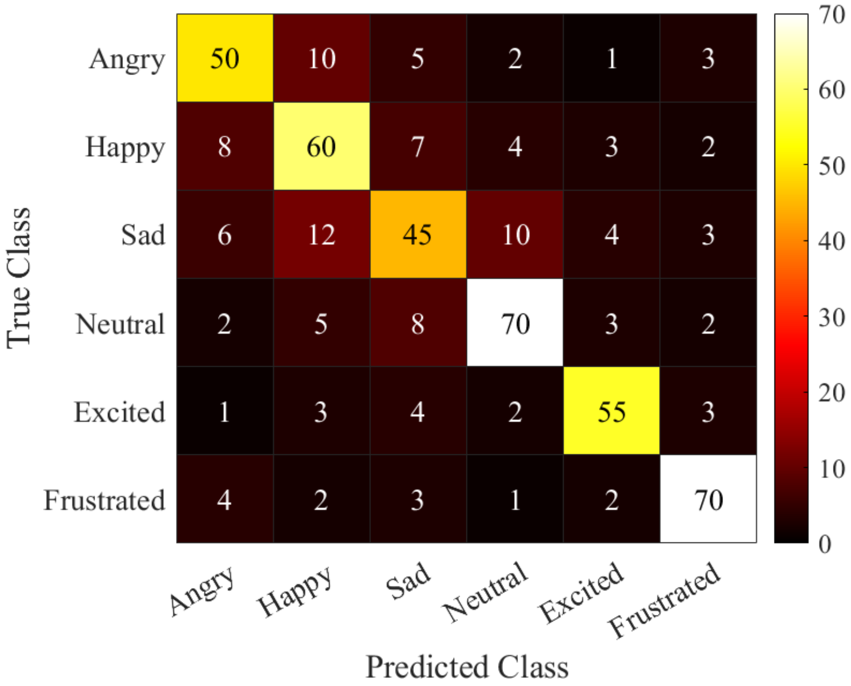

Fig 3. Confusion Matrix of IEMOCAP Dataset.

The IEMOCAP dataset's trimodal emotion recognition model's performance is shown in the **Fig 3**. The matrix contrasts the true and anticipated classes for each of the six emotions: frustrated, sad, excited, neutral, and happy. High diagonal values indicate accurate predictions; "neutral" and "frustrated" are the most properly classified groups.

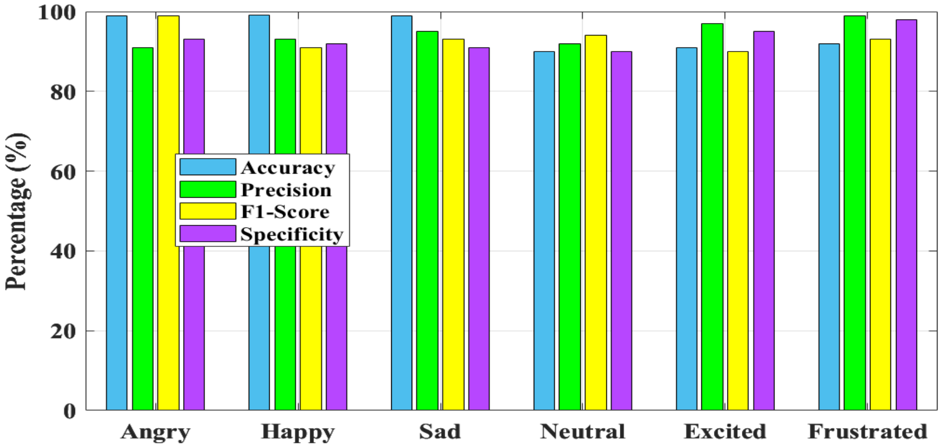

Fig 4. Performance Comparison of IEMOCAP Dataset.

Fig 4 displays the performance comparison of IEMOCAP dataset. The IEMOCAP dataset's six emotions: angry, happy, sad, neutral, excited, and frustrated are represented by the graph performance measures, which include accuracy, precision, F1-score, and specificity. The model is trimodal in nature. The robustness of the model's ability in identifying various emotional states is demonstrated by the consistently high metrics with minimal fluctuations across all emotions.

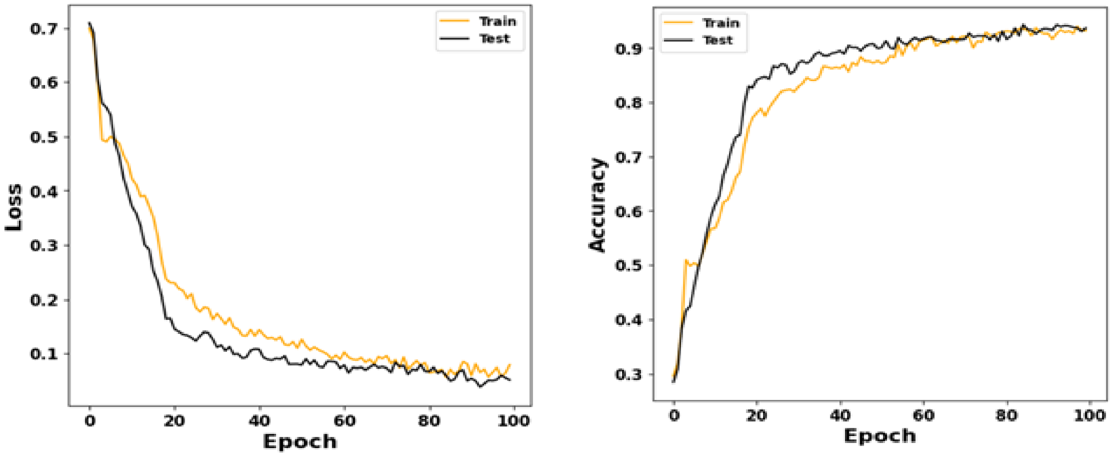

Fig 5. Training and Testing Loss and Accuracy of IEMOCAP Dataset.

The trimodal emotion recognition graphic uses the IEMOCAP dataset to illustrate the "loss" and "accuracy" of 100 epochs for training and testing datasets are shown in **Fig 5**. The "accuracy" graph increases smoothly, indicating the evolution of the model's performance, while the "loss" graph falls rapidly before steadily rising, indicating learning and stability.

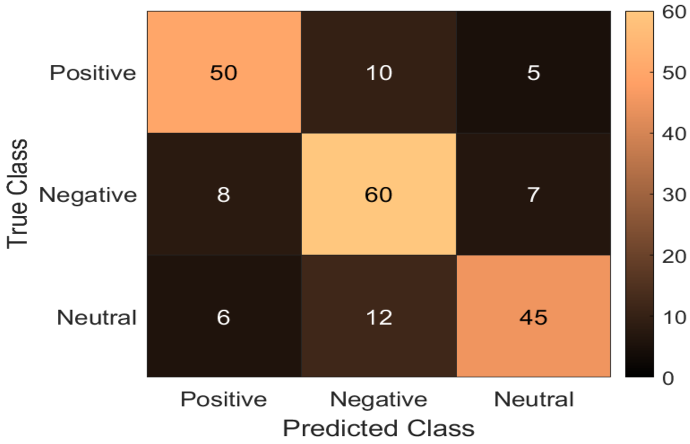

Fig 6. Confusion Matrix of CMU-MOSI Dataset.

Fig 6 displays the confusion matrix of CMU-MOSI dataset. Using the CMU-MOSI dataset, the following comes out as the confusion matrix for the trimodal emotion recognition system: It focuses on false negative, true positive, false positive and true negative coefficients to identify the degrees of accuracy and cases of misclassification of the model.

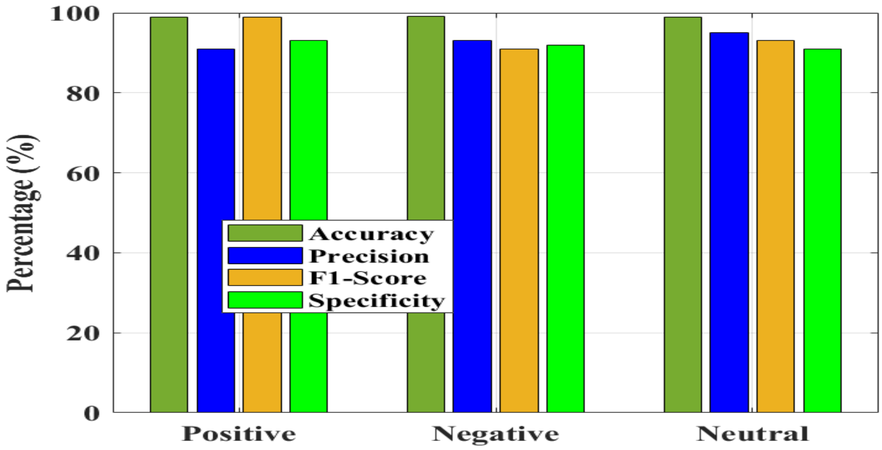

Fig 7. Performance Comparison of CMU-MOSI Dataset.

Fig 7 shows the performance comparison of CMU-MOSI dataset. The model attains almost consistent accuracy (99%, 99.1%, 99 %) in all sentiment categories in the CMU-MOSI dataset. F1 scores rise for positive (99) and fall for negative (91). Precision is highest for neutral (95) and lowest for positive (91). For positive (93) and neutral (91) cases, specificity is a bit higher and lower, respectively.

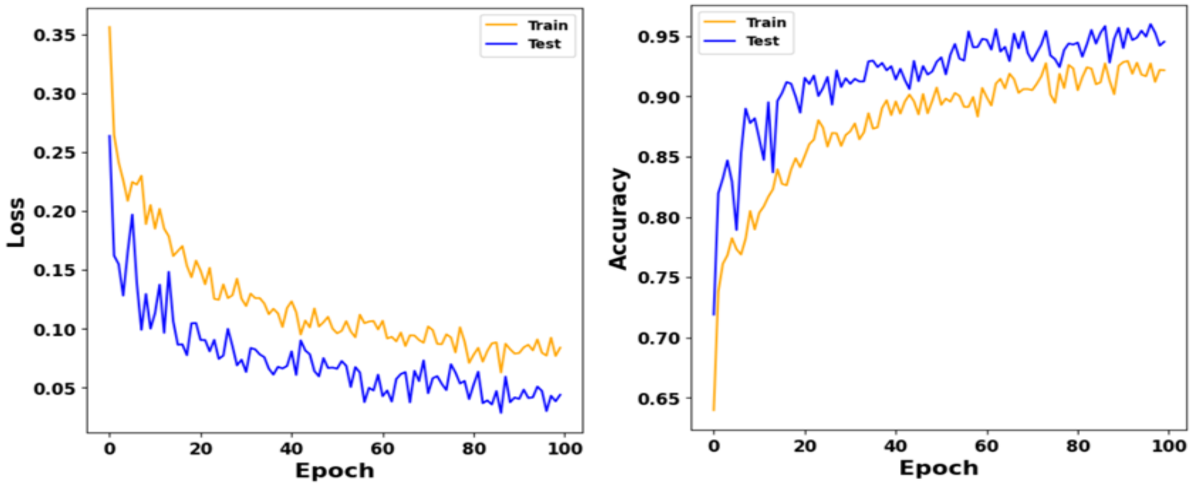

Fig 8. Training and Testing Loss and Accuracy of CMU-MOSI Dataset.

Fig 8 contains a performance comparison of a trimodal emotional recognition system having trained on the CMU-MOSI dataset. The first graph represents the training and testing loss on the left side and it also shows a decreasing trend till 100 epochs of training which also means that the method is performing better. The right graph then shows a trend of improvement of the accuracy when training an/tested and proves that the model can generalize.

Table 3. IEMOCAP Dataset Performance Comparison of Existing Methods

Methods	Accuracy	Precision	Recall	F1-score	Error rate	Computational time
MLP	57.42	56.8	57	56.9	42.58%	Medium
CMC-HF	91.2	90.9	91	90.8	8.8%	High
CA-WGNN	93	93	93	93	7%	Very high
DNN	73.15	72.8	73	72.9	26.85%	Medium
Proposed EMFANN	94.5	94.7	94.6	94.65	5.5%	Very low

When evaluating the several methods to trimodel emotion identification, the proposed EMFANN model performs the best on all measures, with 94.7% precision, 94.6% recall, 94.5% accuracy, and 94.65% F1-score. Notably, although it has a sophisticated architecture, it also has a relatively low computational time, demonstrating its efficiency. On the other hand, conventional models such as MLP and DNN perform much less well, with medium computation durations and accuracy of 57.42% and 73.15%, respectively. The CA-WGNN model requires a very high computing time, even if it performs well, with 93% accuracy and balanced precision, recall, and F1-scores. With an accuracy of 91.2%, CMC-HF likewise produces excellent results, although at a significant computational cost. Overall, EMFANN is the most practical and effective approach in the comparison, outperforming the others not only in accuracy and F1-score but also in efficiency. **Table 3** shows the IEMOCAP dataset performance comparison of existing methods.

Table 4. CMU-MOSI Dataset Performance Comparison of Existing Methods

Methods	Accuracy (%)	Precision(%)	Recall (%)	F1-score(%)	Error rate	Computational time
CA-WGNN	94	94	94	94	6%	Very high
Gated Self-Attentive Recurrent Multimodal Fusion	83.91	83.5	83.7	81.17	16.09%	High
Open Face + VGG16	61.53	60.5	61	60.73	38.47%	Medium
HyCon	85.2	84.9	85	85.1	14.8%	High
Proposed EMFANN	99.5	99.7	99.6	99.65	4.5%	Very low

Tested on the CMU-MOSI dataset, the proposed EMFANN model outperforms the most recent methods such as CA-WGNN, Gated Self-Attentive Recurrent Multimodal Fusion, Open Face + VGG16, and HyCon. EMFANN is the most practical and efficient method for multimodal emotion recognition because it produces excellent accuracy (99.5%) with a very short calculation time. **Table 4** shows the CMU-MOSI dataset performance comparison of existing methods.

Table 5. Performance Metrics for Fusion Methods of IEMOCAP Dataset

	Accuracy	Precision	Recall	F1-Score
Fusion Macro Average	0.94	0.94	0.94	0.94
Fusion Weighted Average	0.94	0.94	0.94	0.94
Fusion Overall Accuracy	0.99	0.94	0.94	0.94

Performance metrics for fusion approaches are displayed in the table. Precision, recall, accuracy, and an F1-score of 0.99 demonstrate that the macro-average, weighted average, and overall accuracy all produce consistent results. This shows that the fusion strategy performs robustly in integrating numerous information sources or modalities and that it is highly effective and balanced across all evaluation metrics. **Table 5** shows the performance metrics for fusion methods of IEMOCAP Dataset.

Table 6. Performance Metrics for Fusion Methods of CMU-MOSI Dataset

	Accuracy	Precision	Recall	F1-Score
Fusion Macro Average	0.95	0.95	0.95	0.95
Fusion Weighted Average	0.95	0.95	0.95	0.95
Fusion Overall Accuracy	0.99	0.99	0.99	0.99

Performance metrics for fusion methods assessed using the CMU-MOSI dataset are shown in the table. It demonstrates that with precision, F1-Score, recall, and an accuracy of 0.99, the macro average, weighted average, and overall accuracy all produce consistent results. This highlights the resilience of the fusion strategy in integrating data from the CMU-MOSI dataset and shows that it operates efficiently and consistently across all measures. **Table 6** shows the performance metrics for fusion methods of CMU-MOSI dataset.

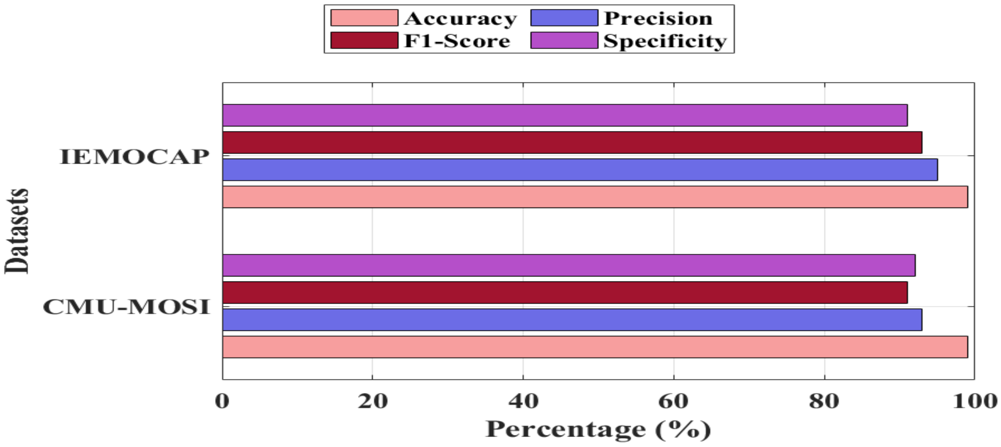

Fig 9. Comparison of IEMOCAP and CMU-MOSI Datasets.

Fig 9 displays the comparison of IEMOCAP and CMU-MOSI datasets. The model performs exceptionally well (99% and 99.1%) on the CMU-MOSI dataset, with variable precision (91 for positive, 95 for neutral) and F1 scores (99 for positive, 91 for negative). There are 91 to 93 levels of specificity. The IEMOCPA dataset, on the other hand, has a little less accuracy (99%), a peak in precision at 95, and F1 scores as high as 93. At 91, specificity is similar. The overall accuracy and precision variability are slightly higher in the CMU-MOSI dataset.

V. CONCLUSION

A significant advance in the recognition of emotions is the proposed Epistemic Multi-Branch Fusion Attention Neural Network (EMFANN), which effectively integrates three advanced pre-processing methods: NLP for text, RFN for audio, and is GIF for video. EMFANN accurately extracts and refines emotional cues by utilizing novel feature extraction techniques such as Inception Transformers, DA-STFT, and class attention mechanisms. While the Epistemic Neural Network improves classification reliability by estimating uncertainty, the Multi-Branch Fusion Attention Network effectively integrates these aspects. EMFANN, enhanced by the Fire Hawk Algorithm, has exceptional resilience and efficiency, as seen by its exceptional precision, F1-score, recall, and accuracy. The model's capacity to precisely identify human emotions in a variety of modalities establishes a new standard in the field, providing notable advancements over current approaches and demonstrating its potential for useful applications in emotion recognition technology. Future research will concentrate on adding more datasets to improve the system's functionality and expanding the spectrum of emotions the model is able to identify. This will enhance the accuracy and resilience of the model in a diverse range of real-world applications.

CRediT Author Statement

The authors confirm contribution to the paper as follows:
Conceptualization: Bangar Raju Cherukuri; **Methodology:** Bangar Raju Cherukuri; **Software:** Bangar Raju Cherukuri; **Data Curation:** Bangar Raju Cherukuri; **Writing- Original Draft Preparation:** Bangar Raju Cherukuri; **Visualization:** Bangar Raju Cherukuri; **Investigation:** Bangar Raju Cherukuri; **Supervision:** Bangar Raju Cherukuri; **Validation:** Bangar Raju Cherukuri; **Writing- Reviewing and Editing:** Bangar Raju Cherukuri; All authors reviewed the results and approved the final version of the manuscript.

Data Availability

No data was used to support this study.

Conflicts of Interests

The author(s) declare(s) that they have no conflicts of interest.

Funding

No funding agency is associated with this research.

Competing Interests

There are no competing interests

References

[1]. H. F. T. Al-Saadawi and R. Das, “TER-CA-WGNN: Trimodel Emotion Recognition Using Cumulative Attribute-Weighted Graph Neural Network,” Applied Sciences, vol. 14, no. 6, p. 2252, Mar. 2024, doi: 10.3390/app14062252.

[2]. A. Aslam, A. B. Sargano, and Z. Habib, “Attention-based multimodal sentiment analysis and emotion recognition using deep neural networks,” Applied Soft Computing, vol. 144, p. 110494, Sep. 2023, doi: 10.1016/j.asoc.2023.110494.

[3]. P. Bhattacharya, R. K. Gupta, and Y. Yang, “Exploring the Contextual Factors Affecting Multimodal Emotion Recognition in Videos,” IEEE Transactions on Affective Computing, vol. 14, no. 2, pp. 1547–1557, Apr. 2023, doi: 10.1109/taffc.2021.3071503.

[4]. G.-N. Dong, C.-M. Pun, and Z. Zhang, “Temporal Relation Inference Network for Multimodal Speech Emotion Recognition,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 9, pp. 6472–6485, Sep. 2022, doi: 10.1109/tcsvt.2022.3163445.

[5]. X. Zhang, M. Li, S. Lin, H. Xu, and G. Xiao, “Transformer-Based Multimodal Emotional Perception for Dynamic Facial Expression Recognition in the Wild,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 5, pp. 3192–3203, May 2024, doi: 10.1109/tcsvt.2023.3312858.

[6]. G. Kaur and A. Sharma, “A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis,” Journal of Big Data, vol. 10, no. 1, Jan. 2023, doi: 10.1186/s40537-022-00680-6.

[7]. S. Lee, D. K. Han, and H. Ko, “Multimodal Emotion Recognition Fusion Analysis Adapting BERT With Heterogeneous Feature Unification,” IEEE Access, vol. 9, pp. 94557–94572, 2021, doi: 10.1109/access.2021.3092735.

[8]. X. Wang, J. He, Z. Jin, M. Yang, Y. Wang, and H. Qu, “M2Lens: Visualizing and Explaining Multimodal Models for Sentiment Analysis,” IEEE Transactions on Visualization and Computer Graphics, vol. 28, no. 1, pp. 802–812, Jan. 2022, doi: 10.1109/tvcg.2021.3114794.

[9]. S. S. Hosseini, M. R. Yamaghani, and S. Poorzaker Arabani, “Multimodal modelling of human emotion using sound, image and text fusion,” Signal, Image and Video Processing, vol. 18, no. 1, pp. 71–79, Aug. 2023, doi: 10.1007/s11760-023-02707-8.

[10]. A. Chaudhari, C. Bhatt, A. Krishna, and C. M. Travieso-González, “Facial Emotion Recognition with Inter-Modality-Attention-Transformer-Based Self-Supervised Learning,” Electronics, vol. 12, no. 2, p. 288, Jan. 2023, doi: 10.3390/electronics12020288.

[11]. A. Yousaf et al., “Emotion Recognition by Textual Tweets Classification Using Voting Classifier (LR-SGD),” IEEE Access, vol. 9, pp. 6286–6295, 2021, doi: 10.1109/access.2020.3047831.

[12]. L. Zhu, X. Zhu, J. Guo, and S. Dietze, “Exploring rich structure information for aspect-based sentiment classification,” Journal of Intelligent Information Systems, vol. 60, no. 1, pp. 97–117, Jul. 2022, doi: 10.1007/s10844-022-00729-1.

[13]. A. Onan, “Bidirectional convolutional recurrent neural network architecture with group-wise enhancement mechanism for text sentiment classification,” Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 5, pp. 2098–2117, May 2022, doi: 10.1016/j.jksuci.2022.02.025.

[14]. N.-H. Ho, H.-J. Yang, S.-H. Kim, and G. Lee, “Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network,” IEEE Access, vol. 8, pp. 61672–61686, 2020, doi: 10.1109/access.2020.2984368.

[15]. L. Schoneveld, A. Othmani, and H. Abdelkawy, “Leveraging recent advances in deep learning for audio-Visual emotion recognition,” Pattern Recognition Letters, vol. 146, pp. 1–7, Jun. 2021, doi: 10.1016/j.patrec.2021.03.007.

[16]. A. I. Middya, B. Nag, and S. Roy, “Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities,” Knowledge-Based Systems, vol. 244, p. 108580, May 2022, doi: 10.1016/j.knosys.2022.108580.

[17]. M. Hao, W.-H. Cao, Z.-T. Liu, M. Wu, and P. Xiao, “Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features,” Neurocomputing, vol. 391, pp. 42–51, May 2020, doi: 10.1016/j.neucom.2020.01.048.

[18]. K. Zhang, Y. Li, J. Wang, E. Cambria, and X. Li, “Real-Time Video Emotion Recognition Based on Reinforcement Learning and Domain Knowledge,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 3, pp. 1034–1047, Mar. 2022, doi: 10.1109/tcsvt.2021.3072412.

[19]. T. Zhang, A. El Ali, C. Wang, A. Hanjalic, and P. Cesar, “CorrNet: Fine-Grained Emotion Recognition for Video Watching Using Wearable Physiological Sensors,” Sensors, vol. 21, no. 1, p. 52, Dec. 2020, doi: 10.3390/s21010052.

[20]. C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J. M. Montero, and F. Fernández-Martínez, “Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning,” Sensors, vol. 21, no. 22, p. 7665, Nov. 2021, doi: 10.3390/s21227665.

[21]. D. Pena, A. Aguilera, I. Dongo, J. Heredia, and Y. Cardinale, “A Framework to Evaluate Fusion Methods for Multimodal Emotion Recognition,” IEEE Access, vol. 11, pp. 10218–10237, 2023, doi: 10.1109/access.2023.3240420.

[22]. S. Chen, J. Tang, L. Zhu, and W. Kong, “A multi-stage dynamical fusion network for multimodal emotion recognition,” Cognitive Neurodynamics, vol. 17, no. 3, pp. 671–680, Jul. 2022, doi: 10.1007/s11571-022-09851-w.

[23]. X. Liu, Z. Xu, and K. Huang, “Multimodal Emotion Recognition Based on Cascaded Multichannel and Hierarchical Fusion,” Computational Intelligence and Neuroscience, vol. 2023, no. 1, Jan. 2023, doi: 10.1155/2023/9645611.

[24]. S. Mai, Y. Zeng, S. Zheng, and H. Hu, “Hybrid Contrastive Learning of Tri-Modal Representation for Multimodal Sentiment Analysis,” IEEE Transactions on Affective Computing, vol. 14, no. 3, pp. 2276–2289, Jul. 2023, doi: 10.1109/taffc.2022.3172360.

[25]. Z. Lian, B. Liu, and J. Tao, “SMIN: Semi-Supervised Multi-Modal Interaction Network for Conversational Emotion Recognition,” IEEE Transactions on Affective Computing, vol. 14, no. 3, pp. 2415–2429, Jul. 2023, doi: 10.1109/taffc.2022.3141237.

[26]. Y. Y. Obaid Al Belushi, P. Jasmin Dennis, S. Deepa, V. Arulkumar, D. Kanchana, and R. Y. P, “A Robust Development of an Efficient Industrial Monitoring and Fault Identification Model using Internet of Things,” 2024 IEEE International Conference on Big Data &amp;amp; Machine Learning (ICBDML), pp. 27–32, Feb. 2024, doi: 10.1109/icbdml60909.2024.10577363.

[27]. N. Wang, H. Cao, J. Zhao, R. Chen, D. Yan, and J. Zhang, “M2R2: Missing-Modality Robust Emotion Recognition Framework With Iterative Data Augmentation,” IEEE Transactions on Artificial Intelligence, vol. 4, no. 5, pp. 1305–1316, Oct. 2023, doi: 10.1109/tai.2022.3201809.

[28]. C. P. Chai, “Comparison of text preprocessing methods,” Natural Language Engineering, vol. 29, no. 3, pp. 509–553, Jun. 2022, doi: 10.1017/s1351324922000213.

[29]. B. Kim, S. Yang, J. Kim, H. Park, J. Lee, and S. Chang, “Domain Generalization with Relaxed Instance Frequency-wise Normalization for Multi-device Acoustic Scene Classification,” Interspeech 2022, pp. 2393–2397, Sep. 2022, doi: 10.21437/interspeech.2022-61.

[30]. L. He, Y. Xie, S. Xie, Z. Jiang, and Z. Chen, “Iterative Self-Guided Image Filtering,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 8, pp. 7537–7549, Aug. 2024, doi: 10.1109/tcsvt.2024.3374758.

[31]. C. Si, W. Yu, P. Zhou, Y. Zhou, X. Wang, and S. Yan, “Inception transformer”, Advances in Neural Information Processing Systems, vol. 35, pp.23495-23509. 2022.

[32]. M. Leiber, Y. Marnissi, A. Barrau, and M. E. Badaoui, “Differentiable Adaptive Short-Time Fourier Transform with Respect to the Window Length,” ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5, Jun. 2023, doi: 10.1109/icassp49357.2023.10095245.

[33]. H. Gu, G. Gu, Y. Liu, H. Lin, and Y. Xu, “Multi-Branch Attention Fusion Network for Cloud and Cloud Shadow Segmentation,” Remote Sensing, vol. 16, no. 13, p. 2308, Jun. 2024, doi: 10.3390/rs16132308.

[34]. I. Osband, Z. Wen, S.M. Asghari, V. Dwaracherla, M. Ibrahimi, X. Lu, and B. Van Roy, “Epistemic neural networks”, Advances in Neural Information Processing Systems, vol.36. 2024.

[35]. M. Azizi, S. Talatahari, and A. H. Gandomi, “Fire Hawk Optimizer: a novel metaheuristic algorithm,” Artificial Intelligence Review, vol. 56, no. 1, pp. 287–363, Jun. 2022, doi: 10.1007/s10462-022-10173-w.