MEDWIN PUBLISHERS
Committed to Create Value for Researchers

Open Access Journal of Criminology Investigation & Justice

DV Emotion Net: An Integrated Multimodal Approach for Emotion Recognition

Dommeti D, Nallapati SRK and Alfaris R*

School of Professional Studies (SPS), Clark University, USA

***Corresponding author:** Rand Alfaris, School of Professional Studies (SPS), Clark University, Greater Boston, Massachusetts, USA, Email: ralfaris@clarku.edu

Research Article

Volume 2 Issue 1

Received Date: April 01, 2024

Published Date: May 06, 2024

DOI: 10.23880/oajcij-16000112

Abstract

This study introduces a novel approach to emotion recognition by amalgamating information from heterogeneous modalities, specifically audio and video. We employed techniques such as energy, zero crossing rate, and Mel-Frequency Cepstral Coefficients (MFCC) for audio feature extraction, which showed promising results. For video feature extraction, spatial-temporal Gaussian kernels were used to organize video frames within a linear scale space, followed by the application of a Gaussian-weighted function to the second momentum matrix for further feature extraction. The Multimodal Feature Aggregation (MFA) fusion method was employed to unify audio and video features, resulting in a comprehensive dataset. Evaluation through the Fusion of Emotion Recognition Convolutional Neural Network (FERCNN) model, supported by the "TPU VM v3-8" accelerator TPU is a Tensor Processing Unit, showcased notable performance improvements. Using the RAVDESS and CREMAD datasets, accuracies of 94.66%, 95.82%, and 94.36% in the RAVDESS dataset and 79.45%, 96.62%, and 70.14% in the CREMAD dataset for audio, video, and multimodal modalities, respectively, were achieved. These outcomes surpass the capabilities of existing multimodal systems, underscoring the efficacy of our proposed approach. Emotion recognition, particularly through multimodal means, plays a critical role in various domains, including human-computer interfaces, healthcare, legal proceedings, and entertainment. Fusing Audio and Video Modalities to Elevate Human-Computer Interaction and Intelligent System Performance is essential for enhancing communication within these domains. The proposed model is termed "DualVision EmotionNet: DV EmotionNet".

Keywords: Emotion Recognition; Multimodal System; Human-Computer Interaction; Intelligent Systems; Fusion Techniques; Emotional Context Understanding

Abbreviations: MFCC: Mel-Frequency Cepstral Coefficients; MFA: Multimodal Feature Aggregation; FERCNN: Fusion of Emotion Recognition Convolutional Neural Network; AI: Artificial Intelligence; DNN: Deep Neural Network's; DBNs: Deep Belief Networks; LSTM: Long Short-Term Memory; RTMRBM: Recurrent Temporal Multimodal Restricted Boltzmann Machine; RNNs: Recurrent Neural Network; CNNs: Convolutional Neural Network; SVM: Support Vector Machine; EEG: Electroencephalography; ZCR: Zero Crossing Rate.

Introduction

Emotion recognition has emerged as a pivotal component in today's intelligent systems, holding immense potential to transform various aspects of human-

computer interaction, healthcare, law, and entertainment. This transformative technology enables machines to comprehend and respond to human emotions, bridging the gap between artificial intelligence and human sentiment. The applications of emotion recognition extend across diverse fields and industries, shaping the way we interact with technology and enhancing human experience. In legal contexts and security measures, emotion recognition assists in identifying deception, assessing witness testimonies, and enhancing surveillance systems. Law enforcement agencies utilize it to analyse facial expressions in real-time during interrogations, aiding in assessing the credibility of witnesses or suspects' emotional states. The entertainment industry harnesses emotion recognition to create immersive and personalized experiences for audiences, adapting video games, virtual reality, and media content to the player's or viewer's emotional responses, thus enhancing engagement. Despite its evident significance, emotion recognition has evolved significantly in recent years with the development of multimodal approaches, transcending experimental stages to become deeply ingrained in our modern society. Traditional algorithms often focused on single data sources, resulting in limited understanding of the full emotional context. Multimodal systems, however, combine information from multiple sources, using fusion techniques to improve accuracy. As we explore emotion recognition, we will delve deeper into its technical aspects, methodologies, and real-world impact, offering insights into how this technology is shaping healthcare, entertainment, and human-computer interaction. The proposed model, termed "Dual Vision EmotionNet; simply, DV EmotionNet," encapsulates the concept of utilizing two different perspectives or modalities, such as audio and video, to capture a more comprehensive understanding of emotions.

Literature Review

Emotion recognition is vital for improving human-computer interaction and the efficiency of intelligent systems. Usually, this task has been approached separately using either audio or video modalities. However, technological advances in recent years have fuelled an already growing interest in creating an integrated multimodal approach which incorporates audio as well as video modalities to accomplish greater precision and reliable emotion recognition. Humans convey and view emotions in a variety of ways. The human sensory system intrinsically fuses multimodal information in a complex manner. The introduction of artificial intelligence (AI) methods, specifically deep learning techniques, facilitated the recent rapid technological development in the human-computer interaction domain. Many AI-based systems can detect users' affective states automatically, resulting in a personalized experience in terms of interaction between humans and computers [1]. Case studies of these methodologies and approaches are discussed with the examples 'Social Robots' by Cavallo F, et al. [2] and 'Emotion monitoring systems for race car drivers' by Katsis CD, et al. [3] AI-based Emotion detection systems supplement traditional information in a wide range of application scenarios, including healthcare, customer service, advertising, security, and identifying fraud. According to Tzirakis P, et al. [4], each individual is distinctive and may convey emotions in their own distinct manner, based on their cultural background, age, gender, or earlier life experiences [4]. According to Ji Q, et al. [5], emotional states can be used to track and predict fatigue. Emotion recognition in speech recognition can be utilized by call centers to detect the psychological condition of the caller while offering feedback on service quality [6]. It is mentioned in Tzirakis P, et al. [4] that, Deep Neural Network's (DNN) have significantly improved the efficiency of pattern recognition models. Recently, a number of new neural network architectures, such as autoencoder networks [7], Convolutional Neural Networks [8], Deep Belief Networks (DBNs) [9], and memory enhanced models of neural networks such as Long Short-Term Memory (LSTM) [10], have been revitalized. These Models have made the process of pattern recognition easier and more reliable than before. In their work, Hu D, et al. [11] put forward, a temporal multimodal network called Recurrent Temporal Multimodal Restricted Boltzmann Machine (RTMRBM) to model audiovisual data sequences in another study. In a study, Huang Y, et al. [12] combined DNNs and hypergraphs to propose a transductive learning framework for image-based emotion recognition. Specifically, after training the DNN for the emotion classification task, each node in the final fully connected layer was treated as a characteristic and used to form a hyperedge in a hypergraph. In their Study, Kahou S E, et al. [13] used Recurrent Neural Network (RNN's) and Convolutional Neural Network (CNN's) to recognize categorized emotions in videos. A CNN was trained to recognize emotion in static images. The characteristics extracted from the CNN were then used to train an RNN to generate an emotion to the entire video. According to Katsis CD, et al. [14], Emotions frequently involve both thinking and feeling, which are both cognitively experienced events as well as physical changes to the body. Although no technology can truly read our thoughts, there are a growing number of sensors that use AI and Deep Learning techniques to record various physical manifestations of emotion, such as footage of facial expressions and posture or gesture changes [15,16]. In their study, Kim Y, et al. [17] put forward four DBN architectures, one of which was a basic 2-layer DBN, and the others were variations on it. The basic architecture first learns the audio and video features separately, then joins these features from the two modalities and uses them to learn the second layer. A Support Vector Machine, or SVM, was used to evaluate the features. According to the authors Radoi A, et al. [1], Emotion recognition systems that use only visual

information (i.e., video frames) can be divided into static and dynamic methods based on feature representations. Static-based methods encode features with spatial data from single frames without regard for temporal extent, whereas dynamic-based methods account for the temporal relationship between continuous frames in the input sequence. For feature extraction in static-based methods, state-of- the-art deep neural network architectures (VGG [18], ResNet [19]) have been proposed, while emotion classification is carried out using a Support Vector Machine classifier [20]. Wei W, et al. [21] investigated a multimodal facial recognition approach that combined low-level facial key point features with a high-level self-learning feature. Experiments revealed that this proposed method outperformed single-modal features in face recognition, demonstrating its effectiveness. Similarly, The Multichannel Convolutional Neural Network technique was put forward and tested on the FER dataset in a similar study. The results demonstrate that the proposed MCCNN outperforms traditional CNN-based architectures. Furthermore, the efficiency of multimodal deep learning algorithms, which use multiple modalities such as text, image, audio, and video, outperforms that of single modalities (i.e., unimodal) frameworks [22]. In Guo JJ, et al. [23], classified five emotions using eye images, eye movements, and Electroencephalography (EEG) signals. According to the findings, the three modes may enhance the ability to identify the five complementary emotions. Authors Santamaria-Granados L, et al. [24] used a CNN to extract characteristics from a range of physiological signals and predict sentiment using fully connected network layers. Trials show that this method achieves higher precision in emotional state classification. Authors He G, et al. [25] describes a data augmentation technique that uses a large labelled visual data set to increase the quantity of audio-based emotion detection data. Similarly, video facial expressions can be used to increase awareness and prediction tracking of emotions in audio data, resulting in cross-modal transfer of knowledge among audio and facial modalities throughout the emotional context [26]. The openSMILE toolkit, widely used in autonomous emotion recognition from speech [27], was one of the most effective methods for audio extraction and categorization of features of speech. Similarly, spectrogram representations of emotional speech perform well in automatic speech emotion recognition [28]. Additional findings have shown that, emotional state is reflected in human biosignals; thus, emotion recognition methodologies based on the classification of extracted series of features from these biosignals are gaining popularity [29-32].

Proposed Method

The performance of the current work done is compared with the already existing work done. It has been observed that the DV EmotionNet performed better, and a detailed description of the comparisons given in Table 1.

Name of the Author	Datasets Utilized	Test Accuracy Percentage	Accuracy of DV EmotionNet
Fu Z, et al. [33]	Ravdess	75.76	94.36%
Cao H, et al. [34]		76.79	
R Chatterjee, et al. [35]		90.48	
M Xu, et al.[36]		92.49	
Chang, X. et al. [37]		91.4	
Livingstone SR, et al.[38]		93.5	
Wang W, et al. [39]		89.8	
K DONUK, et al. [40]		59.27	
Beard R, et al. [41]		58.33	
Samadiani N, et al. [42]	Cremad	68	70.14%
Ghaleb E, et al. [43]		66.5	
Beard R, et al. [41]		65	
He G, et al. [44]		64	

Table 1: Accuracy comparison of DV EmotionNet Model with other state of the art methods.

Datasets Description

Experimentation and evaluation in this study utilize the CREMAD and RAVDESS datasets, both encompassing emotional expressions of actors in both audio and video formats. Common emotions such as anger, disgust, fear, happiness, neutrality, and sadness are present in both datasets across both modalities. Notably, the RAVDESS

audio data introduces two additional emotions, calm and surprise. The CREMAD dataset comprises 22,326 and 60,359 instances of emotions related to audio and video, respectively. In comparison, the RAVDESS dataset includes 4,321 and 45,225 instances of emotions in audio and video, respectively. Further details about these datasets can be found in the Table 2 provided below. The Table 2 provides a comprehensive breakdown of emotion types present in the CREMAD and RAVDESS datasets across different modes of data, namely Audio and Video/Image. In the CREMAD dataset, six emotions- Angry, Disgust, Fear, Happy, Neutral, and Sad are explored in both audio and video modes. The corresponding numbers of instances for each emotion in the audio and video modes are detailed, revealing the dataset's diversity in emotional expressions. For instance, the angry emotion has 3,510 instances in audio and 10,472 instances in video mode. Similar patterns are observed for other emotions within CREMAD. For the RAVDESS dataset, seven emotions- Angry, Calm, Disgust, Fear, Happy, Neutral, Sad, and Surprise are considered. Notably, Calm and Surprise are unique to the RAVDESS dataset. The table highlights the distribution of emotions across audio and video modes in RAVDESS, emphasizing the varying number of instances for each emotion. This detailed presentation serves as a valuable resource for understanding the composition and diversity of emotional data in these datasets, providing essential insights for researchers and practitioners working on emotion recognition tasks.

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	Angry	476	7603
	Calm	524	NA
	Disgust	628	7885
	Fear	542	7394
	Happy	610	7784
	Neutral	385	7419
	Sad	559	7140
	Surprise	596	NA

Table 2: Description of CREMAD and RAVDESS Datasets.

Image Feature Extraction

Emotion recognition is a complex task that requires the extraction and analysis of various modalities, including image features. In this section, we delve into the process of image feature extraction and its significance in enhancing human-computer interaction and intelligent system performance. The first step in image feature extraction is to pre-process the input images. This involves resizing, normalizing, and enhancing the images to ensure consistency and improve their quality. Additionally, noise reduction techniques are applied to minimize any distortions that could affect accurate feature extraction. Once the images are pre-processed, a multitude of features can be extracted for emotion recognition. These features include but are not limited to facial landmarks, which capture key points on the face such as the eyes, nose, and mouth. Other important features involve texture analysis to identify patterns in facial expressions that correlate with specific emotions. An intriguing aspect of image feature extraction is the utilization of deep learning techniques such as CNNs. These networks are trained using massive datasets and possess remarkable capabilities in learning hierarchical representations from raw image data. By leveraging CNN models for emotion recognition tasks, we can achieve greater accuracy and robustness in identifying subtle emotional cues. In the process of analysing a multimodal dataset comprising video sequences, it is imperative to transform these video segments into a series of images, subsequently extracting relevant facial features from these images. The following is a detailed description of the procedures involved in feature extraction from the provided facial emotion videos.

Video to Image Conversion and Linear Scale Space Representation: Firstly, from the provided set of facial emotion videos denoted as $fvideo$ within the multimodal dataset, a pivotal step involves the conversion of these videos into a sequence of images. Similar to the equations proposed by Kamarol SKA, et al. [45], these images are subsequently represented within a linear scale space termed $L_{scalespace}$, which is achieved through the application of a convolution process to the original video sequence f_{video} with a three-dimensional Gaussian kernel. Mathematically, the transformation of f_{vIdeo} to $L_{scalespace}$ is expressed as following Table 3.

$$L_{scalespace}(.;\sigma^2_{L_{scalespace}}, \mathrm{T}^2_{L_{scalespace}} = Gau_k(.;\sigma^2_{L_{scalespace}}, \mathrm{T}^2_{L_{scalespace}} \times f_{Video}(.) \quad (1)$$

Mathematical Representation	Significance
$L_{scalespace}$	Linear Scale Space
f_{video}	Facial Video Sequence
$\sigma^2 L_{scalespace}$	Spatial Variance
$r^2 L_{scalespace}$	Temporal Variance
Gau_k	Spatio-Temporal Gaussian Kernel

Table 3: Description of Mathematical Representation in (1).

A. Spatio-Temporal Gaussian Kernel (Table 4)**:**

B.

$$Gau_k(x, y, t_d : \sigma^2_{L_{scalespace}}, T^2_{L_{scalespace}}) = \exp(-(x^2 + y^2)/2\sigma^2_{L_{scalespace}} - t_d^2/2T^2_{L_{scalespace}}) \quad (2)$$

Mathematical Representation	Significance
x, y	spatial axes, corresponding to the frames derived from $ffvvvdeo$
t_d	Temporal axis within the temporal domain
$\sigma^2_{Lscalespace}$	Spatial Variance, which affects the spatial distribution
$T^2_{Lscalespace}$	Temporal Variance, which influences the temporal aspects of the representation

Table 4: Description of Mathematical Representation in (2).

Through this methodology, the videos are effectively transformed into a series of images, which are then further represented in the linear scale space $L_{scalespace}$. This representation is crucial for subsequent feature extraction and analysis, forming the foundation for advanced multimodal emotion recognition systems.

A technique put forth by Harris C, et al. [36] involves the utilization of a Gaussian window for the identification of salient points within an image. These salient points, in turn, enable the determination of locations within a video sequence, denoted as f_{video}, where significant changes in image intensity occur within defined spatial and temporal domains. This identification is achieved by slightly adjusting the Gaussian window in various directions. The discernible points are subsequently detected through the convolution of the Spatial-Temporal Second Momentum matrix with a designated Gaussian weighted function denoted as $Gau_k(.; \sigma^2_i, T^2_i)$. The Spatial-Temporal Second Momentum matrix is a 3x3 dimensional matrix, which serves as a pivotal component in this method. Its structure and properties are essential for the accurate detection of significant changes in intensity over spatial and temporal dimensions. For simplicity $X^2_{Lscalespace}$ is written as X^2_{Lss} below.

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

And the distinct point's identification is given by

$$u_{ch} = Gau_k(., \sigma^2, r^2) \times \left(\begin{bmatrix} L^2_{ssx} & L_{ssx}L_{ssy} & L_{ssx}L_{sst} \\ L_{ssx}L_{ssy} & L^2_{ssy} & {}_{ssy}L_{sst} \\ L_{ssx}L_{sst} & L_{ssy}L_{sst} & L^2_{ssz} \end{bmatrix} \right) \quad (4)$$

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

$$L_{ssx}(., \sigma^2_{lss}, T^2_{lss}) = \partial_x(Gau_k \times f_{Vid}) \quad (5)$$

$$L_{ssy}(., \sigma^2_{lss}, T^2_{lss}) = \partial_y(Gau_k \times f_{Vid}) \quad (6)$$

$$L_{ssz}(., \sigma^2_{lss}, T^2_{lss}) = \partial_z(Gau_k \times f_{Vid}) \quad (7)$$

Where $\sigma^2_w = S_{ssk} \times \sigma^2_{lss}, T^2_i = S_{ssk} \times r^2_{lss}$ *and* S_{ssk} is constant.

The existence of distinct points in the f_{vid} is indicated by the eigen values λ_1, λ_2, λ_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

$$H_{ffn} = |u_{ch}| - K \times trace^3_{lss}(u_{ch}) = \lambda_1 \times \lambda_2 \times \lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3) \quad (8)$$

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.

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.

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.

$$Z_{t_n} = \frac{1}{2N} \sum_{n=1}^{N} |sign_{Aud}(x_{Audt}(n)) - sign_{Aud}(x_{Audt}(n-1))| \quad (9)$$

$$sign_{Aud}(x_{Audt}) = \begin{cases} 1 & if\ x_{Audt} > 0 \\ 0 & otherwise \end{cases} \quad (10)$$

Where, $t_n \in [t_{n1}, t_{n2}]$, $x_{Audt}(_{tn})$ 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^{th}_{n} Segments time
Sequence.

MFCC (Mel Frequency Cestrum Coefficient): The coefficients of the corresponding spectral form of the audio stream are represented using a nonlinear Mel scale. The Mel

frequency was used to analyze 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 calculated.

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

Step 4: Logarithm of the Resulted output of Step 3 is calculated.

Step 5: Discrete Cosine Transform (DCT) for each frame is calculated.

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

Pitch: It depicts the signal's fundamental frequency [47]. 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 [48]. Autocorrelation is used to calculate the pitch of the audio signal and is given as follows.

Let $x_{Aud}[n]$ be a Stochastic Process Sinusoidal function given as $x_{Aud}[n] = Cos(w_0 n + \phi)$ and the autocorrelation of $x_{Aud}[n]$ is given below

$$R_{Aud}[t] = E\left\{x^*_{Aud}[n] \times x^*_{Aud}[n+t]\right\} \quad (11)$$

$$= \frac{1}{2} cos(w_0 t) \quad (12)$$

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

$$\hat{R}_{Aud}[t] = \frac{1}{S_{Aud}} \times \sum_{s_{Aud}=0}^{s_{Aud}-|t|} (W_{Aud}[s_{Aud}] \times xS_{Aud} \times W_{Aud}[s_{Aud}+|t|]) \quad (13)$$

$W_{Aud}[S_{Aud}]$ is a window of lengh S_{Aud} and the Expected Value of $R_{Aud}[t]$ is given as

$$E_{Aud}\left\{\hat{R}_{Aud}[t]\right\} = (1 - \frac{|t|}{S_{Aud}}) \times \frac{cos(w_{Aud} \times s_{Aud})}{2}, |t| < S_{Aud} \quad (14)$$

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 [49]. The energy of the audio signal $x_{Aud}(n)$ is given as

$$Energy_{Aud} = \sqrt{\frac{1}{n} \times \sum_{n=1}^{N} (x_{Aud}(n^2))} \quad (15)$$

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 maps them into the required latent space. By maximising the cross correlation of the respective features and by minimising the feature distance or by taking the normalisation 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 intraclass is given as

$$S_{compact} = \sum_i \sum_{i \in N^+_{k1(j)}} \left\| W^T_{AV} x_i - W^T_{AV} x_i \right\|^2 \quad (16)$$

$$= 2W^T_{AV} X_{AV} (D^{AV} - S^{AV}) \times X^T_{AV} W_{AV} \quad (17)$$

$X_{AV} = \{x_1, x_2, ..., x_n\}$ are the set of frames, N is the total samples and $N_{k1}{}^+$ is k1 neighbours of x_i in the same class

$$S^{AV}_{ij} = \begin{cases} 1 & if\ i \in N^+_{k1(j)} \\ 0 & otherwise \end{cases} \quad (18)$$

$$D^{AV}_{ij} = \sum_j S^{AV}_{ij} \quad (19)$$

And the Inter-Class Separability is given by

$$ICP_p = \sum_i \sum_{(i,j) \in p_{k2(ci)}} \left\| W^T_{AV} x_i - W^T_{AV} x_i \right\|^2 \quad (20)$$

$$= 2W^T_{AV} x_{AV} (D^P_{AV} - S^P_{AV}) \times W^T_{AV} W_{AV} \quad (21)$$

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

$$S^P_{AV_{ij}} = \begin{cases} 1 & if\ (i,j) \in p_{k2}(ci) \\ 0 & otherwise \end{cases} \quad (22)$$

And the objective function is given as follows

$$W_{AV}^{\wedge} = \arg_{w_{AV}} \left\{ \min \left\{ \frac{W_{AV}^T X_{AV} (D^{AV} - S^{AV}) X_{AV}^T W_{AV}}{W_{AV}^T X_{AV} (D_{AV}^T - S_{AV}^P) X_{AV}^T W_{AV}} \right\} \right\} \quad (23)$$

And the optimal solution is given by

$$Y_{AV} = X_{AV}^T W_{AV} \quad (24)$$

$$L_{AV} \times Y_{AV} = \lambda L_{AV}^P \times Y_{AV} \quad (25)$$

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_{AV}^P

Algorithm for DV EmotionNet Model

Input: Audio Data / Video Data Output: Emotion classification Begin
Step 1: Preprocess and collect audio data
Step 2: Extract audio features
Step 3: Preprocess and collect video data
Step 4: Extract image features
Step 5: Combine audio and image features into a multimodal latent space
Step 6: Define DV EmotionNet CNN architecture:

- Initialize layers and parameters
- Define convolutional, max pooling, dropout, and dense layers
- Utilize Accelerator TPU VM v3-8 for training.

Step 7: Forward pass through the CNN model:

- Apply convolution and max pooling operations
- Flatten the output
- Apply dense layers with Rectified Linear Unit (ReLU) activation
- Apply dropout
- Final classification with softmax activation

Step 8: Backpropagation and optimization:

- Compute loss
- Update weights using backpropagation algorithm
- Utilize TPU accelerator for parallel processing.

Step 9: Repeat steps 7-8 for multiple epochs until convergence
Step 10: Evaluate the trained model on a validation set
Step 11: Test the model on unseen data
Step 12: Analyse results and interpret emotion classification performance End

DV EmotionNet Model Description in Detail

The DV EmotionNet 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 pooling layers, and the representation of the first fully connected layer is given as

$$Out_{Conv1} = Act\left(\sum_i L_{AV} \times W_{i,j}^n\right) \quad (26)$$

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.

$$Out_{Convf1} = Max\ polling\left(Out_{Conv1}\right) \quad (27)$$

$Out_{convff1}$ is the output obtained from the max pooling layer, where the input is Out_{conv1} a first convolutional layer output. The output of the first fully connected layer $Out_{Maxpoll1}$ is given as input to the second fully connected layer, which consists of convolutional, max pooling, and dropout layers, and the representation of the second fully connected layer is given as

$$Out_{Conv2} = Act\left(\sum_i Out_{Max\ poll1} \times W_{i,j}^{2n}\right) \quad (28)$$

$$Out_{Max\ poll2} = Max\ polling\left\{Out_{Conv2}\right\} \quad (29)$$

$$Out_{Conv2f} = Act\left((\mathrm{Out}_{Max\ poll2} \times Drop(0.2))\right) \times W^{[2n+1]} \quad (30)$$

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

Out_{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

$$Out_{Conv3f} = Act\left((\mathrm{Out}_{Max\ poll3} \times Drop(0.2))\right) \times W^{[2n+2]} \quad (31)$$

Out_{conv3} is given as input to the fourth fully connected layer which consists of a convolution and max pooling layers and the output is given as

$$Out_{Conv4} = Act\left(\sum_i Out_{Conv3f} \times W_{i,j}^n\right) \quad (32)$$

$$Out_{Conv4f} = Max\ polling\left\{Out_{Conv4}\right\} \quad (33)$$

The output of the fourth fully connected layer is flattened

by giving to a flatten layer and the output is represented as

$$Flatten_{CNN} = Flatten(a_1 Out_{Conyf1}, a_2 Out_{Conyf2}, a_3 Out_{Conyf3}, a_4 Out_{Conyf4}) \quad (34)$$

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. Finally, the obtained features are given as input to a dense layer where the features a classified as an output. A Relu activation function is used in the dense layers that are used in between, and a SoftMax activation function is used in the final output layer. The representation of the dense, dropout, and final output layers is as follows.

$$Out^{1}_{Dense1} = Dense(Den_N, Act_{\text{Re}lu}((Flatten_{CNN})) \quad (35)$$

$$Out^{1}_{Drop} = Act(Out^{1}_{Dense1}. \times Drop(0.2)) \times \text{W} \quad (36)$$

$$Out^{1}_{F} = Dense\left(Den_c, Act_{soft\max}(Out^{1}_{Drop})\right) \quad (37)$$

Out^{l}_{Dense1} is the output of the dense layer, Out^{l}_{Drop} is the output of dropout layer Out^{l}_{F} is the final classified output. Figure 1 gives the architecture of the DV EmotionNet CNN.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d (Conv1D)             (None, 160, 256)          1024

 max_pooling1d (MaxPooling1  (None, 53, 256)           0
 D)

 conv1d_1 (Conv1D)           (None, 51, 128)           98432

 max_pooling1d_1 (MaxPoolin  (None, 17, 128)           0
 g1D)

 dropout (Dropout)           (None, 17, 128)           0

 conv1d_2 (Conv1D)           (None, 15, 64)            24640

 max_pooling1d_2 (MaxPoolin  (None, 5, 64)             0
 g1D)

 dropout_1 (Dropout)         (None, 5, 64)             0

 conv1d_3 (Conv1D)           (None, 3, 64)             12352

 max_pooling1d_3 (MaxPoolin  (None, 3, 64)             0
 g1D)

 conv1d_4 (Conv1D)           (None, 1, 32)             6176

 max_pooling1d_4 (MaxPoolin  (None, 1, 32)             0
 g1D)

 flatten (Flatten)           (None, 32)                0

 dense (Dense)               (None, 256)               8448

 dropout_2 (Dropout)         (None, 256)               0

 dense_1 (Dense)             (None, 6)                 1542

=================================================================
Total params: 152614 (596.15 KB)
Trainable params: 152614 (596.15 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Figure 1: DV EmotionNet Architecture.

Accelerator TPU VM v3-8: The accelerator employed in this study is the "TPU VM v3-8," specifically designed for handling deep learning tasks. It finds support in Tensorflow 2.1, accessible through the Keras high-level API and, at a more granular level, within models employing a custom training loop. Within Kaggle, TPUs serve as network-connected accelerators, and the initial step involves locating them on the network, a task facilitated by TPUClusterResolver.connect(). Subsequently, a TPU-Strategy is instantiated, encompassing the essential distributed training code tailored for TPUs and their 8 compute cores. To leverage the TPU-Strategy, finalize the process by instantiating model within the strategy's scope. This action ensures that the model is established on the TPU itself. It's important to note that the size of the model is confined solely by the TPU's RAM and isn't constrained by the memory available on the VM running the Python code. Throughout model creation and training, the standard Keras APIs can be seamlessly utilized. Throughout model creation and training, the accelerator supported the model by simplifying the development process and harnessing the computational power of the TPU for accelerated deep learning task.

Data Pre-Processing

The datasets used in the proposal contain data related to audio and video as described in the dataset description section. 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 this so that the same numbers of features are 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 feature- level fusion techniques. The detailed description of how the feature level fusion is being done is explained in the section Feature-Level Fusion. The resultant multimodal datasets are given for the CNN model for evaluation. The description of the model is given in the DV EmotionNet Model Description section. Figure 2 gives the workflow of the proposed work done in this paper.

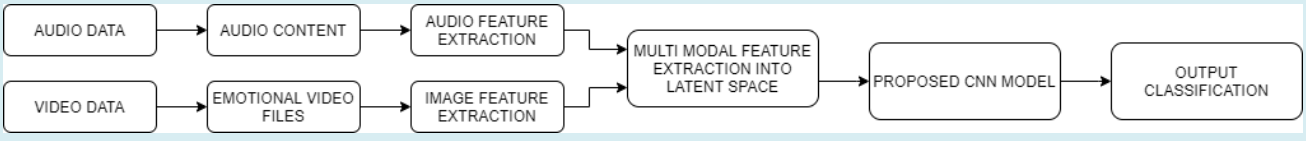

Figure 2: Workflow of the Our Work.

Experimentation and Results

Figures 3.a to 3.f give the train and test accuracies, train and test accuracies losses, and confusion matrices on the RAVDESS dataset. Figures 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 CREMAD dataset, train and test accuracies and train and test accuracies losses are shown. Figures 4a to 4f represent training and testing accuracy and loss comparisons in audio, video, and multimodal modes. A detailed description of the results is given in Table 5.

Name of Dataset	Type of Data	Train Accuracy	Test Accuracy	Train Loss	Test Loss
RAVDESS	Audio	91.11	94.66	0.193	0.1644
	Video	95.92	95.82	0.1676	0.1632
	Multi Modal	90.87	94.36	0.2392	0.15
CREMAD	Audio	74.77	79.45	0.8334	0.6071
	Video	94.31	96.62	0.2088	0.1644
	Multi Modal	65.74	70.14	0.7607	0.6333

Table 5: Accuracy and Loss of RAVDESS and CREMAD Datasets on different Types of Data in the Dataset.

The Table 5 presents detailed performance metrics for two datasets, RAVDESS and CREMAD, across different modalities such as audio, video, and a combination of both (multi-modal). For the RAVDESS dataset, in the audio modality, the training accuracy is 91.11%, the test accuracy is 94.66%, the training loss is 0.1930, and the test loss is 0.1644. In the video modality, the corresponding values are 95.92%, 95.82%, 0.1676, and 0.1632. The multi-modal results for RAVDESS show a training accuracy of 90.87%, a test accuracy of 94.36%, a training loss of 0.2392, and a test loss of 0.1500. Moving to the CREMAD dataset, the audio modality exhibits a training accuracy of 74.77%, a test accuracy of 79.45%, a training loss of 0.8334, and a test loss of 0.6071. In the video modality, the metrics are 94.31%, 96.62%, 0.2088, and 0.1644, while the multi-modal results are 65.74%, 70.14%, 0.7607, and 0.6333. These values provide a comprehensive overview of the model's performance on different datasets and modalities, serving as a valuable resource for evaluating and comparing. The effectiveness of the implemented models.

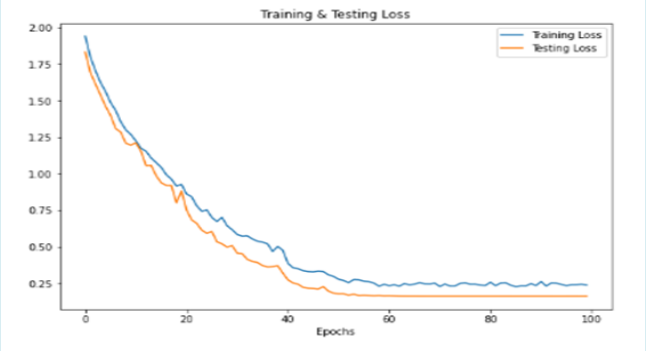

Figure 3a: Training and Testing Loss of Audio Data in RAVDESS Dataset.

Figure 3b: Training and Testing Accuracy of Audio Data in RAVDESS Dataset.

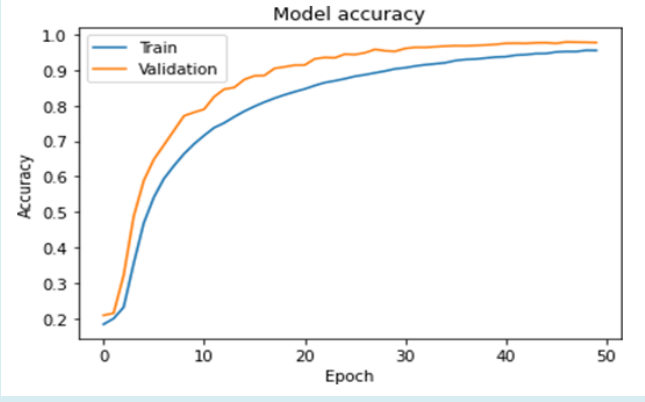

Figure 3c: Training and Testing Accuracy of Video Data in RAVDESS Dataset.

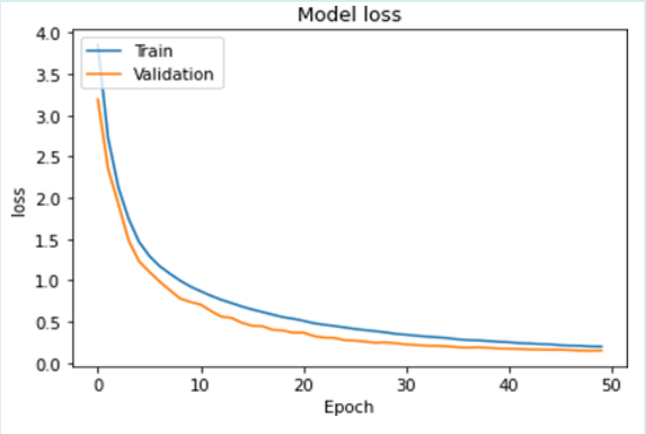

Figure 3d: Training and Testing Loss of Video Data in RAVDESS Dataset.

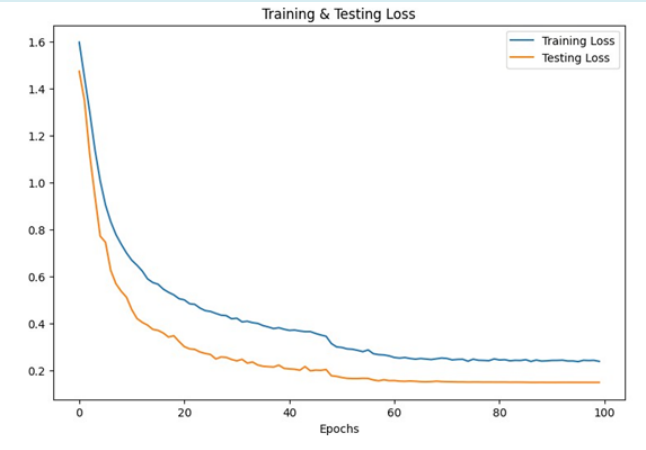

Figure 3e: Training and Testing Loss of Multi Modal Data in RAVDESS Dataset.

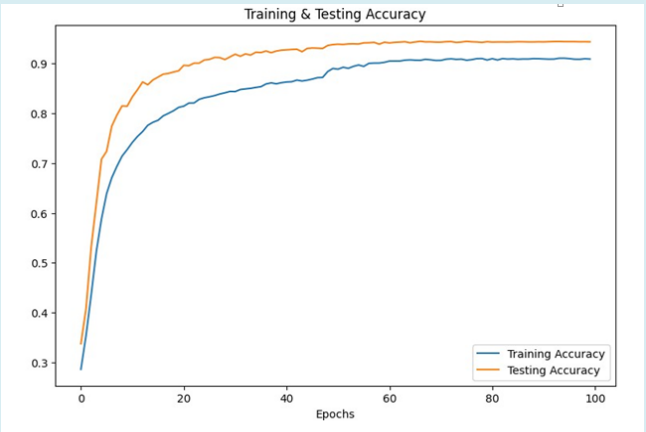

Figure 3f: Training and Testing Accuracy of Multi Modal Data in RAVDESS Dataset.

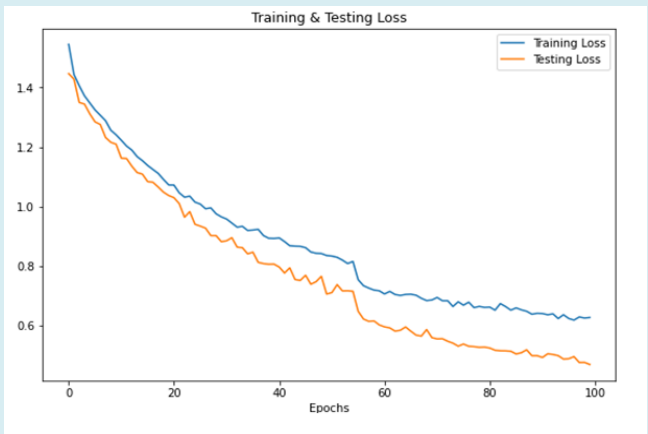

Figure 4a: Training and Testing Loss of Audio Data in CREMAD Dataset.

Figure 4b: Training and Testing Accuracy of Audio Data in CREMAD Dataset.

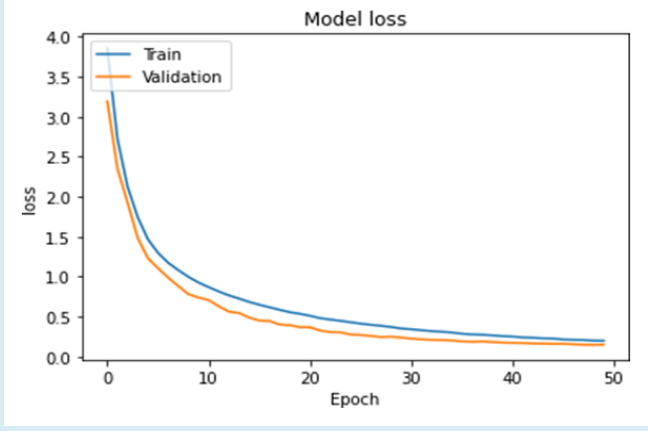

Figure 4c: Training and Testing Loss of Video Data in CREMAD Dataset.

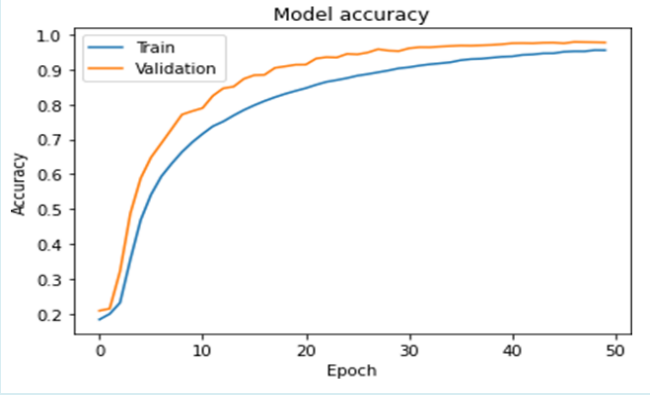

Figure 4d: Training and Testing Accuracy of Video Data in CREMAD Dataset.

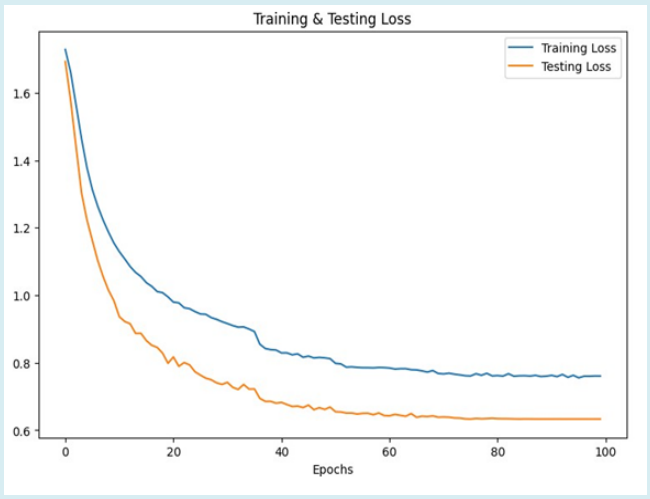

Figure 4e: Training and Testing Loss of Multi Modal Data in CREMAD Dataset.

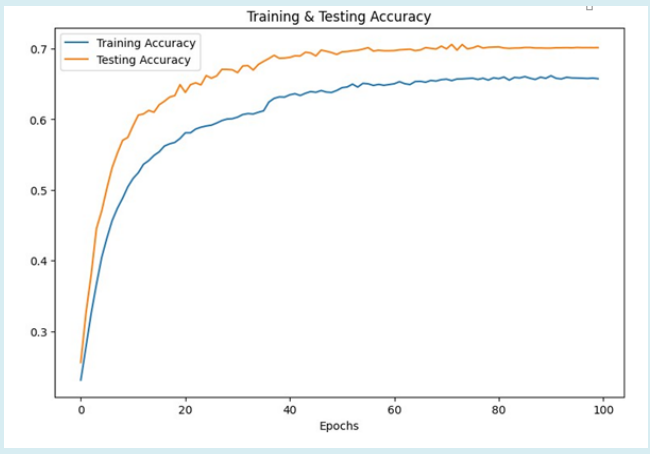

Figure 4f: Training and Testing Accuracy of Multi Modal Data in CREMAD Dataset.

Figures 5a & b represent the confusion matrix/ classification report of how the classes are classified during the testing phase on audio data of the RAVDESS and CREMAD datasets. The detailed description of various emotions and their respective performance measure values of RAVDESS and CREMAD audio data is given in Table 5. The table presents a performance evaluation of emotion classification on two datasets, RAVDESS and CREMAD, using various performance metrics such as precision, recall, and F1-score for different emotions. The "Name of the Dataset" column specifies the dataset under consideration, while the "Type of Emotion" column lists the specific emotions analysed. For the RAVDESS dataset, the precision values range from 0.92 to 0.99, recall values range from Table 6 (0.91 to 0.97), and F1-scores range from 0.92 to 0.97. The support column indicates the number of instances for each emotion. Similarly, for the CREMAD dataset, precision values vary from 0.73 to 0.94, recall values range from 0.73 to 0.87, and F1-scores range from 0.78 to 0.90. The results showcase the model's effectiveness in distinguishing emotions within each dataset, providing insights into the classification performance for different emotional states. The precision metric indicates the accuracy of positive predictions, while recall measures the ability to capture all relevant instances, and the F1-score is a harmonic mean of precision and recall. These metrics collectively offer a comprehensive evaluation of the emotion classification model's performance on the given datasets.

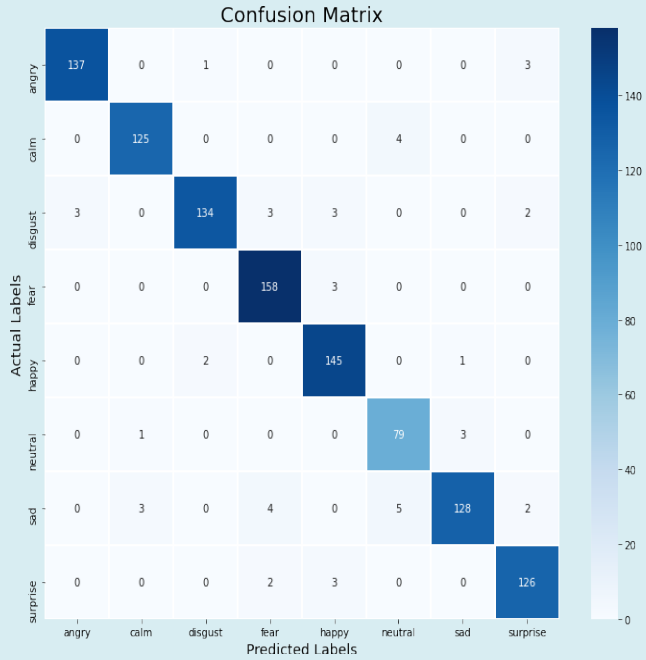

Figure 5a: Confusion Matrix/Classification Report of RAVDESS Dataset Audio Data.

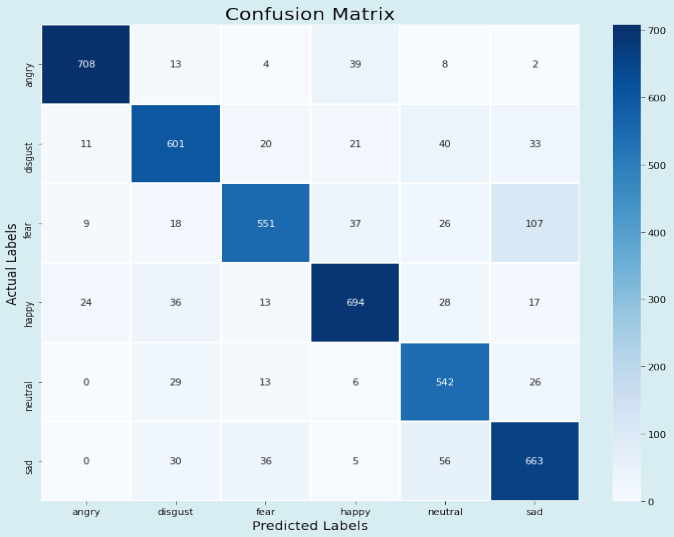

Figure 5b: Confusion Matrix/Classification Report of CREMAD Dataset Audio Data.

Name of the Dataset	Type of Emotion	Performance Metrics			
		Precision	recall	f1-score	Support
RAVDESS	Angry	0.97	0.96	0.97	564
	Calm	0.96	0.96	0.96	516
	Disgust	0.99	0.91	0.94	580
	Fear	0.94	0.97	0.96	644
	Happy	0.93	0.97	0.96	592
	Neutral	0.94	0.94	0.92	332
	Sad	0.95	0.91	0.94	568
	Surprise	0.92	0.94	0.96	524
CREMAD	Angry	0.94	0.87	0.9	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

Table 6: Performance Metrics of RAVDESS and CREMAD Datasets on Audio Data.

Figures 6a & b present below represent the confusion matrix/classification report of how the classes are classified during the testing phase on video data of the RAVDESS and CREMAD datasets. The detailed description of various emotions and their respective performance measure values of RAVDESS and CREMAD video data is given in Table 7. The table provides a detailed performance evaluation of emotion recognition using two datasets, RAVDESS and CREMAD, across various emotions. For RAVDESS, the emotions evaluated include Angry, Disgust, Fear, Happy, Neutral, and Sad, while CREMAD encompasses Angry, Disgust, Fear, Happy, Neutral, and Sad as well. The performance metrics considered are Precision, Recall, and F1-score, along with the Support value, indicating the number of instances for each emotion in the dataset. For RAVDESS, the precision scores range from 0.98 to 0.99, recall scores from 0.96 to 0.98, and F1-scores from 0.97 to 0.98. The Support values vary for each emotion, reflecting the dataset's diversity. Similarly, for CREMAD, precision ranges from 0.96 to 0.99, recall from 0.97 to 0.98, and F1-score from 0.96 to 0.98. The Support values again vary across

emotions. The results suggest high overall performance in emotion recognition for both datasets, with slight variations in performance metrics among different emotions. These metrics provide a comprehensive understanding of the models' ability to correctly identify specific emotions within each dataset, aiding in the assessment and improvement of emotion recognition systems.

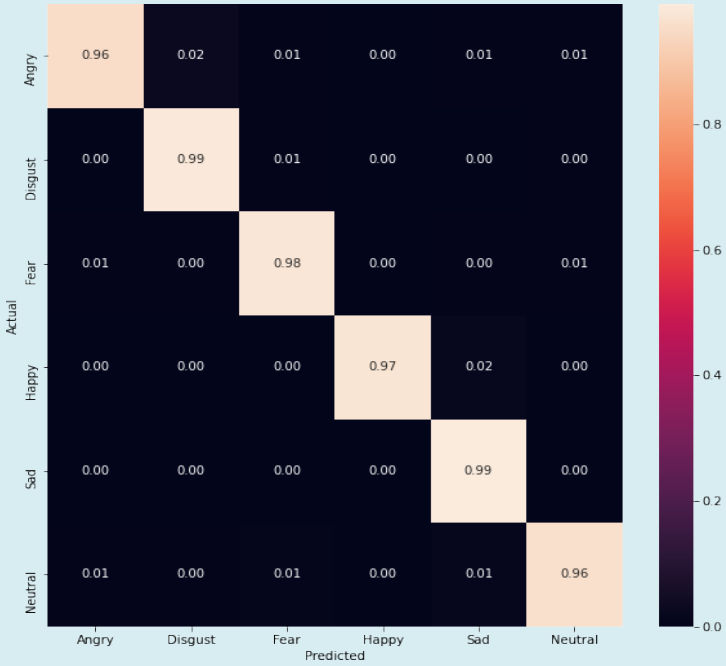

Figure 6a: Confusion Matrix/Classification Report of RAVDESS Dataset Video Data.

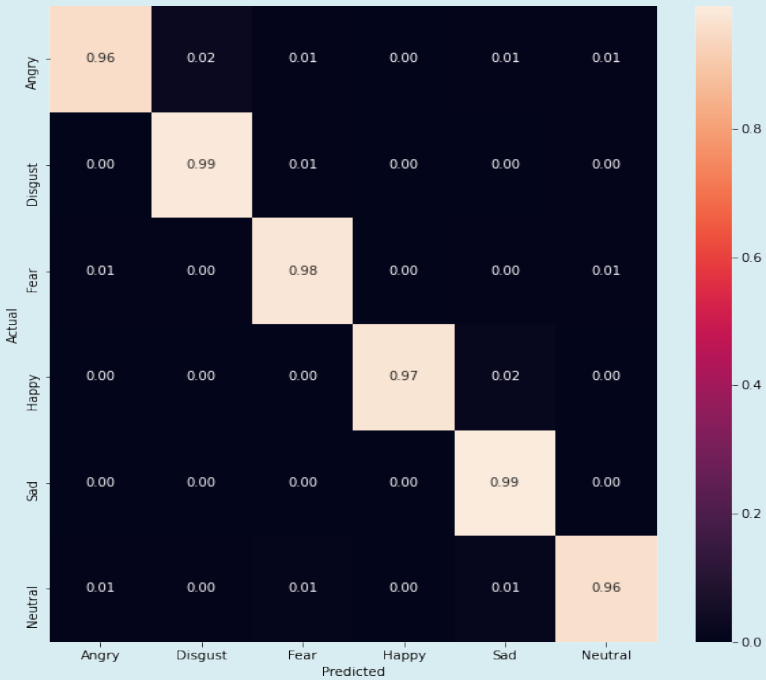

Figure 6b: Confusion Matrix/Classification Report of CREMAD Dataset Video Data.

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

Table 7: Performance Metrics of RAVDESS and CREMAD Datasets on Video Data.

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. Figures 7a & 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. The detailed description of the results is explained in Table 8. The table presents performance metrics for emotion recognition on two datasets: RAVDESS and CREMAD. The first dataset, RAVDESS, is evaluated for six emotions: Angry, Disgust, Fear, Happy, Neutral, and Sad. Precision, recall, and f1-score are reported for each emotion, along with the corresponding support values. RAVDESS achieves high performance across emotions, with particularly notable precision and f1-score scores for Angry (0.97 and 0.97, respectively) and Neutral (0.99 and 0.96, respectively). Disgust and Sad also demonstrate strong performance, though Disgust exhibits a slightly lower precision of 0.90. The second dataset, CREMAD, is assessed for the same set of emotions using the same metrics. However, CREMAD generally exhibits lower performance compared to RAVDESS. Notably, the precision for Angry in CREMAD is considerably lower at 0.56, indicating a higher rate of false positives. Additionally, Sad has a lower f1-score of 0.60. The variations in performance between the two datasets underscore the importance of dataset selection in emotion recognition tasks, as different datasets may pose unique challenges that impact model performance. Further analysis and consideration of factors influencing these results would be necessary for a comprehensive understanding of the effectiveness of emotion recognition models on these datasets.

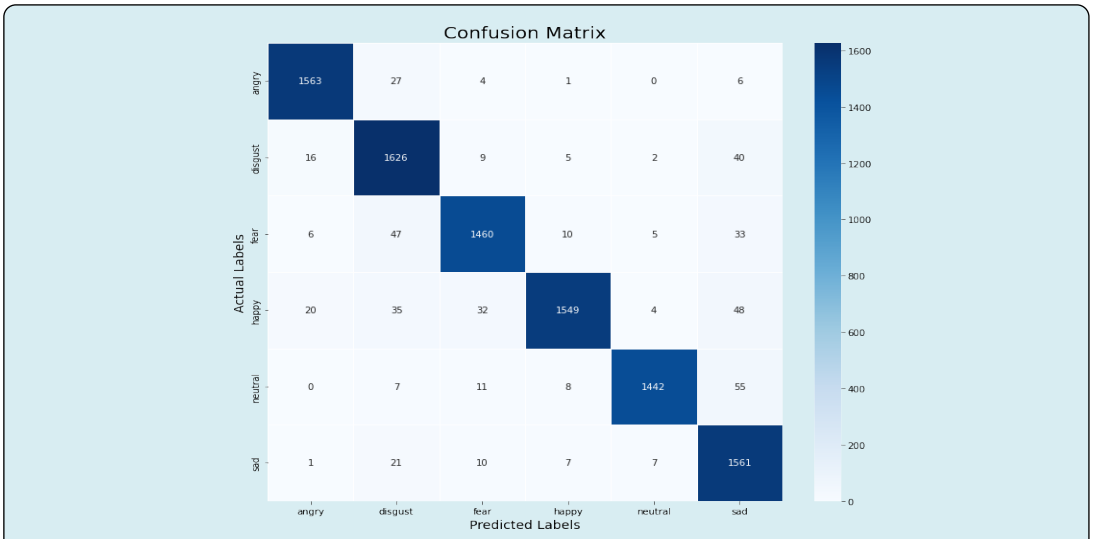

Figure 7a: Confusion Matrix/Classification Report of RAVDESS Dataset on Multimodal Data.

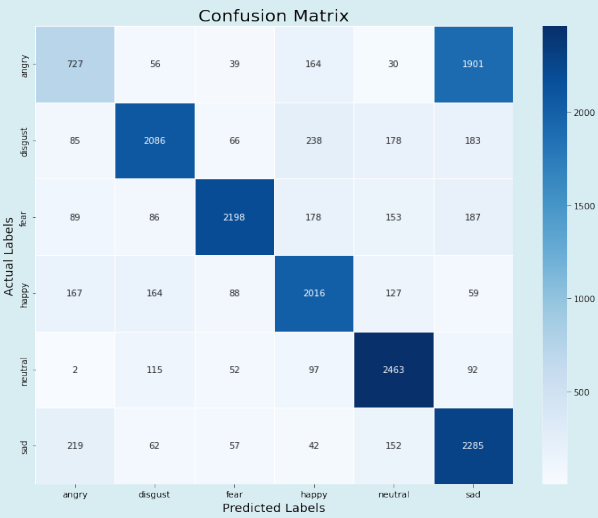

Figure 7b: Confusion Matrix/Classification Report of CREMAD Dataset on Multimodal Data.

Name of the Dataset	Type of Emotion	Performance Metrics			
		Precision	recall	f1-score	Support
RAVDESS	Angry	0.97	0.97	0.97	1601
	Disgust	0.9	0.95	0.93	1698
	Fear	0.96	0.92	0.94	1561
	Happy	0.97	0.92	0.94	1688
	Neutral	0.99	0.94	0.96	1523
	Sad	0.89	0.96	0.92	1607
CREMAD	Angry	0.56	0.26	0.36	2917
	Disgust	0.8	0.76	0.78	2836
	Fear	0.9	0.76	0.82	2891
	Happy	0.75	0.77	0.76	2621
	Neutral	0.8	0.88	0.84	2821
	Sad	0.48	0.79	0.6	2817

Table 8: Performance Metrics of RAVDESS and CREMAD Datasets on Multimodal Data.

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 9. The table presents performance metrics for emotion recognition models trained on two different datasets-RAVDESS and CREMAD across three types of data: Audio, Video, and Multimodal (a combination of both audio and video). For the RAVDESS dataset, the models achieved high macro average accuracy for all three data types, with scores ranging from 0.94 to 0.96. Specifically, in the Audio category, precision, recall, and f1-score were 0.95, 0.94, and 0.95, respectively, with a support of 1080 instances. The Video and Multimodal categories demonstrated similarly impressive results. Moving on to the CREMAD dataset, the models performed well but generally exhibited lower scores compared to RAVDESS. In the Audio category, precision, recall, and f1-score were 0.79, 0.79, and 0.78, with a support of 4466 instances. Video and Multimodal categories showed higher precision, recall, and f1-score values, with weighted averages consistently outperforming macro averages. This table provides a comprehensive overview of the model's classification accuracy across different datasets and modalities, aiding in the assessment of their effectiveness in emotion recognition tasks.

Name of the Dataset	Type of Data	Macro Average Accuracy				Weighted Average Accuracy			
		Precision	recall	f1- score	support	Precision	recall	f1- score	support
RAVDESS	Audio	0.95	0.94	0.95	1080	0.94	0.94	0.94	1080
	Video	0.96	0.95	0.96	9292	0.95	0.95	0.95	9232
	Multimodal	0.95	0.94	0.94	9678	0.94	0.94	0.94	9678
CREMAD	Audio	0.79	0.79	0.78	4466	0.8	0.79	0.8	4466
	Video	0.95	0.96	0.95	12642	0.97	0.96	0.96	12642
	Multimodal	0.72	0.7	0.69	16903	0.72	0.7	0.69	16903

Table 9: Macro Average and Weighted Accuracies of performance metrices in different modes on RAVDESS and CREMAD datasets.

Conclusion

In conclusion, this research paper introduces a novel multimodal system named DV EmotionNet for emotion recognition, leveraging both audio and video information to enhance the accuracy of emotion classification. The DV EmotionNet approach involves extracting audio features through the Mel-Frequency Cepstral Coefficients (MFCC) technique and converting videos into images stored in a spatial-temporal space, with image features extracted using a gaussian weighted function. The Multimodal Fusion and Attention (MFA) technique is employed to merge audio and video features, and the resulting integrated features are utilized for training and evaluating the FERCNN Model. The experiments conducted on the RAVDESS and CREMAD datasets, comprising audio and video data, demonstrate the effectiveness of the proposed approach. Upon examining Table 8, it is also evident that the multimodal approach yields notable improvements in emotion recognition accuracy compared to unimodal approaches. Specifically, the macro and weighted average accuracy scores across various metrics for both RAVDESS and CREMAD datasets consistently outperform the individual audio and video modalities. The ability of the multimodal system to capture complementary information from both audio and visual cues contributes to its superior performance. This underscores the importance of incorporating multiple modalities for a more comprehensive understanding of emotional expressions. Moreover, the DV EmotionNet, combining the MFA fusion technique and the FERCNN Model, demonstrates robust performance across diverse datasets, highlighting its potential applicability in real-world scenarios requiring accurate emotion recognition. This research not only advances the field of emotion recognition but also emphasizes the significance of multimodal approaches in enhancing the reliability and versatility of such systems. Furthermore, we plan to explore this work on DV EmotionNet further by entertaining and developing the possibility of real- time detection applications that can contribute to a variety of fields in general and criminology in particular.

References

1. Radoi A, Birhala N, Ristea C, Dutu LC (2021) An End-to-End Emotion Recognition Framework Based on Temporal Aggregation of Multimodal Information. IEEE Access 9: 135559-135570.

2. Cavallo F, Semeraro F, Fiorini L, Magyar G, Sincak P, et al. (2018) Emotion Modelling for Social Robotics Applications: A Review. J Bionic Eng 15(2): 185-203.

3. Katsis CD, Katertsidis N, Ganiatsas G, Fotiadis DI (2008) Toward Emotion Recognition in Car-Racing Drivers: A Biosignal Processing Approach. IEEE Trans Syst Man Cybern A Syst Humans 38(3): 502-512.

4. Tzirakis P, Trigeorgis G, Nicolaou MA, Schuller B, Zafeiriou S (2017) End-to-End Multimodal Emotion Recognition Using Deep Neural Networks. Journal of Latex Class Files 14(8): 1-9.

5. Ji Q, Zhu Z, Lan P (2004) Real-Time Nonintrusive Monitoring and Prediction of Driver Fatigue. IEEE Trans Veh Technol 53(4): 1052-1068.

6. Burkhardt F, Ajmera J, Englert R, Stegmann J, Burleson W (2006) Detecting Anger in Automated Voice Portal Dialogs. Proc Annu Conf Int Speech Commun Assoc 2: 1053-1056.

7. Zemel RS, Hinton GE (1994) Autoencoders, Minimum Description Length and Helmholtz Free Energy. Proc Neural Inf Process Syst, USA, pp: 3-10.

8. LeCun Y (1989) Generalization and Network Design Strategies. In: R Pfeifer, et al. (Eds.), Connectionism in Perspective, Elsevier, Zurich, Switzerland, pp: 143-155.

9. Hinton GE, Osindero S, The YW (2006) A Fast Learning Algorithm for Deep Belief Nets. Neural Comput 18(7): 1527-1554.

10. Hochreite S, Schmidhuber J (1997) Long Short-Term

Memory. Neural Comput 9(8): 1735-1780.

11. Hu D, Li X, Lu X (2016) Temporal Multimodal Learning in Audiovisual Speech Recognition. Proc IEEE Conf Comput Vis Pattern Recognit, Las Vegas, NV, USA, pp: 3574-3582.

12. Huang Y, Lu H (2016) Deep Learning Driven Hypergraph Representation for Image-Based Emotion Recognition. Proc Int Conf Multimodal Interaction, Tokyo, Japan, pp: 243-247.

13. Kahou SE, Michalski V, Konda K, Memisevic R, Pal C (2015) Recurrent Neural Networks for Emotion Recognition in Video. Proc Int Conf Multimodal Interaction, Seattle, WA, USA, pp: 467-474.

14. Katsis CD, Katertsidis N, Ganiatsas G, Fotiadis DI (2008) Toward Emotion Recognition in Car- Racing Drivers: A Biosignal Processing Approach. IEEE Trans Syst Man Cybern A Syst Humans 38(3): 502-512.

15. Essa, Pentland A (1994) A Vision System for Observing and Extracting Facial Action Parameters. Proc CVPR, pp: 76-83.

16. Essa, Gardner A (1997) Prosody Analysis for Speaker affect Determination. Proc Workshop Perceptual User Interfaces, Banff, Alberta, Canada, pp: 45-46.

17. Kim Y, Lee H, Provost EM (2013) Deep Learning for Robust Feature Generation in Audiovisual Emotion Recognition. Proc Int Conf Acoust, Speech, Signal Process, Vancouver, Canada, pp: 3687-3691.

18. Simonyan K, Zisserman A (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. 3rd International Conference on Learning Representations, pp: 1409-1556.

19. He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. Proc IEEE Conf Comput Vis Pattern Recognit (CVPR), pp: 770-778.

20. Bargal SA, Barsoum E, Ferrer CC, Zhang C (2016) Emotion Recognition in the Wild from Videos using Images. Proc 18th ACM Int Conf Multimodal Interact, pp: 433-436.

21. Wei W, Jia Q, Feng Y, Chen G, Chu M (2019) Multi-Modal Facial Expression Feature based on Deep- Neural Networks. J Multimodal User Interfaces 14(1): 17-23.

22. Hamester, Barros P, Wermter S (2015) Face Expression Recognition with a 2-Channel Convolutional Neural Network. Proc Int Joint Conf Neural Netw (IJCNN), pp: 1-8.

23. Guo JJ, Zhou R, Zhao LM, Lu BL (2019) Multimodal Emotion Recognition from Eye Image, Eye Movement and EEG using Deep Neural Networks. Proc 41st Annu Int Conf IEEE Eng Med Biol Soc (EMBC), Berlin, Germany, pp: 3071-3074.

24. Santamaria-Granados L, Munoz-Organero M, Ramirez-Gonzalez G, Abdulhay E, Arunkumar N (2019) Using Deep Convolutional Neural Network for Emotion Detection on a Physiological Signals Dataset (AMIGOS). IEEE Access 7: 57-67.

25. He G, Liu X, Fan F, You J (2020) Classification-Aware Semi-Supervised Domain Adaptation. Proc. IEEE/CVF Conf Comput Vis Pattern Recognit Workshops (CVPRW), pp: 4147-4156.

26. Athanasiadis C, Hortal E, Asteriadis S (2020) Audio-Visual Domain Adaptation using Conditional Semi-Supervised Generative Adversarial Networks. Neurocomputing 397: 331-344.

27. Eyben, Wollmer M, Schuller B (2010) Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. Proc Int Conf Multimedia (MM), pp: 1459-1462.

28. Zhao Z, Bao Z, Zhao Y, Zhang Z, Cummins N, et al. (2019) Exploring Deep Spectrum Representations via Attention-Based Recurrent and Convolutional Neural Networks for Speech Emotion Recognition. IEEE Access 7: 97515-97525.

29. Dawson ME, Schell AM, Filion DL (2000) The Electrodermal System. Handbook of Psychophysiology, Cambridge University Press, New York.

30. Haag A, Gornzy S, Schaich P, Williams J (2004) Emotion Recognition using Bio-Sensors: First Steps towards an Automatic System. Affective Dialogue Systems. Springer-Verlag, New York, 3068: 36-48.

31. Katsis CD, Katersidis N, Ganiatsas G, Fotiadis DI (2005) Aubade-A Wearable EMG Augmentation System for Robust Emotional Understanding. Proc ICICTH, pp: 292-297.

32. Kim KH, Bang SW, Kim SR (2004) Emotion Recognition System using Short-term Monitoring of Physiological Signals. Med Biol Eng Comput 42(3): 419-427.

33. Fu Z, Liu F, Wang H, Qi J, Fu X, et al. (2021) A Cross-Modal Fusion Network based on Self-Attention and Residual Structure for Multimodal Emotion Recognition. arXiv: 2111.02172.

34. Cao H, Cooper DG, Keutmann MK, Gur RC, Nenkova A, et al. (2014) CREMA-D: Crowd-Sourced Emotional

Multimodal Actors Dataset. IEEE Trans Affect Comput 5(4): 377-390.

35. Chatterjee R, Mazumdar S, Sherratt RS, Halder R, Maitra T, et al. (2021) Real-Time Speech Emotion Analysis for Smart Home Assistants. IEEE Transactions on Consumer Electronics 67(1): 68-76.

36. Xu M, Zhang F, Zhang W (2021) Head Fusion: Improving the Accuracy and Robustness of Speech Emotion Recognition on the IEMOCAP and RAVDESS Dataset. IEEE Access 9: 74539-74549.

37. Chang X, Skarbek W (2021) Multi-Modal Residual Perceptron Network for Audio-Video Emotion Recognition. Sensors 21(16): 5452.

38. 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): e0196391.

39. Wang W, Tran D, Feiszli M (2020) What Makes Training Multi-Modal Classification Networks Hard? Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp: 12692-12702.

40. Donuk K (2022) CREMA-D: Improving Accuracy with BPSO-Based Feature Selection for Emotion Recognition Using Speech. JSCAI 3(2): 51-57.

41. Beard R, Das R, Ng RW, Gopalakrishnan PK, Eerens L, et al. (2018) Multi-modal Sequence Fusion via Recursive Attention for Emotion Recognition. Proceedings of the 22nd Conference on Computational Natural Language Learning, pp: 251-259.

42. Samadiani N, Huang G, Cai B, Luo W, Chi CH, et al. (2019) A Review on Automatic Facial Expression Recognition Systems Assisted by Multimodal Sensor Data. Sensors (Basel) 19(8): 1863.

43. Ghaleb E, Popa M, Asteriadis S (2019) Metric Learning-based Multimodal Audio-Visual Emotion Recognition. IEEE Multimedia 27(1): 37-48.

44. He G, Liu X, Fan F, You J (2020) Image 2 audio: Facilitating Semi-supervised Audio Emotion Recognition with Facial Expression Image. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp: 912-913.

45. Kamarol SKA, Jaward MH, Parkkinen J, Parthiban R (2016) Spatiotemporal Feature Extraction for Facial Expression Recognition. IET Image Processing 10(7): 534-541.

46. Harris C, Stephens M (1988) A Combined Corner and Edge Detector. Proc Alvey Vision Conf, Manchester, pp: 147-152.

47. Zhang S, Huang Q, Jiang S, Gao W, Tian Q (2010) Affective Visualization and Retrieval for Music Video. IEEE Trans Multimedia 12(6): 510-522.

48. Fragopanagos N, Taylor JG (2015) Emotion Recognition in Human- Computer Interaction. Neural Netw 18(4): 389-405.

49. Leymani, Pantic M, Pun T (2012) Multimodal Emotion Recognition in Response to Videos. IEEE Trans Affect Comput 3(2): 211-223.