Human Emotion Recognition by Integrating Facial and Speech Features: An Implementation of Multimodal Framework using CNN

P V V S Srinivas, Pragnyaban Mishra
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation (KLEF)
Guntur, India

***Abstract*—This Emotion recognition plays a prominent role in today's intelligent system applications. Human computer interface, health care, law, and entertainment are a few of the applications where emotion recognition is used. Humans convey their emotions in the form of text, voice, and facial expressions, thus developing a multimodal emotional recognition system playing a crucial role in human-computer or intelligent system communication. The majority of established emotional recognition algorithms only identify emotions in unique data, such as text, audio, or image data. A multimodal system uses information from a variety of sources and fuses the information by using fusion techniques and categories to improve recognition accuracy. In this paper, a multimodal system to recognise emotions was presented that fuses the features from information obtained from heterogenous modalities like audio and video. For audio feature extraction energy, zero crossing rate and Mel-Frequency Cepstral Coefficients (MFCC) techniques are considered. Of these, MFCC produced promising results. For video feature extraction, first the videos are converted to frames and stored in a linear scale space by using a spatial temporal Gaussian Kernel. The features from the images are further extracted by applying a Gaussian weighted function to the second momentum matrix of linear scale space data. The Marginal Fisher Analysis (MFA) fusion method is used to fuse both the audio and video features, and the resulted features are given to the FERCNN model for evaluation. For experimentation, the RAVDESS and CREMAD datasets, which contain audio and video data, are used. Accuracy levels of 95.56, 96.28, and 95.07 on the RAVDESS dataset and accuracies of 80.50, 97.88, and 69.66 on the CREMAD dataset in audio, video, and multimodal modalities are achieved, whose performance is better than the existing multimodal systems.**

Keywords*—*Emotion recognition; multimodal; fusion; MFCC; MFA; FERCNN; CREMAD; RAVDESS

I. Introduction

Emotion recognition is the process of determining a person's emotional state. Affective computing and human-computer interaction (HCI) applications rely heavily on it [1]. In recent studies, emotion identification has sparked increased attention in academics and the commercial sector [2]. It is used in a variety of applications, including analysis of Twitter, tutoring systems, playing video games, prediction of consumer satisfaction, and military healthcare [3-5].

Speech or audio emotion recognition has been employed in medical studies to examine the changes in the emotions of depressed patients and in children who are having communication difficulties. It can also be used to warn the drivers during driving when the condition of the driver is fatigued to avoid accidents. Low level information from the speech or audio signals is extracted by the speech or audio emotion recognition system to comprehend the emotion status. Compilation of databases related to emotions, extraction of emotional features from the speech or audio signal, reduction of features by using dimensionality reduction techniques, and classification of emotions into respective classes are all part of this classification problem based on speech or audio signal sequences. K nearest neighbour, Gaussian mixture model, Support vector machines, and artificial neural networks are some of the traditional techniques that are used for speech or audio emotional recognition and are not that efficient because human emotions have high complexity and uncertainty [6].

About 93% of communication with humans is done through nonverbal means such as voice tone, facial expressions, and body language [7]. Identifying emotions through facial expressions which has been extensively studied [8][9] resulted in higher accuracies by making the changes at the pre-processing stage. To reduce overfitting during the training stage, adding dropout to the CNN model plays a prominent role in reducing overfitting during training [10]. Extracting of faces from the chain of video sequences and extracting the features from the resulted images are the steps followed in general to detect the emotions of the faces in the video sequences [11]. The robust face detection algorithm [12], the AdaBoost learning algorithm [13], and the spatial template tracker [14] are some of the techniques used in detecting the faces in the video. Fisher vectors, Active Shape model, Active Appearance model, local binary patterns, principal component analysis [15] and Gaussian mixture model [16] are some of the methods that are used for feature extraction in facial images. Occlusions and light changes may also lead the identification technique to be misled. If the emotion is to be identified through speech, ambient noise and differences in the voices of different participants are major factors that might affect the final recognition result. According to both physiological and psychological research, humans need both audio and visual signals to correctly understand emotions for which multimodal systems that fuse audio and video signals can be used.

Thanks to recent research interest in multimodal systems, the limitations of monomodal systems [17] [18] have been overcome. The information obtained from different modalities at different levels of fusion was fused by multimodal systems. The different fusion levels are classified into two different categories, namely: matching prior to fusion and matching after fusion. Feature level and sensor level fusion techniques [19] come under the first category, and decision, rank, and score level fusion techniques come under the second category. To combine the audio and video features of the multimodal, a fusion method that takes advantage of both decision and feature-level fusion was developed. Latent space fusion methods preserve analytical or numerical correlation between the different modalities and store them in a common latent space.

II. Related Studies and Motivations

Many attempts have been made by researchers to enhance emotion identification using a combination of audio and visual information [20]. According to [21], audio-visual emotion detection may be categorized as kernel-based, feature level, model-level, decision-level, score-level, and hybrid level fusion techniques. In this paper, we focus on latent-space fusion methods and multimodal recognition to detect emotions. Multimodal emotion recognition systems consistently outperform unimodal systems [22], [23], and [24]. Although there are certain benefits to using multimodal affective systems, they also face some important challenges [24]. Selecting the modalities that result in the best combinations is the area that has been focused on in recent studies [25]. CREMA-D [26], RAVDESS [27], and SAVEE [28] are some of the existing multimodal datasets that have been considered for research in recent times. A multimodal method by Cid et al. [29] used tempo, pitch, and energy feature extraction techniques to extract the audio features and a Bayesian classification method to classify the emotions. Edge-based characteristics are obtained from visual images to classify them in the SAVEE database.

Gharavian et al. [30] evaluated the performance of a neural network called FAMNN. MFCC, Zero Crossing Rate, and pitch are some of the audio feature extraction techniques used to extract the audio features. Visual information is obtained by using marker positions on the face concept, and the resulted features are given to a feature selection algorithm (FCBF). For audio features, Huang et al. [31] used prosodic and frequency domains, while for facial expression description, they used geometry and appearance-based features. Using a back-propagation neural network, each feature vector was utilised to train a single-modal classifier. They suggested a genetic learning-based collaborative decision-making model, which was compared to concatenated equal weighted choice fusion, BPN learning-based weighted decision fusion, and feature fusion methods. The audio spectrum features are obtained from BERT and CNN and are combined in parallel to form a multimodal [32].

A HGFM method was proposed by Xu [33], which fuses the hand-crafted features and the features extracted from the gated recurrent unit. The key frame videos are summarized by the method proposed by Noroozi [34] which uses a CNN model and the concept of stack fashion or late fusion for detecting the emotions. Xu et al. [35] proposed a multi-hop memorized network that describes the single-modality and cross-modality interactions among the three different feature domains in aspect-level sentimental analysis of a multimodal system. Zadeh et al. [36] introduced a tensor fusion network that uses the product of audio, visual and image elements to represent multimodal fusion information.

RMFN, a multistage recurrent network for fusion described by Liang et al. [37], divides the multimodal fusion into various stages that utilize LSTM to record multimodal interactions in both synchronous and asynchronous modes. Liu et al. [38] lowered the computational complexity of the parameters by using a low-rank multimodal fusion approach that employs a low-rank tensor to relieve the increased computational cost of considering all three modalities. Poria et al. [39] used LSTM to isolate audio, video and text elements before combining them in a multi-level architecture. Ghosal et al. [40] developed a multi-attention recurrent network architecture for multimodal representation that learns features through attention. Tsai et al. [41] suggested learning interactions between modalities by employing multimodal transformers to construct an attention-based cross-modal architecture.

By using the RAVDESS dataset Fu Z et al. [47], R. Chatterjee et al. [48], Chang X et al. [49], Wang W et al. [50] achieved test accuracies of 75.76, 90.48, 91.4, and 89.8 on their respective multimodal systems. Ghaleb E et al. [52], He G et al. [53] proposed multimodal systems which resulted in test accuracies of 66.5 and 64 on the CREMAD dataset. Rory Beard et al. [51] proposed a multimodal where CREMAD and RAVDSR datasets are used for experimentation and resulted in test accuracies of 65.0 and 58.3, respectively.

III. Research Method

A. Dataset Description

CREMAD and RAVDESS datasets are used for experimentation and evaluation purposes. Both datasets consist of data related to the emotions of actors in both audio and video modes. Angry, disgust, fear, happy, neutral, and sad are the common emotions present in both datasets in both modes, whereas RAVDESS audio data consists of two more emotions, calm, and surprise. CREMAD consists of 22326 and 60359 emotions related to audio and video. RAVDESS consists of 4321 and 45225 emotions related to audio and video. A detailed overview of the datasets is given in Table I below.

B. Image Feature Extraction

From the given set of video sequences of the multimodal dataset the videos should be converted into images and then facial features should be extracted from the images. The detailed description of the features is extracted from the videos is given below.

From the given set of facial emotion videos f_{vid} of a multimodal dataset, the images are represented in linear scale space L_{ss} which is obtained by convoluting f_{vid} with 3 dimensional Gaussian Kernel.

TABLE I. DESCRIPTION OF CREMAD AND RAVDESS DATASETS

Name of The Dataset	Emotion Type	Data mode and Number of Emotions	
		Audio Mode	*Video/Image Mode*
CREMAD Dataset	Angry	3510	10472
	Disgust	4116	10098
	Fear	3918	10626
	Happy	3709	9661
	Neutral	3666	10867
	Sad	3417	8635
RAVDESS Dataset	Angry	476	7603
	Calm	524	NA
	Disgust	628	7885
	Fear	542	7394
	Happy	610	7784
	Neutral	385	7419
	Sad	559	7140
	Surprise	596	NA

$$L_{ss}(.;\ \sigma^2_{L_{ss}}, \tau^2_{L_{ss}}) = Gau_k(.;\sigma^2_{L_{ss}}, \tau^2_{L_{ss}}) * f_{vid}(.) \tag{1}$$

linear scale space, f_{vid} is video sequence, $\sigma^2_{L_{ss}}$is Spatial variance, $\tau^2_{L_{ss}}$ is Temporal variance, Gau_k is Spatial Temporal Gaussian Kernel.

$$Gau_k(x, y, t_d : \sigma^2_{Lss}, \tau^2_{Lss}) = \exp(-(x^2 + y^2)/2\sigma^2_{Lss} - t^2_d \,/\, 2\,\tau^2_{Lss}) \tag{2}$$

Whereas x and y represents the axis of the frames that are obtained from the facial input video sequence f_{vid}, t_d denotes the axis if time in the temporal domain

A method proposed by Forstner and Harris [42] [43] considers a Gaussian window to identify distinct points of the image which in turn determines the locations in f_{vid} when there are significant changes in the intensity of image in the given space and time domains when sliding the Gaussian window in various directions. The distinct points can be detected by convoluting Spatial-Temporal Second Momentum matrix with the given Gaussian weighted function $Gau_k(.;\sigma^2_i, \tau^2_i)$.

The Spatial-Temporal Second Momentum matrix is 3×3 dimensional matrix and is given as

$$\begin{bmatrix} L^2_{ssx} & L_{ssx}L_{ssy} & L_{ssx}L_{sst} \\ L_{ssx}L_{ssy} & L^2_{ssy} & L_{ssy}L_{sst} \\ L_{ssx}L_{sst} & L_{ssy}L_{sst} & L^2_{ssz} \end{bmatrix} \tag{3}$$

And the distinct points identification is given by

$$\mu_{ch} = Gau_k(.,\sigma^2_i, \tau^2_i) * \left(\begin{bmatrix} L^2_{ssx} & L_{ssx}L_{ssy} & L_{ssx}L_{ssz} \\ L_{ssx}L_{ssy} & L^2_{ssy} & L_{ssy}L_{ssz} \\ L_{ssx}L_{ssz} & L_{ssy}L_{ssz} & L^2_{ssz} \end{bmatrix}\right) \tag{4}$$

Where L_{ssx}, L_{ssy} & L_{sst} are first order derivatives that are defined as follows

$$L_{ssx}(.,\sigma^2_{lss}, \tau^2_{lss}) = \partial_x(Gau_k * f_{vid}) \tag{5}$$

$$L_{ssy}(.,\sigma^2_{lss}, \tau^2_{lss}) = \partial_y(Gau_k * f_{vid}) \tag{6}$$

$$L_{ssz}(.,\sigma^2_{lss}, \tau^2_{lss}) = \partial_z(Gau_k * f_{vid}) \tag{7}$$

Where $\sigma^2_i = S_{ssk} * \sigma^2_{lss}$, $\tau^2_i = S_{ssk} * \tau^2_{lss}$ and S_{ssk} is a constant

The existence of distinct points in the f_{vid} is indicated by the eigen values $\lambda_1, \lambda_2, \lambda_3$ that can hold larger values. In the Spatial-Temporal domain the variations that are existing in the intensity of image are obtained by concatenating the $trace_{lss}$ and determinant of μ_{ch} which is given as

$$H_{fn} = |(\mu_{ch})| - K * trace^3_{lss}(\mu_{ch})$$

$$= \lambda_1 * \lambda_2 * \lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3) \tag{8}$$

K is a constant and the function $H_{fn}$is normalized such that the effect of variations in the images due to illumination can be removed

C. Audio Feature Extraction

Zero crossing rate (ZCR), Mel Frequency Spectrum Coefficient (MFCC), pitch and energy are some of the feature extraction techniques used to extract the features of the emotions from the given audio signal.

1) Zero crossing rate: The number of times the audio signal crosses the zero-line, x-axis, is referred to as the zero-crossing rate, and it is stated as follows.

$$Z_{t_n} = \frac{1}{2N} * \sum_{n=1}^{N} \left| \begin{array}{l} Sign_{Aud}(x_{Audt}(n)) - \\ Sign_{Aud}(x_{Audt}(n-1)) \end{array} \right| \tag{9}$$

$$Sign_{Aud}(x_{Audt}) = \begin{Bmatrix} 1 & \text{if } x_{Audt} > 0 \\ 0 & \text{Otherwise} \end{Bmatrix} \tag{10}$$

Where, $t_n \in [t_{n1}, t_{n2}]$, $x_{Audt(t_n)}$ is the respective audio signal that was divided into segments by using a sliding window that was having al length of T, $n \in [0, N]$ and $x_{Audt}(n)$ is the t_n^{th} Segments time sequence

2) MFCC (Mel frequency cestrum coefficient): The coeeficients of the corresponding spectral form of the audio stream are represented using a nonlinear Mel scale. The Mel frequency was used to analyse cepstral coefficients, and the steps below were followed.

$Step\ 1$: $Audio\ Signals\ are\ splitted\ into\ frames\ by\ using\ fixed\ shift\ and\ window\ sizes.$

$Step\ 2$: $Fast\ Fourier\ Transform\ (FFT)\ for\ each\ frame\ is\ calcullated.$

$Step\ 3$: $Frequencies\ are\ based\ on\ the\ Mel\ Scale\ used.$

$Step\ 4$: $Logarithm\ of\ the\ resulted\ output\ of\ Step\ 3\ is\ calcullated.$

$Step\ 5$: $Discrete\ Cosine\ Transform\ (DCT)\ for\ each\ frame\ is\ calcullated.$

Acoustic tube characteristics pitch and energy are exhibited by MFCC that contains great amount of emotional information which plays a key role in emotion recognition.

3) Pitch: It depicts the signal's fundamental frequency [44]. The valence of an audio stream is connected to its rhythm and average pitch from an emotional standpoint. For example, higher amount of pitch may be associated to discomfort, lower standard deviation to sadness and usually happiness and discomfort are having higher talk and pitch rates whereas sadness can be represented by lower talk and pitch rates [45]. Autocorrelation is used to calculate the pitch of the audio signal and is given as follows.

$x_{Aud}[n]$ be Stochastic Process Sinusoidal function

$x_{Aud}[n] = \mathrm{Cos}(w_0 n + \emptyset)$ and the autocorrelation of $x_{Aud}[n]$ is given as

$$R_{Aud}[t] = E\{ x^*_{Aud}[n] * x^*_{Aud}[n+t] \} \tag{11}$$

$$= \frac{1}{2}\cos(w_0 t)$$

Maximum of the autocorrelation value is used to calculate the pitch, S_{Aud} Samples are used to calculate the estimate of $R_{Aud}[t]$

$$\hat{R}_{Aud}[t] = \frac{1}{S_{Aud}} * \sum_{S_{Aud=0}}^{S_{Aud}-|t|}(W_{Aud}[S_{Aud}] * xS_{Aud} * W_{Aud}[S_{Aud} + |t|]) \tag{12}$$

$W_{Aud}[S_{Aud}]$ is window length of S_{Aud} the Expected value of $\hat{R}_{Aud}[t]$ is given as

$$E_{Aud}\{\hat{R}_{Aud}[t]\} = \left(1 - \frac{|t|}{S_{Aud}}\right) * \frac{\mathrm{Cos}(w_{Audo} * S_{Aud})}{2}, |t| < S_{Aud} \tag{13}$$

4) Energy: It represents the signal's intensity or total energy. From an emotional standpoint, an audio signal having exciting emotions (e.g., pain or happiness) has more energy than an audio signal containing sadness or fatigued feelings [46]. The energy of the audio signal $x_{Audt}(n)$ is given as

$$\mathrm{Energy}_{Aud} = \sqrt{\frac{1}{N} * \sum_{n=1}^{N}\left(x_{Audt}(n^2)\right)} \tag{14}$$

D. Feature Level Fusion

From the features obtained from audio and video signals, only a few portions of the features are related to emotions. Personality, age, gender, and many other features are obtained from audio and video signals, which may impact the quality of recognition of the emotions that are used in the model for training. Feature Level Latent Space methods are one of the existing categories of methods that are used to find the common features related to emotions and map them into the required latent space. By maximizing the cross correlation of the respective features and by minimizing the feature distance or by taking the normalization of the features, they can be used in feature level fusion. Marginal Fisher Analysis (MFA) is a supervised method that is used for audio video feature level for fusion by extracting the required features from the respective modalities. The process of MFA^s feature level fusion is given as below.

Information related to class labels is used in latent space generation. The compactness in the intra class is given as

$$S_{compact} = \sum_{i} \sum_{i \in N^+_{k1(j)}} \left\| W^T_{AV} x_i - W^T_{AV} x_i \right\|^2$$

$$= 2\, W^T_{AV} X_{AV} (D^{AV} - S^{AV}) * X^T_{AV} w_{AV} \tag{15}$$

$X_{AV} = \{x_1, x_2, \dots, x_n\}$ is the frame set, N is the total samples and N^+_{k1} is k_1 in the same class.

$$S^{AV}_{ij} = \begin{cases} 1 & \text{if } i \in N^+_{k1}(j) \\ 0 & \text{Otherwise} \end{cases} \tag{16}$$

$$D^{AV}_{ij} = \sum_j S^{AV}_{ij} \tag{17}$$

And the Inter-Class Separability is given by

$$Icp_P = \sum_i \sum_{(i,j) \in P_{k2(c_i)}} \left\| W^T_{AV} x_i - W^T_{AV} x_j \right\|^2 \tag{18}$$

$$= 2\, W^T_{AV} X_{AV} \left(D^P_{AV} - S^P_{AV}\right) * W^T_{AV} W_{AV}$$

$c_i$is the emotion of class i, $P_{k2}(c_i)$ is the set of K_2 nearest pairs and S_{AV} is given by

$$S^P_{AV_{ij}} = \begin{cases} 1 & \text{if}(i,j) \in P_{k2}(c_i) \\ 0 & \text{Otherwise} \end{cases} \tag{19}$$

And the objective function is given as follows

$$\hat{W}_{AV} = \arg_{W_{AV}} \left\{ \min \left\{ \frac{W^T_{AV} X_{AV} (D^{AV} - S^{AV}) X^T_{AV} W_{AV}}{W^T_{AV} X_{AV} (D^P_{AV} - S^P_{AV}) X^T_{AV} W_{AV}} \right\} \right\} \tag{20}$$

And the optimal solution is given by

$$Y_{AV} = X^T_{Av} W_{AV} \tag{21}$$

$$L_{AV} \,.\, Y_{AV} = \lambda\, L^P_{AV}\,. \tag{22}$$

Where $L_{AV} = D^{AV} - S^{AV}$ and $L^P_{AV} = D^P_{AV} - S^P_{AV}$ are called Laplacian matrices for W_{AV} and W^P_{AV}

E. Proposed CNN Architecture

The proposed CNN architecture consists of four fully connected layers, one flattening layer and two dense layers. All the fully connected layers are interconnected with each other where the output features obtained from each fully connected layer are given as an input to the next fully connected layer. The inputs to the first fully connected layer are audio, video, and multimodal features that are obtained during pre-processing by applying the audio feature, image feature, and feature level fusion extraction techniques described in the above sections. The first fully connected layer consists of convolution and max polling layers, and the representation of the first fully connected layer is given as

$$\mathrm{Out}_{conv1} = \mathrm{Act}\left(\sum_i L_{AV} * W^n_{i,j}\right) \tag{23}$$

Where Out_{conv1} is the output of the convolutional layer, Act is the activation function, L_{AV} is the latent space or latent features obtained after applying feature level fusion, and $W^n_{i,j}$ is the set of weights associated with the convolutional layer

$$\mathrm{Out}_{convf1} = \text{Max polling}\{\mathrm{Out}_{conv1}\} \tag{24}$$

$\mathrm{Out}_{\mathrm{convf1}}$ is the output obtained from the max polling layer, where the input is $\mathrm{Out}_{\mathrm{conv1}}$, the first convolutional layer output. The output of the first fully connected layer $\mathrm{Out}_{\mathrm{Maxpoll1}}$ is given as input to the second fully connected layer, which consists of convolutional, max polling, and dropout layers, and the representation of the second fully connected layer is given as

$$\mathrm{Out}_{\mathrm{conv2}} = \mathrm{Act}\left(\sum_i \mathrm{Out}_{\mathrm{Maxpoll1}} * W_{i,j}^{2n}\right) \tag{25}$$

$$\mathrm{Out}_{\mathrm{Maxpoll2}} = \mathrm{Max\ polling}\{\mathrm{Out}_{\mathrm{conv2}}\} \tag{26}$$

$$\mathrm{Out}_{\mathrm{conv2f}} = \mathrm{Act}\left(\left(\mathrm{Out}_{\mathrm{Maxpoll2}} .* \mathrm{Drop}(0.2)\right)\right) * W^{[2n+1]} \tag{27}$$

$\mathrm{Out}_{\mathrm{conv2f}}$ is the output of the second fully connected layer, $\mathrm{Drop}(0.2)$ means that 20% of the features are dropped from the output of the max polling layer, and $W^{[2n+1]}$ are associated weights used.

$\mathrm{Out}_{\mathrm{conv2f}}$ the output of second fully connected layer, is given as input to the third fully connected layer which consists of the same layers as second fully connected layer and the output of the third fully connected layer is given as

$$\mathrm{Out}_{\mathrm{conv3f}} = \mathrm{Act}\left(\left(\mathrm{Out}_{\mathrm{Maxpoll3}} .* \mathrm{Drop}(0.2)\right)\right) * W^{[2n+2]} \tag{28}$$

$\mathrm{Out}_{\mathrm{conv3f}}$ is given as input to the fourth fully connected layer which consists of a convolution and max polling layers and the output is given as

$$\mathrm{Out}_{\mathrm{conv4}} = \mathrm{Act}\left(\sum_i \mathrm{Out}_{\mathrm{conv3f}} * W_{i,j}^{n}\right) \tag{29}$$

$$\mathrm{Out}_{\mathrm{conv4f}} = \mathrm{Max\ polling}\{\mathrm{Out}_{\mathrm{conv4}}\} \tag{30}$$

The output of the fourth fully connected layer is flattened by giving to a flatten layer and the output is represented as

$$\mathrm{Flatten}_{\mathrm{CNN}} = \mathrm{Flatten}(a_1\mathrm{Out}_{\mathrm{convf1}}, a_2\mathrm{Out}_{\mathrm{convf2}}, a_3\mathrm{Out}_{\mathrm{convf3}}, a_4\mathrm{Out}_{\mathrm{convf4}}) \tag{31}$$

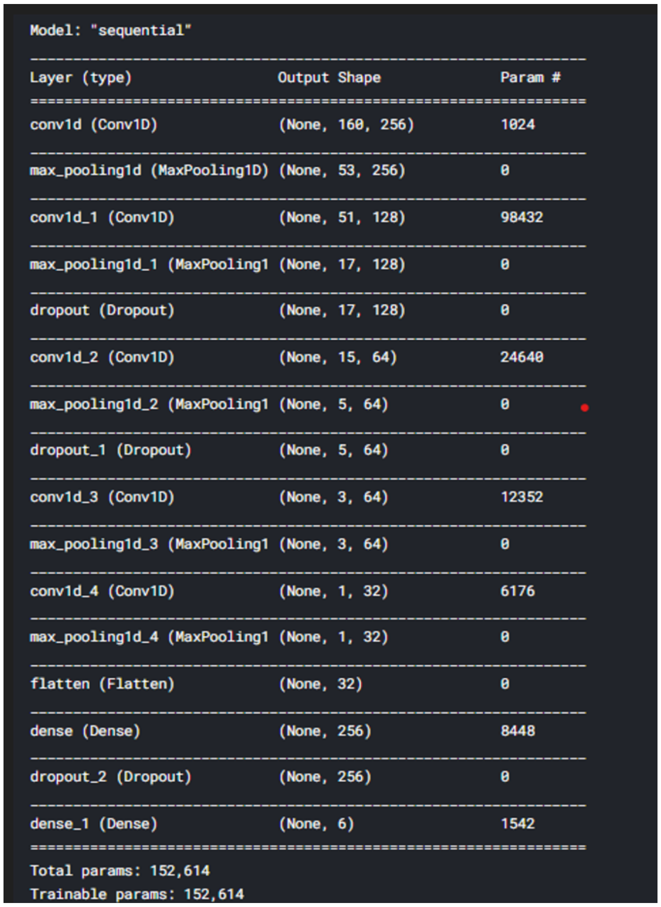
The output of a flattening layer is given to a dense layer and a dropout of 20% is applied to the output obtained from the dense layer. The resultant features are given as input to the next dense layer where the output is classified. Relu activation function is used in the dense layers that are used in between, and a SoftMax activation function is used in the final dense output layer. The representation of the dense, dropout, and final output layers is as follows:

$$\mathrm{Out}_{\mathrm{Dense1}}^{l} = \mathrm{Dense}\left(\mathrm{Den}_{\mathrm{N}}, \mathrm{Act}_{\mathrm{Relu}}(\mathrm{Flatten}_{\mathrm{CNN}})\right) \tag{32}$$

$$\mathrm{Out}_{\mathrm{Drop}}^{l} = \mathrm{Act}\left(\mathrm{Out}^{l}{}_{\mathrm{Dense1}} .* \mathrm{Drop}(0.2)\right) * W \tag{33}$$

$$\mathrm{Out}_{\mathrm{F}}^{l} = \mathrm{Dense}\left(\mathrm{Den}_{\mathrm{C}}, \mathrm{Act}_{\mathrm{Softmax}}\left(\mathrm{Out}_{\mathrm{Drop}}^{l}\right)\right) \tag{34}$$

$\mathrm{Out}_{\mathrm{Dense1}}^{l}$ is the output of the dense layer, $\mathrm{Out}_{\mathrm{Drop}}^{l}$ is the output of dropout layer $\mathrm{Out}_{\mathrm{F}}^{l}$ is the final classified output. The architecture of the proposed CNN is given in the Fig. 1.



```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d (Conv1D)              (None, 160, 256)          1024
_________________________________________________________________
max_pooling1d (MaxPooling1D) (None, 53, 256)           0
_________________________________________________________________
conv1d_1 (Conv1D)            (None, 51, 128)           98432
_________________________________________________________________
max_pooling1d_1 (MaxPooling1 (None, 17, 128)           0
_________________________________________________________________
dropout (Dropout)            (None, 17, 128)           0
_________________________________________________________________
conv1d_2 (Conv1D)            (None, 15, 64)            24640
_________________________________________________________________
max_pooling1d_2 (MaxPooling1 (None, 5, 64)             0
_________________________________________________________________
dropout_1 (Dropout)          (None, 5, 64)             0
_________________________________________________________________
conv1d_3 (Conv1D)            (None, 3, 64)             12352
_________________________________________________________________
max_pooling1d_3 (MaxPooling1 (None, 3, 64)             0
_________________________________________________________________
conv1d_4 (Conv1D)            (None, 1, 32)             6176
_________________________________________________________________
max_pooling1d_4 (MaxPooling1 (None, 1, 32)             0
_________________________________________________________________
flatten (Flatten)            (None, 32)                0
_________________________________________________________________
dense (Dense)                (None, 256)               8448
_________________________________________________________________
dropout_2 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_1 (Dense)              (None, 6)                 1542
=================================================================
Total params: 152,614
Trainable params: 152,614
```

Fig. 1. Proposed CNN Architecture.

F. Data Preprocessing

For experimentation, the RAVDESS and CREMAD datasets are used in this paper. The datasets contain data related to audio and video emotions of various actors, and the description of the data is given in the dataset description section of the same module. The features of the video and audio data are obtained by using the image feature extraction and audio feature extraction methods explained above. There is dissimilarity in the number of features obtained from audio and video datasets. There are more features in the resultant dataset of video images when compared to audio files. A dimensionality reduction technique is applied to the image set to reduce the number of features so that the same number of features is present in the audio and video resultant datasets. Finally, a multimodal dataset is obtained by combining the resultant features of audio and video from the respective datasets by using the feature-level fusion technique that was explained in Section D, namely the "Feature Level Fusion" of the same module. The features obtained after applying Feature Level Fusion are given to the proposed CNN Model for Evaluation, and the description of the proposed CNN Model is explained in Section E, named "Proposed CNN Architecture." Fig. 2 gives the workflow of the proposed work done in this paper.

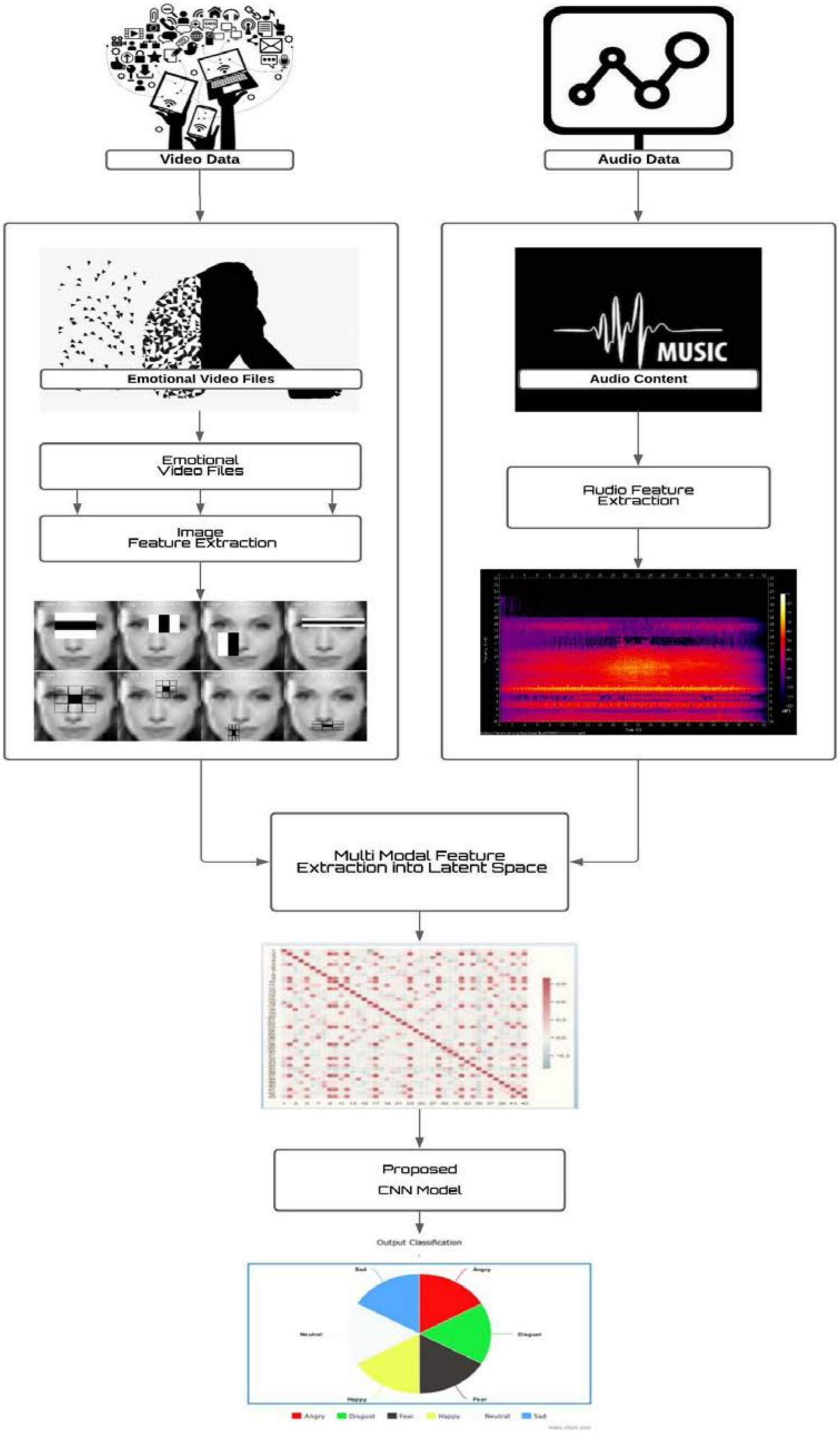
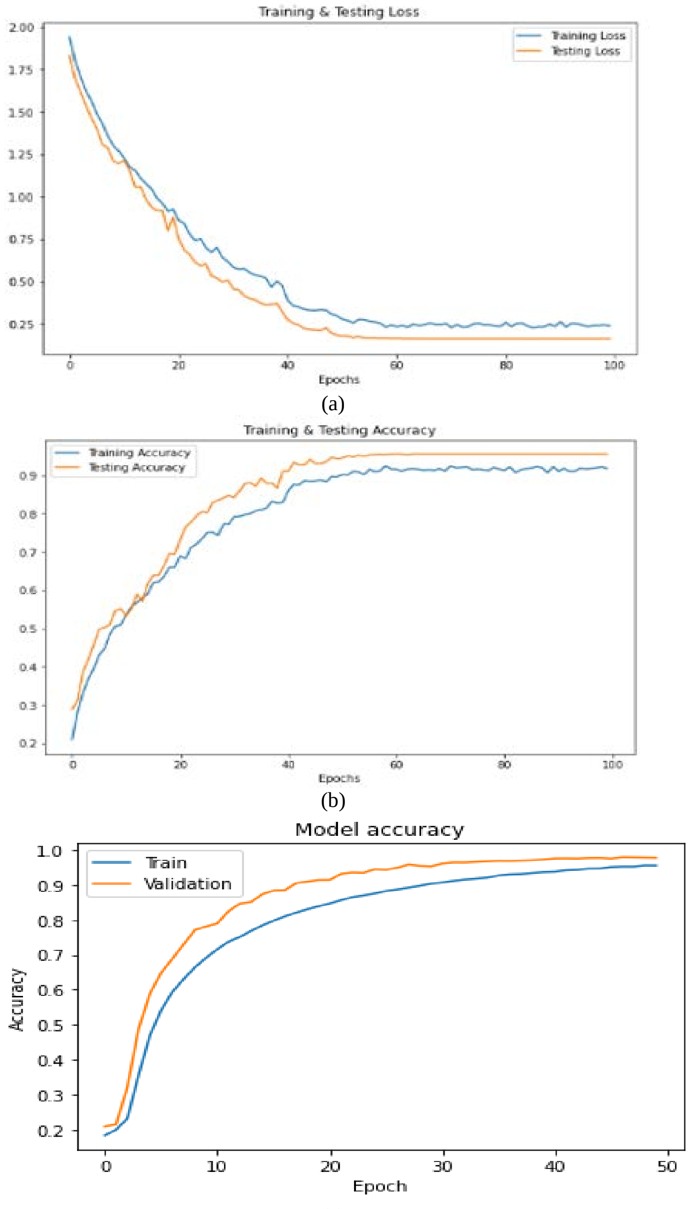


Fig. 2. Workflow of the Proposed Method.

IV. Experimentation and Results

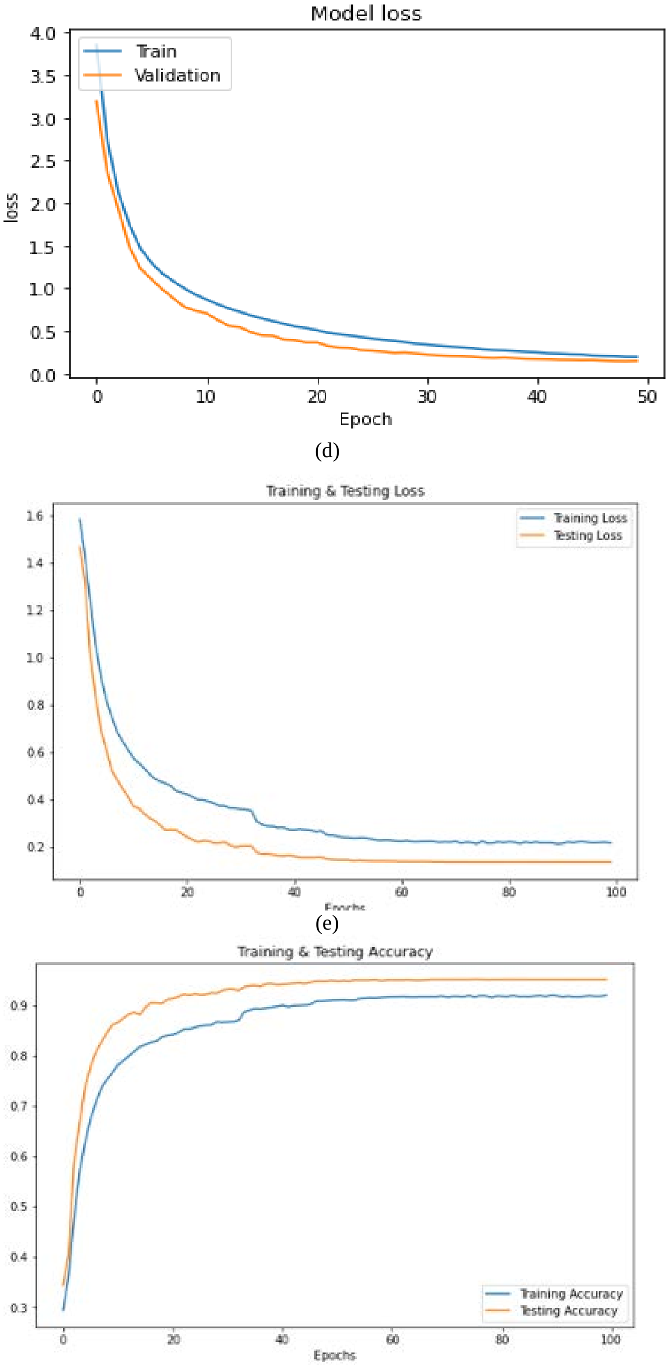
Fig. 3(a) and (b), 3(c) and (d) and 3(e) and (f) represent training and testing accuracy and loss comparisons in audio, video, and multimodal modes on the RAVDESS dataset. Test accuracies of 95.96, 96.28, and 95.07 were observed. On the CREMAD dataset, train and test accuracies and train and test accuracies losses are shown in Fig. 4(a) and (b), 4(c) and (d), and 4(e) and (f) represent training and testing accuracy and loss comparisons in audio, video, and multimodal modes. Test accuracies of 80.70, 97.88, and 69.66 were observed. A detailed description of the results is given in Table II.



(a)

(b)

(c)

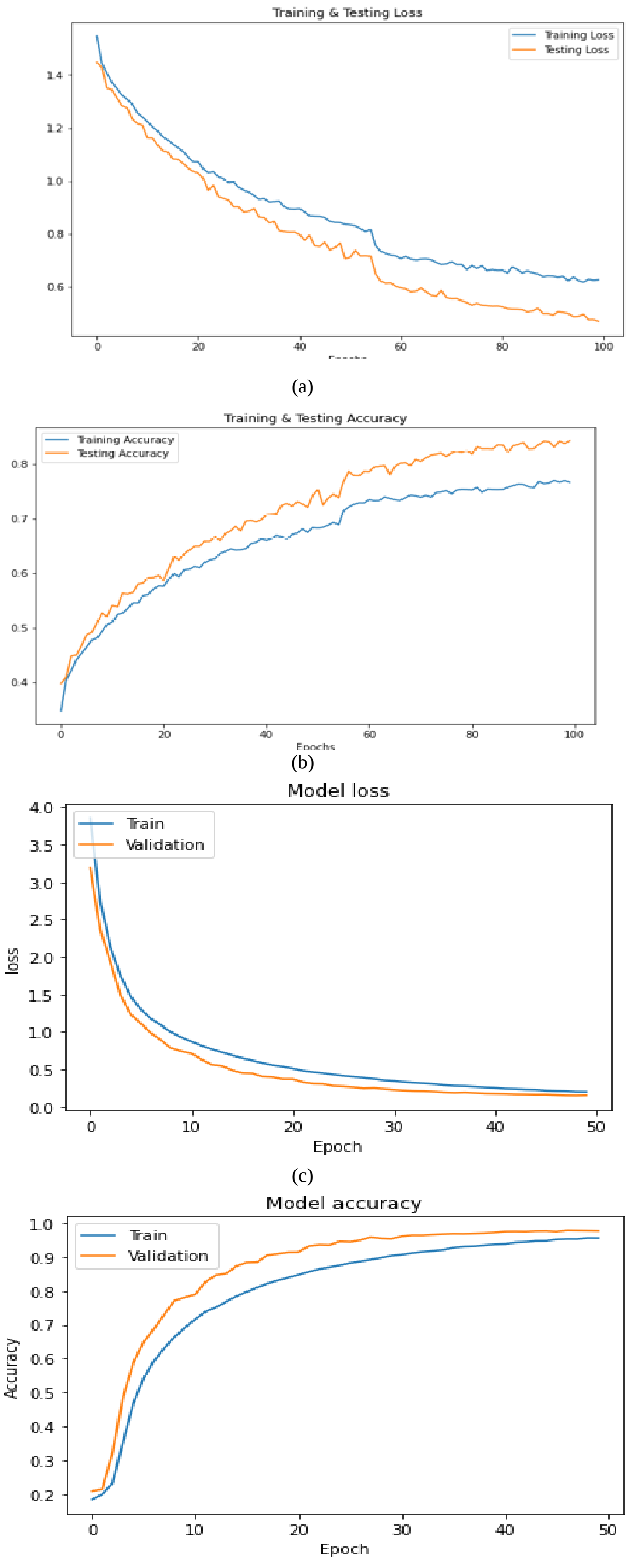


(d)

(e)

(f)

Fig. 3. (a) Training and Testing Loss of Audio Data in RAVDESS Dataset, (b) Training and Testing Accuracy of Audio Data in RAVDESS Dataset, (c) Training and Testing Accuracy of Video Data in RAVDESS Dataset, (d) Training and Testing Loss of Video Data in RAVDESS Dataset, (e) Training and Testing Loss of Multi Modal Data in RAVDESS Dataset, (f) Training and Testing Accuracy of Multi Modal Data on.

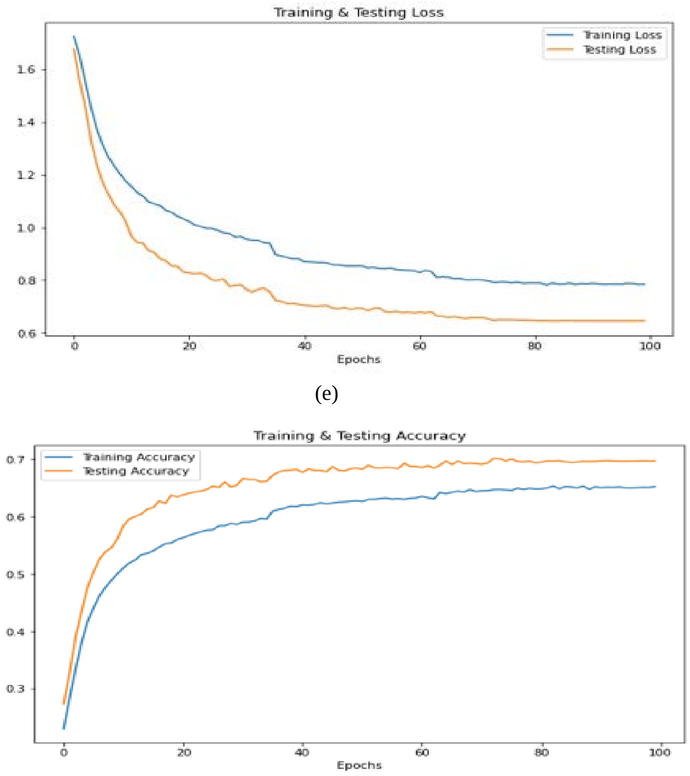


(a)

(b)

(c)

(d)



(e)

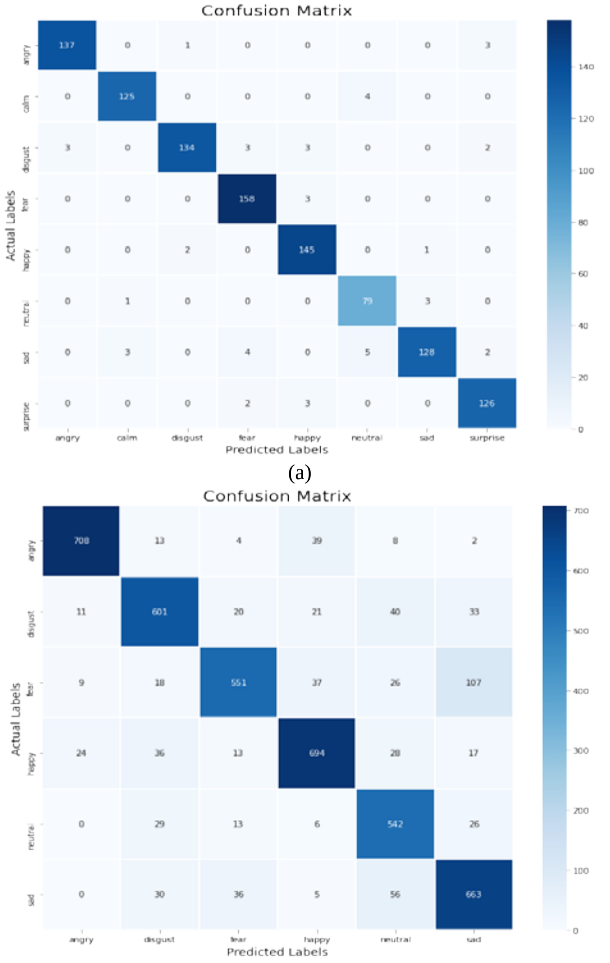
(f)

Fig. 4. (a) Training and Testing Loss of Audio Data in CREMAD Dataset, (b) Training and Testing Accuracy of Audio Data in CREMAD Dataset, (c) Training and Testing Loss of Video Data in CREMAD Dataset, (d) Training and Testing Accuracy of Video Data in CREMAD Dataset, (e) Training and Testing Loss of Multi Modal Data in CREMAD Dataset, (f) Training and Testing Accuracy of Multi Modal Data in CREMAD Dataset.

TABLE II. ACCURACY AND LOSS OF RAVDESS AND CREMAD DATASETS ON DIFFERENT TYPES OF DATA IN THE DATASET

Name of Dataset	**Type of Data**	**Train Accuracy**	**Test Accuracy**	**Train Loss**	**Test Loss**
RAVDESS	Audio	92.19	95.56	0.2430	0.161
	Video	96.92	96.28	0.1716	0.173
	Multi Modal	91.95	95.07	0.2155	0.136
CREMAD	Audio	72.70	80.50	0.7293	0.588
	Video	95.33	97.88	0.1978	0.151
	Multi Modal	65.20	69.66	0.7840	0.646

Fig. 5(a) and 5(b) represent the confusion matrix/classification report of how the classes are classified during the testing phase on audio data of the RVDSR and CREMAD datasets. It has been observed that an average precision, recall, f1-score and support values of 0.955, 0.954, 0.953 and 540 on RAVDESS, 0.807, 0.806, 0.805 and 744 on CREMAD datasets, respectively. The detailed description of various emotions and their respective performance measure values of RAVDESS and CREMAD audio data is given in Table III.



Confusion Matrix

Actual Labels \ Predicted Labels	angry	calm	disgust	fear	happy	neutral	sad	surprise
angry	137	0	1	0	0	0	0	3
calm	0	125	0	0	0	4	0	0
disgust	3	0	134	3	3	0	0	2
fear	0	0	0	158	3	0	0	0
happy	0	0	2	0	145	0	1	0
neutral	0	1	0	0	0	79	3	0
sad	0	3	0	4	0	5	128	2
surprise	0	0	0	2	3	0	0	126

(a)

Confusion Matrix

Actual Labels \ Predicted Labels	angry	disgust	fear	happy	neutral	sad
angry	708	13	4	39	8	2
disgust	11	601	20	21	40	33
fear	9	18	551	37	26	107
happy	24	36	13	694	28	17
neutral	0	29	13	6	542	26
sad	0	30	36	5	56	663

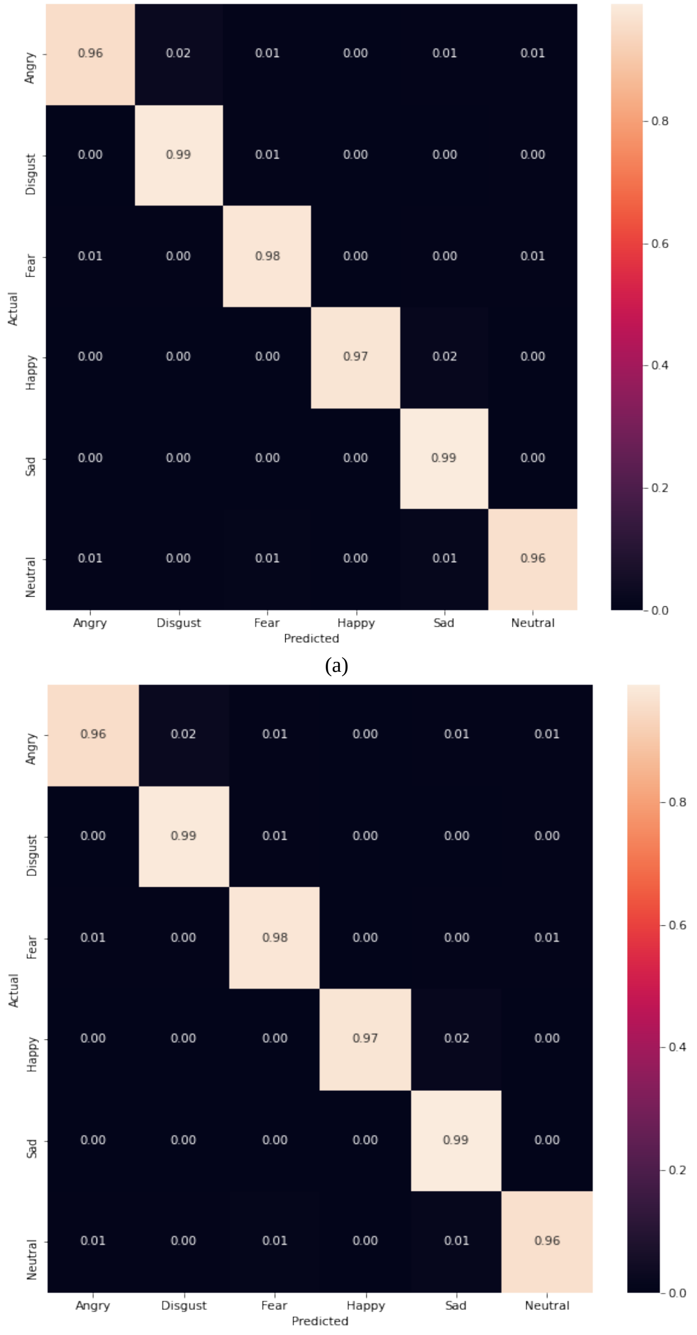
(b)

Fig. 5. (a) Confusion Matrix/Classification Report of RAVDESS Dataset Audio Data, (b) Confusion Matrix/Classification Report of CREMAD Dataset Audio Data.

TABLE III. PERFORMANCE METRICS OF RAVDESS AND CREMAD DATASETS ON AUDIO DATA

Name Of the Dataset	**Type of Emotion**	**Performance Metrics**			
		Precision	***recall***	***f1-score***	***Support***
RAVDESS	Angry	0.98	0.97	0.98	564
	Calm	0.97	0.97	0.97	516
	Disgust	0.98	0.92	0.95	580
	Fear	0.95	0.98	0.96	644
	Happy	0.94	0.98	0.96	592
	Neutral	0.90	0.95	0.92	332
	Sad	0.97	0.90	0.93	568
	Surprise	0.95	0.96	0.95	524
CREMAD	Angry	0.94	0.87	0.90	774
	Disgust	0.75	0.82	0.78	726
	Fear	0.82	0.73	0.78	748
	Happy	0.81	0.81	0.81	812
	Neutral	0.73	0.85	0.78	616
	Sad	0.79	0.76	0.78	790

Fig. 6(a) and 6(b) represent the confusion matrix/classification report of how the classes are classified during the testing phase on video data of the RVDSR and CREMAD datasets. It has been observed that an average precision, recall, f1-score and support values of 0.98, 0.985, 0.985 and 997 on RAVDESS, 0.973, 0.975, 0.975 and 1027 on CREMAD datasets, respectively. The detailed description of various emotions and their respective performance measure values of RAVDESS and CREMAD video data is given in Table IV.



Actual \ Predicted	Angry	Disgust	Fear	Happy	Sad	Neutral
Angry	0.96	0.02	0.01	0.00	0.01	0.01
Disgust	0.00	0.99	0.01	0.00	0.00	0.00
Fear	0.01	0.00	0.98	0.00	0.00	0.01
Happy	0.00	0.00	0.00	0.97	0.02	0.00
Sad	0.00	0.00	0.00	0.00	0.99	0.00
Neutral	0.01	0.00	0.01	0.00	0.01	0.96

(a)

Actual \ Predicted	Angry	Disgust	Fear	Happy	Sad	Neutral
Angry	0.96	0.02	0.01	0.00	0.01	0.01
Disgust	0.00	0.99	0.01	0.00	0.00	0.00
Fear	0.01	0.00	0.98	0.00	0.00	0.01
Happy	0.00	0.00	0.00	0.97	0.02	0.00
Sad	0.00	0.00	0.00	0.00	0.99	0.00
Neutral	0.01	0.00	0.01	0.00	0.01	0.96

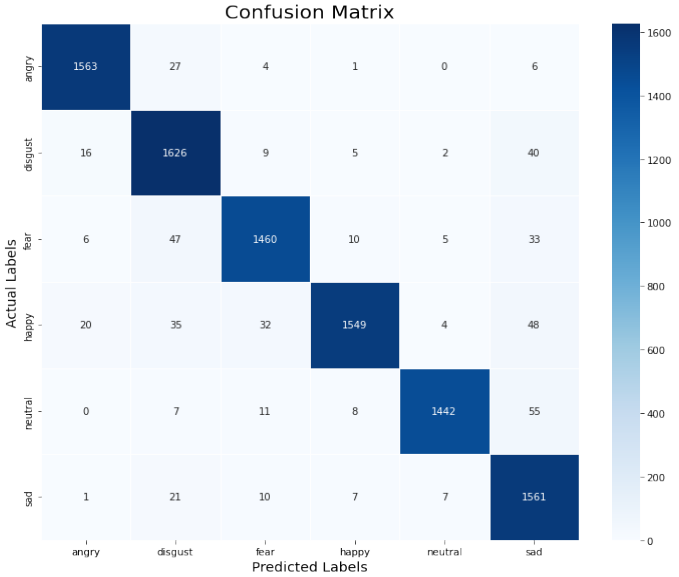
(b)

Fig. 6. (a) Confusion Matrix/Classification Report of RAVDESS Dataset Video Data, (b) Confusion Matrix/Classification Report of CREMAD Dataset Video Data.

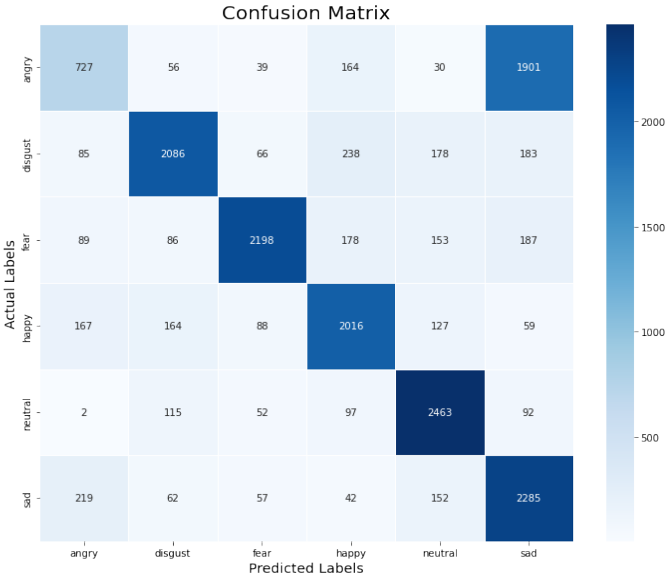
TABLE IV. PERFORMANCE METRICS OF RAVDESS AND CREMAD DATASETS ON VIDEO DATA

Name Of the Dataset	**Type of Emotion**	**Performance Metrics**			
		Precision	***recall***	***f1-score***	***Support***
RAVDESS	Angry	0.99	0.98	0.98	1028
	Disgust	0.97	0.99	0.99	1012
	Fear	0.98	0.99	0.98	1075
	Happy	0.99	0.99	0.99	940
	Neutral	0.97	0.99	0.99	1082
	Sad	0.98	0.97	0.98	842
CREMAD	Angry	0.99	0.96	0.97	1068
	Disgust	0.96	0.99	0.98	1031
	Fear	0.97	0.98	0.97	1084
	Happy	0.99	0.97	0.98	987
	Neutral	0.96	0.99	0.98	1108
	Sad	0.97	0.96	0.97	884

A multi-modal dataset has been obtained by combing the features of audio and video by using the feature level fusion techniques described in the feature level fusion section of the proposed method on the RAVDESS and CREMAD datasets. Fig. 7(a) and 7(b) give the classification report/confusion matrix obtained from the proposed CNN architecture during the evaluation stage. The classification report shows how the six classes, namely angry, disgust, fear, happy, neutral, and sad, are properly classified during their test by the proposed CNN architecture. An average precision of 0.953 & 0.716, recall of 0.953 & 0.7, f1-support of 0.951 & 0.688, and support of 1613 & 2817 were observed by the proposed CNN architecture during the evaluation phase on RAVDESS & CREMAD multimodal. The detailed description of the results is explained in Table V.



(a)



(b)

Fig. 7. (a) Confusion Matrix/Classification Report of RAVDESS Dataset on Multimodal Data, (b) Confusion Matrix/Classification Report of CREMAD Dataset on Multimodal Data.

TABLE V. PERFORMANCE METRICS OF RAVDESS AND CREMAD DATASETS ON MULTIMODAL DATA

Name of the Dataset	**Type of Emotion**	**Performance Metrics**			
		Precision	***recall***	***f1-score***	***Support***
RAVDESS	Angry	0.97	0.98	0.97	1601
	Disgust	0.92	0.96	0.94	1698
	Fear	0.96	0.94	0.95	1561
	Happy	0.98	0.92	0.95	1688
	Neutral	0.99	0.95	0.97	1523
	Sad	0.90	0.97	0.93	1607
CREMAD	Angry	0.56	0.25	0.35	2917
	Disgust	0.81	0.74	0.77	2836
	Fear	0.88	0.76	0.82	2891
	Happy	0.74	0.77	0.75	2621
	Neutral	0.79	0.87	0.83	2821
	Sad	0.49	0.81	0.61	2817

A detailed description of the macro average and weighted average accuracies Precision, recall, f1-score and support of RAVDESS and CREMAD datasets in all the three modes (Audio, video, and multimoded) are given in Table VI.

The performance of the current work done has been compared with earlier work. It has been observed that the proposed method performed better, and a detailed description of the comparisons is given in Table VII.

TABLE VI. MACRO AVERAGE AND WEIGHTED ACCURACIES OF PERFORMANCE METRICES IN DIFFERENT MODES ON RAVDESS AND CREMAD DATASETS

Name of The Dataset	**Type of Accuracy**	**Type of Data**	**Performance Metrics**			
			Precision	***recall***	***f1-score***	***Support***
RAVDESS	Macro Average Accuracy	Audio	0.95	0.96	0.95	1080
		Video	0.97	0.96	0.97	9292
		Multimodal	0.95	0.95	0.95	1578
	Weighted Average Accuracy	Audio	0.96	0.96	0.96	1080
		Video	0.97	0.97	0.97	9232
		Multimodal	0.95	0.95	0.95	1578
CREMAD	Macro Average Accuracy	Audio	0.81	0.81	0.80	4466
		Video	0.95	0.96	0.95	12642
		Multimodal	0.71	0.70	0.69	16903
	Weighted Average Accuracy	Audio	0.81	0.80	0.81	4466
		Video	0.97	0.96	0.96	12642
		Multimodal	0.71	0.70	0.69	16903

TABLE VII. ACCURACY COMPARISON OF PROPOSED METHOD WITH ALREADY EXISTING RESULTS

Name of The Author	**Datasets Used**	**Percentage of Test Accuracy**	Proposed Method **Accuracy**
Fu, Ziwang, et al. [47]	RAVDSR	75.76	95.07%
R. Chatterjee et al. [48]		90.48	
Chang X et al. [49]		91.4	
Wang W et al. [50]		89.8	
Rory Beard et al [51]		58.33	
Rory Beard et al [51]	CREMAD	65.0	69.66%
Ghaleb E et al. [52]		66.5	
He G et al. [53]		64	

V. CONCLUSION AND FUTURE WORK

A multimodal system for emotion recognition was proposed in the current work. Audio and video information are used here. Audio features are obtained by the Mel-Frequency Cepstral Coefficients extraction technique, and all the videos are converted into images and stored in a spatial-temporal space. The image features are extracted by using a Gaussian weighted function. The MFA fusion technique is to fuse the audio and video features, and the resultant features are given to the FERCNN Model for training and evaluation. For experimentation, the RAVDESS and CREMAD datasets, which consist of audio and video data, are used. Test accuracies of 95.07 and 69.66 were obtained on the RAVDSR and CREMAD datasets in multimodal mode. Even though many multimodal emotional datasets exist, only two of them are considered. An efficient multimodal system that is generic to all types of multimodal emotional databases can be designed, and the maximum multimodal data accuracy on the CREMAD dataset is 69.66%, which can be further improved.

REFERENCES

[1] E. Avots, T. Sapiński, M. Bachmann, and D. Kamińska, ‘‘Audiovisual emotion recognition in wild,’’ Mach. Vis. Appl., vol. 30, no. 5, pp. 975–985, 2019.

[2] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, ‘‘Multimodal sentiment analysis: Addressing key issues and setting up the baselines,’’ IEEE Intell. Syst., vol. 33, no. 6, pp. 17–25, Nov./Dec. 2018.

[3] N. Colneric and J. Demsar, ‘‘Emotion recognition on Twitter: Comparative ˆ study and training a unison model,’’ IEEE Trans. Affective Comput., to be published.

[4] K. P. Seng and L.-M. Ang, ‘‘Video analytics for customer emotion and satisfaction at contact centers,’’ IEEE Trans. Human-Mach. Syst., vol. 48, no. 3, pp. 266–278, May 2017.

[5] S. Tokuno, G. Tsumatori, S. Shono, E. Takei, T. Yamamoto, G. Suzuki, S. Mituyoshi, and M. Shimura, ‘‘Usage of emotion recognition in military health care,’’ in Proc. Defense Sci. Res. Conf. Expo (DSR), Aug. 2011, pp. 1–5.

[6] Wang, X., Chen, X., & Cao, C. (2020). Human emotion recognition by optimally fusing facial expression and speech feature. *Signal Processing: Image Communication*, *84*, 115831.

[7] Anbarjafari, G., Noroozi, F., Marjanovic, M., Njegus, A., & Escalera, S. (2019). Audio-Visual Emotion Recognition in Video Clips.

[8] Srinivas, P. V. V. S., & Mishra, P. (2021). Facial Expression Detection Model of Seven Expression Types Using Hybrid Feature Selection and Deep CNN. In International Conference on Intelligent and Smart Computing in Data Analytics: ISCDA 2020 (pp. 89-101). Springer Singapore.

[9] Mishra, Pragnyaban, and Srinivas, P. V. V. S. “Facial Emotion Recognition Using Deep Convolutional Neural Network and Smoothing, Mixture Filters Applied during Preprocessing Stage.” IAES International Journal of Artificial Intelligence (IJ-AI), vol. 10, no. 4, 1 Dec. 2021, pp. 889–900., https://doi.org/10.11591/ijai.v10.i4.pp889-900.

[10] P V V S Srinivas and Pragnyaban Mishra, “An Improvised Facial Emotion Recognition System using the Optimized Convolutional Neural Network Model with Dropout” International Journal of Advanced Computer Science and Applications (IJACSA), 12(7), 2021.

[11] T. Wu, S. Fu, and G. Yang, “Survey of the facial expression recognition research,” in Proc. Int. Conf. Brain Inspired Cognitive Syst., 2012, pp. 392–402.

[12] C. Ding, J. Choi, D. Tao, and L. S. Davis, “Multi-directional multilevel dual-cross patterns for robust face recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 3, pp. 518–531, Mar. 2016.

[13] Y. N. Chae, T. Han, Y.-H. Seo, and H. S. Yang, “An efficient face detection based on color-filtering and its application to smart devices,” Multimedia Tools Appl., vol. 75, pp. 1–20, 2016.

[14] K. Anderson and P. W. McOwan, “A real-time automated system for the recognition of human facial expressions,” IEEE Trans. Syst. Man Cybern. Part B (Cybern.), vol. 36, no. 1, pp. 96–105, Feb. 2006.

[15] Fuad, M., Hasan, T., Fime, A. A., Sikder, D., Iftee, M., Raihan, A., ... & Islam, M. (2021). Recent Advances in Deep Learning Techniques for Face Recognition. *arXiv preprint arXiv:2103.10492*.

[16] Kamarol, S. K. A., Jaward, M. H., Parkkinen, J., & Parthiban, R. (2016). Spatiotemporal feature extraction for facial expression recognition. *IET Image Processing*, *10*(7), 534-541.

[17] J. Yan, W. Zheng, Q. Xu, G. Lu, H. Li, and B. Wang, ‘‘Sparse kernel reduced-rank regression for bimodal emotion recognition from facial expression and speech,’’ IEEE Trans. Multimedia, vol. 18, no. 7, pp. 1319–1329, Jul. 2016. [9].

[18] S. Nemati and A. R. Naghsh-Nilchi, ‘‘Exploiting evidential theory in the fusion of textual, audio, and visual modalities for affective music video

retrieval,'' in Proc. 3rd Int. Conf. Pattern Recognit. Image Anal. (IPRIA), Apr. 2017, pp. 222–228.

[19] K. P. Seng, L.-M. Ang, and C. S. Ooi, ''A combined rule-based & machine learning audio-visual emotion recognition approach,'' IEEE Trans. Affective Comput., vol. 9, no. 1, pp. 3–13, Jan./Mar. 2018.

[20] S. E. Kahou, et al., "EmoNets: Multimodal deep learning approaches for emotion recognition in video," J. Multimodal User Interfaces, vol. 10, pp. 1–13, 2015. [26].

[21] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel crossmodal factor analysis for information fusion with application to bimodal emotion recognition," IEEE Trans. Multimedia, vol. 14, no. 3, pp. 597–607, Jun. 2012.

[22] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, ''A review of affective computing: From unimodal analysis to multimodal fusion,'' Inf. Fusion, vol. 37, pp. 98–125, Sep. 2017.

[23] E. Cambria, ''Affective computing and sentiment analysis,'' IEEE Intell. Syst., vol. 31, no. 2, pp. 102–107, Mar./Apr. 2016.

[24] S. K. D'Mello and J. Kory, ''A review and meta-analysis of multimodal affect detection systems,'' ACM Comput. Surv., vol. 47, no. 3, 2015, Art. no. 43.

[25] F. A. Salim, F. Haider, O. Conlan, and S. Luz, ''An approach for exploring a video via multimodal feature extraction and user interactions,'' J. Multimodal User Interfaces, vol. 12, no. 4, pp. 285–296, 2018.

[26] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova and R. Verma, "CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset," in IEEE Transactions on Affective Computing, vol. 5, no. 4, pp. 377-390, 1 Oct.-Dec. 2014, doi: 10.1109/TAFFC.2014.2336244.

[27] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): 0196391. https://doi.org/10.1371/journal.pone.0196391.

[28] Jackson, P., & Haq, S. (2014). Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*.

[29] F. Cid, L. J. Manso, and P. Nunez, "A novel multimodal emotion recognition approach for affective human robot interaction," Proc. FinE, pp. 1–9, 2015.

[30] D. Gharavian, M. Bejani, and M. Sheikhan, "Audio-visual emotion recognition using FCBF feature selection method and particle swarm optimization for fuzzy ARTMAP neural networks," Multimedia Tools Appl., vol. 76, pp. 1–22, 2016.

[31] K.-C. Huang, H.-Y. S. Lin, J.-C. Chan, and Y.-H. Kuo, "Learning collaborative decision-making parameters for multimodal emotion recognition," in Proc. IEEE Int. Conf. Multimedia Expo, 2013, pp. 1–6

[32] S. Lee, D. K. Han, and H. Ko, "Fusion-convbert: Parallel convolution and bert fusion for speech emotion recognition," Sensors, vol. 20, no. 22, p. 6688, 2020.

[33] Y. Xu, H. Xu, and J. Zou, "Hgfm: A hierarchical grained and feature model for acoustic emotion recognition," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6499–6503.

[34] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, ''Audio-visual emotion recognition in video clips,'' IEEE Trans. Affective Comput., vol. 10, no. 1, pp. 60–75, Jan./Mar. 2019.

[35] N. Xu, W. Mao, and G. Chen, "Multi-interactive memory network for aspect based multimodal sentiment analysis," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, 2019, pp. 371–378.

[36] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," arXiv preprint arXiv:1707.07250, 2017.

[37] P. P. Liang, Z. Liu, A. Zadeh, and L.-P. Morency, "Multimodal language analysis with recurrent multistage fusion," arXiv preprint arXiv:1808.03920, 2018.

[38] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.- P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," arXiv preprint arXiv:1806.00064, 2018.

[39] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.- P. Morency, "Context-dependent sentiment analysis in user-generated videos," in Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers), 2017, pp. 873–883.

[40] D. Ghosal, M. S. Akhtar, D. Chauhan, S. Poria, A. Ekbal, and P. Bhattacharyya, "Contextual inter-modal attention for multi-modal sentiment analysis," in proceedings of the 2018 conference on empirical methods in natural language processing, 2018, pp. 3454–3466.

[41] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," arXiv preprint arXiv:1806.06176, 2018.

[42] Harris, C., Stephens, M.: 'A combined corner and edge detector'. Proc. Alvey Vision Conf., Manchester, 1988, pp. 147–152.

[43] Förstner, W., Gülch, E.: 'A fast operator for detection and precise location of distinct points, corners and centres of circular features. Proc. ISPRS Inter Commission Conf. Fast Processing of Photogrammetric Data, Interlaken, June 1987, pp. 281–305.

[44] S. Zhang, Q. Huang, S. Jiang, W. Gao, and Q. Tian, ''Affective visualization and retrieval for music video,'' IEEE Trans. Multimedia, vol. 12, no. 6, pp. 510–522, Oct. 2010.

[45] N. Fragopanagos and J. G. Taylor, ''Emotion recognition in human–computer interaction,'' Neural Netw., vol. 18, no. 4, pp. 389–405, 2015.

[46] leymani, M. Pantic, and T. Pun, ''Multimodal emotion recognition in response to videos,'' IEEE Trans. Affect. Comput., vol. 3, no. 2, pp. 211–223, Apr. 2012.

[47] Fu, Z., Liu, F., Wang, H., Qi, J., Fu, X., Zhou, A., & Li, Z. (2021). A cross-modal fusion network based on self-attention and residual structure for multimodal emotion recognition. arXiv preprint arXiv:2111.02172.

[48] R. Chatterjee, S. Mazumdar, R. S. Sherratt, R. Halder, T. Maitra, and D. Giri, "Real-time speech emotion analysis for smart home assistants," IEEE Transactions on Consumer Electronics, vol. 67, no. 1, pp. 68–76, 2021.

[49] Chang, X., & Skarbek, W. (2021). Multi-Modal Residual Perceptron Network for Audio–Video Emotion Recognition. Sensors, 21(16), 5452.

[50] Wang, W.; Tran, D.; Feiszli, M. What Makes Training Multi-Modal Classification Networks Hard? In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–18 June 2020; pp. 12692–12702.

[51] Beard, R., Das, R., Ng, R. W., Gopalakrishnan, P. K., Eerens, L., Swietojanski, P., & Miksik, O. (2018, October). Multi-modal sequence fusion via recursive attention for emotion recognition. In Proceedings of the 22nd Conference on Computational Natural Language Learning (pp. 251-259).

[52] Ghaleb, E., Popa, M., & Asteriadis, S. (2019). Metric learning-based multimodal audio-visual emotion recognition. Ieee Multimedia, 27(1), 37-48.

[53] He, G., Liu, X., Fan, F., & You, J. (2020). Image2audio: Facilitating semi-supervised audio emotion recognition with facial expression image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 912-913).