International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7464
***Article Received**: 21 May 2023 **Revised:** 14 July 2023 **Accepted:** 02 August 2023*

Critical Analysis on Multimodal Emotion Recognition in Meeting the Requirements for Next Generation Human Computer Interactions

Shwetkranti Taware[1], Anuradha Thakare[2]
[1]Research Scholar, Department of Computer Engineering,
Pimpri Chichwad College of Engineering
Pune, India
shweta.taware@gmail.com
[2]Department of Computer Engineering,
Pimpri Chichwad College of Engineering,
Pune, India.
anuradha.thakare@pccoepune.org

Abstract—Emotion recognition is the gap in today's Human Computer Interaction (HCI). These systems lack the ability to effectively recognize, express and feel emotion limits in their human interaction. They still lack the better sensitivity to human emotions. Multi modal emotion recognition attempts to addresses this gap by measuring emotional state from gestures, facial expressions, acoustic characteristics, textual expressions. Multi modal data acquired from video, audio, sensors etc. are combined using various techniques to classify basis human emotions like happiness, joy, neutrality, surprise, sadness, disgust, fear, anger etc. This work presents a critical analysis of multi modal emotion recognition approaches in meeting the requirements of next generation human computer interactions. The study first explores and defines the requirements of next generation human computer interactions and critically analyzes the existing multi modal emotion recognition approaches in addressing those requirements.

Keywords- Human Computer Interaction (HCI); multimodal emotion; context awareness; micro expression; personalization.

I. INTRODUCTION

Communication between humans is 70-90% non verbal with 55% visual and 38% vocal [1]. Thus there human communication has its significant influence on non verbal components. The non verbal behavior does have any linguistic content but they have variety of communicative behaviors like facial expression, smile, eye glance, body postures, pitch changes, speech dysfluency, loudness etc. These non verbal communication means carry more emotional information [2]. Some of the non verbal communication means and the emotion they convey in a discussion are illustrated below in Table 1. These normal verbal communication means can be used to identify if the meaning conveyed through verbal communication is authentic or fake. To illustrate, in a discussion a person may say he agrees but frown on his face can be used to identify his true intention that he is not happy with the discussion.

Table 1 Expression emotion correlation

Expression	Emotion
Frown	Disapproval
Smile	Agreement
Blank	Boredom

An earlier attempt to make use of human emotion detection was mood ring [3]. Mood ring was a ring with stone. The color of stone changed in response to temperature, potentially indicating wearer's emotion. Colors were mapped to various human emotions. Rosalind Picard was a pioneer in area of affective computing. He attempted to model emotion using non linear sigmoid function [4]. As a prelude, various automatic emotion recognition systems have been developed over last two decades. Emotion recognition has been increasingly used in various commercial applications in domains of banking, market research, advertising, health care etc. Many attempts have been made by various researches to correlate physiological and behavioral responses to human emotions.

These researches have attempted to detect emotions using various modalities like speech, facial expressions, text, body gestures and movements, Autonomic Nervous System (ANS) and physiological signals [5]. Earlier emotion detection approaches were unimodal and they used one type of input data to detect emotions. Emotion detection from text is basically sentiment analysis task. Machine learning and artificial intelligence concepts are used for sentiment analysis tasks. The sentiment analysis methods detected the polarity of text (in

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7464
***Article Received**: 21 May 2023 **Revised:** 14 July 2023 **Accepted:** 02 August 2023*

three class of positive, negative and neutral) and valence score indicating the strength of polarity. Sentiment analysis involves following process of pre-processing and analysis. Preprocessing involves part of speech tagging and extracting relationship between entities. Analysis in various levels of document, aspect and sentience level is conducted using various supervised, semi supervised and hybrid methods to learn the sentiments ([6]-[8]). Facial expression based emotion detection system were based Facial Action Coding System (FACS) which has its foundation on Ekman's theory of basic emotion. It is a dominant emotion theory classifying facial expressions [9]. Facial action coding system (FACS) mapped the series of muscular movements in face to basic emotions. It was proposed in 1978, updated in 1992 and revised in 2002[10]. FACS classifies emotions based on the characteristics or movements of various actions units in Face. Ekman and Friesen defined 44 facial action units and its was later refined to 68 different facial action units. Some of these facial action units are not correlated to any specific emotion. Emotion detection from speech were based on examining the paralinguistic characteristics like tone of voice, intensity etc. Features like speech signal energy, spectral variables and pitch contour are extracted from speech and classified to emotions ([11], [12]). Emotion detection from ANS involves ANS responses like heart rate [13], skin conductance levels [14]. Emotion detection using physiological signals involve using signals captured from human body like Electromyogram (EMG), Galvanic Skin Response (GSR), Respiratory Volume (RV), Skin Temperature (SKT), Blood Volume Pulse (BVP), Heart Rate (HR), Electro Cardiogram (ECG) and Photo Plethysmography (PPG) etc. to classify the emotions [15]. Multi modal emotion detection is a transition from unimodal and many initial works found higher classification accuracy by incorporating multimodal signals ([16-20]). Use of multi modal features to describe emotions was found to be more comprehensive and detailed ([21-22]). It can supplement for less emotional information in one modality with another modality thereby improving the classification accuracy ([23-24]). The future of human computer interaction is use of multi modal features for emotion recognition.

The interaction between text, video and speech modalities helps to predict the emotion with higher accuracy compared to individual modalities. The bias affecting individual modalities has higher chance of being removed in multi modalities. Multi modal modalities are able to gather comprehensive information from complementary information sources [25]. Many multi modal emotional recognition systems have been designed using various combinations of modalities like text, video, speech and physiological signals ([26])

This work makes a critical analysis of existing multi modal emotion detection approaches. Though there were many such recent survey on multi modal emotion detection approaches, most of them were based on comparison in terms of number and type of modalities, datasets, evaluation metrics, classifier and the number of emotion labels classified([27]-[30]).

Differing from it, this survey explores the upcoming requirements for multi modal emotion recognition by the state of art applications and defines a set of application requirements as shown in Table 2 on multi modal emotion recognition. The existing multi modal emotion detection approaches are evaluated in terms of those requirements and the gaps in existing approaches are identified

Table 2 Dimensions for critical analysis

Critical Analysis Dimensions	
Context awareness (R1)	Ability to detect emotions based on shift in environment and demographics
Robustness to real world scenarios (R2)	Ability to recognize emotions in presence of various disturbances and noises in data acquisition
Sensitivity to micro expression (R3)	Ability to detect subtle expressions or emotion faking
Personalization (R4)	Ability to adapt to emotion expressing characteristics of a person
Emotions class coverage (R5)	Ability to cover more emotions in Plutchik wheel of emotions as in Figure 1.

.

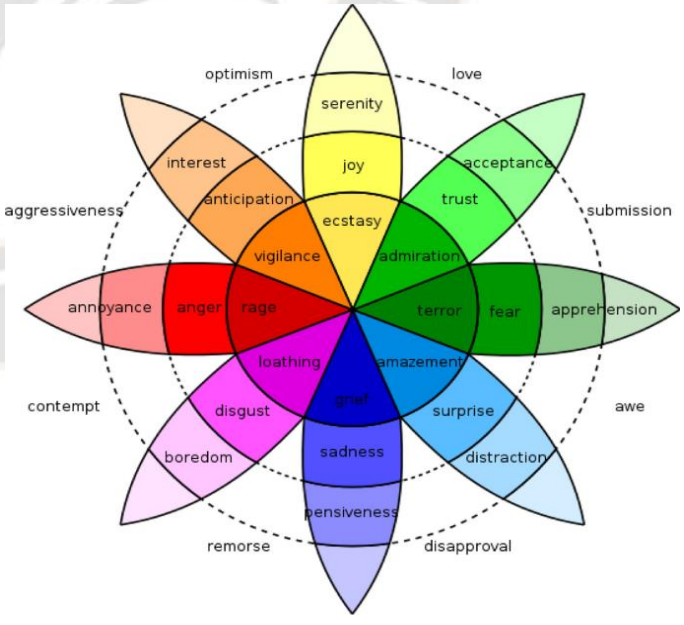

Figure 1 Plutchik Wheel of emotions [36]

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7464
***Article Received**: 21 May 2023 **Revised:** 14 July 2023 **Accepted:** 02 August 2023*

II. CRITICAL ANALYSIS PROCESS

The critical analysis process followed in this work is given in Figure 2. A secondary data analysis is conducted over the next generation human computer interactions in aim to identify important requirements to be addressed by the emotion recognition systems. These requirements are defined. Critical analysis on existing multi modal emotion recognition solutions are conducted for both methodology and dataset in terms of requirement dimensions defined. From the critical analysis gaps in existing solutions to meet the defined requirements are found. Recommendation model for multi modal emotion recognition is framed to address the gaps. The possible directions for improving the dataset to test the recommendation model are also presented.

The next generation human computer interactions scenarios and the requirements on emotion detection systems are discussed in this section.

The current Chatbot systems like Alexa, Siri to name a few lacks emotional cognizance and they are not able to involve in an emotional interaction with humans [31]. But Chatbot applications need to evolve towards emotion based interaction as applications like psychological counseling, patient health care etc. need interaction at an emotional level. To achieve it, smart human machine dialog generation must be designed. This advanced dialog generation facilitates fluent and highly interactive emotional conversation. One of importance challenge in emotion recognition for effective human machine dialog generation is that emotion recognition systems do not have contextual awareness or don't are not able to detect change in conversational behavior.

Most of the emotion recognition applications perform well in highly controlled environments. But these conditions hardly exist in real world applications. With emotion recognition models trained on dataset of controlled environments, it becomes difficult to generalize these models to natural recordings made in unconstrained settings. For models to work in realistic environments, they should adapt to various factors like pose changes, illumination variations etc.

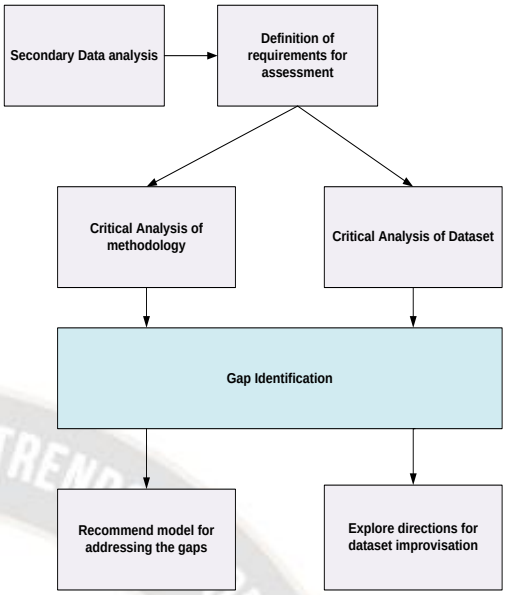

Figure 2 Critical analysis process

Micro Expression (ME) analysis is gaining importance in various applications like security and forensic applications [31]. ME occur very briefly and very subtle. They are caused generally due involuntary expressions and can be hidden with false emotion expression after its subtle occurrence. These expressions last only between 1/25 and 1/5 of a second. Micro expressions can occur during macro expressions. In presence of micro expression, emotion detection becomes erroneous [32]. Emotion recognition systems must detect micro expressions and distinguish them when they co-occur or even overlap to recognize emotions accurately.

There is a difference in expression of emotion by a person depending on situation, interaction partner and even the time of day ([34-35]). But most datasets used for training emotion recognition systems are very generic and model trained with these datasets could not address the need for personalization across gender, age or race etc. But with emotion detection based services gaining in roots into various services like health care it is necessary to accommodate personalization into emotion recognition.

Most the existing emotion recognition systems are based on Ekman's six basic emotion theory and they were trained to classify six basic emotions of anger, disgust, fear, joy, sadness and surprise. HCI (Human Computer Interaction) systems which can evolve continuously based on emotional profile of humans its interacting with is a latest happening in HCI research. These systems continuously improvise themselves with aim to improve the interaction experience. These kinds of systems need to recognize more emotion classes then covered by Ekman's six basic emotion model. If the emotion recognition system can detect as much emotions in the Robert Plutchik wheel of emotions [36], advanced HCI based adaptive systems

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7464
***Article Received**: 21 May 2023 **Revised:** 14 July 2023 **Accepted:** 02 August 2023*

can be realized.

Based on the requirements of the next generation HCI applications, this work intend to critically analyse the existing multimodal approaches in terms of five dimensions listed in Table 2.

III. CRITICAL ANALYSIS ON METHODLOGY

Zheng et al. [37] proposed a multi modal emotion recognition system called EmotionMeter combining movement of eyes and EEG. Six electrodes were placed above the ears to collect EEG signals. The EEG and eye movement are combined and analyzed to detect the internal cognitive states. The method was able to classify four emotions: happy, sad, fear and neutral. EEG has higher sensitivity to happy emotional state and eye movement has higher sensitivity to fear. The method was able to achieve 72.39% accuracy in classifying the four emotions. Deep neural networks were used for classification. This approach is more suitable for controlled environment like lie detection in forensics as the data acquisitions is invasive. The approach was not adaptive to EEG fluctuations and environmental changes. Factors like noise, displacement in electrode positions introduces errors in emotion recognition. The approach lacked context awareness. The emotions were recognized only based on measurements without using any reasoning on context knowledge and the environment impact on the eye movements. Due to larger windowing in feature extraction, the approach cannot detect micro expressions. The training was not adapted to any personal characteristics, due to which the approach cannot be adapted for emotion characteristics of a person. Only four emotions were classified without situational awareness and past emotions states, thus more derived emotions cannot be learnt.

Nguyen et al. [38] proposed a multimodal emotion recognition system based on spatio temporal modeling different from earlier works on feature fusion or decision fusion. A novel 3 dimensional neural network was proposed to model the spatio temporal information across audio and video inputs. The solution classified six different emotions of anger, disgust, fear, happiness, sadness and surprise with an average accuracy of 82.83%. The emotions are classified for 16 frames window and it limits this approach in detection of micro expression. No preprocessing were done on frames for correcting illumination or pose variations and this accuracy will be low for real world scenarios. Though the approach classifies only six basic emotions, it has ability to cover more emotions due to its use of deep belief networks in training stage and score level fusion. By changing the training set, the approach can be extended to cover more emotions. Training process does not have any provisions for personalization.

Tzirakis et al. [39] used auditory and visual modalities to detect emotions using deep neural network. Convolutional neural network features are extracted from speech. Deep residual features are extraction from videos. The features are classified to emotion using a Long Short Term Memory(LSTM). The method was able to classify two emotions of valence and arousal with accuracy of 62%. LSTM can remember only short term context and there is no way to feed long term context information for better emotion detection. The training is generic and it does not accommodate personalization. The approach cannot be extended for more emotion class as it feature discriminating ability is limited due to it decision level fusion. Since features are learnt over longer window duration, the solution does not ability to classify micro expressions.

Zhang et al. [40] extracted various handcrafted features from the multiple physiological signals and built a ensemble dense embedding of multi modal features. The fused features are then used for classification of arousal and valence using traditional machine learning classifier. The method was able to achieve an accuracy of 60%. Neither short nor long term contexts are realized in this method. The method was not sensitive to physiological signal characteristics during the duration of micro expressions. Due to use of physiological signals, the method can be easily adopted for personalization as these signals don't show much diversity. Due to use of handcrafted features and without learning correlation between features to different emotion classes, the approach is difficult to extend for multiple emotion classes. Most of the signals used in this work are invasive and the approach can work only for controlled environment like response to a video etc. The approach lacks ability to be applied to non invasive real world scenarios.

Chen et al. [41] exploited the advantage of temporal and spatial consistency using a learning network for multi modal emotion recognition involving speech and text modalities. The approach was able to recognize the short term contextual information but it is very narrow for longer emotional inconsistency detection. Personalization is not considered in learning. Also the bias in texts and emotional deceiving in form of micro expressions were not focused. Context information of text and the expected emotional trigger were not considered. Bu the approach is non invasive and can be used in real world environments.

Cimtay et al. [42] proposed a multimodal emotion recognition system combining facial cues, skin responses and brain waves. CNN (Convolution Neural Network) features extracted from each modality are classified using random forest classifier and decision of each module is fused. With usage of hybrid fusion, the method was able to achieve an average accuracy of 74% for three basic emotions of happiness, sad and neutral.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7464
***Article Received**: 21 May 2023 **Revised:** 14 July 2023 **Accepted:** 02 August 2023*

Personalization in terms of tuning the datasets or learning process is not considered in this works. Also the approach lacks both short term and long term context information in decision making. The approach is limited to six basic emotions. Due to large window in feature extraction it misses out micro expressions. The approach works only in controlled environment and it does not consider noises during acquisition in real world scenarios.

Zhou et al. [43] proposed a novel multimodal fusion attention network for emotion recognition. It used speech and visual features. Deep learning features are extracted separately from speech and video, and feature level fusion is done with adaptive weighting for speech and visual features. The method was able to achieve an average accuracy of 61%. Longer window makes the approach unsuitable for micro expression detection. The approach is snap shot based and does not use the context information both short and long term. The approach is suited for real world scenarios as the encoder can segregate between emotion and non emotional parts in the modalities. The approach covers only six basic emotions.

Wu et al. [44] used three multimodal features of strength, clustering coefficient and eigenvector centrality for multimodal emotion recognition. Through experiments, author's inferred strength feature has higher correlation to emotions compared to other modalities. This strength feature of EEG and eye movements is passed to Deep Canonical Correlation Analysis Model to classify emotions. The approach was able to achieve average accuracy of 85% for five emotions of happy, neutral, sad, fear and disgust. The EEG connectivity features were not correlated to micro expression detection. Personalization can realized as there is no higher deviation in EEG strength feature across person. The approach is more suitable for controlled environment and it does not accommodate non static characteristics in EEG data acquisition.

Dai et al. [45] proposed a weakly supervised Multi-Task Learning (MTL) for multimodal emotion recognition. MTL was used to solve the relative smaller dataset problem. Class labels predicted for each task is fused to recognize the final emotion. Three tasks of emotion recognition, sentiment analysis and sarcasm recognition are used to recognize emotion. The approach was able to achieve an average accuracy of 71% for four emotions of neutral, happy, sad and angry. Short term context based decision is made possible using LSTM in this work. Due to its task based fusion, personalization is difficult to realize in this approach. The approach is easy to extend for more emotions due to its task based emotion classification. Micro expression detection can be added a task and results from it can be used for fusion.

Lee et al. [46] proposed a multi modal emotion recognition system combining speech and visual features. Bidirectional Encoder Representations from Transformers **(**BERT**)** was trained with the heterogeneous features learnt from speech and visual modalities. The method was able to achieve an average accuracy of 85.49% for four emotions of happy, sad, angry and neutral. Context information was not involved in emotion recognition. But the approach can be extended for micro expression due to it fine grained feature extraction. The training process did not accommodate any provisions for personalization. The emotion coverage is limited but it can be extended due to its fine grained feature extraction.

Noroozi et al. [47] used speech and visual cues for designing a multimodal emotion recognition system. Handcrafted features are extracted from each of the modalities. The handcrafter features extracted from speech were Mel Frequency Cepstrum Coefficient(MFCC), Filter bank energy and prosodic features. The handcrafted features extracted from face were distance between facial landmarks and angles between facial landmarks. Instead of extracting hand crafted features from each frames of video, a novel key frame selection algorithm is proposed. This novel key frame selection algorithm discriminated between the frames based on their visual information. Classifiers in different combination of features are trained and one with higher confidence is selected for emotion prediction. For six basic emotions, the method was able to achieve an accuracy of 95.34%. Due to reduction of frames, the approach lacks sensitivity to micro expressions. Handcrafted features used in this work are limited and there is no correlation to personalization. The approach has no consideration for both short terms and long term contexts. The approach can be easily extended for some emotion classes by modifying the training dataset.

Zhang et al. [48] proposed a feature fusion algorithm based on Discriminative Canonical correlation analysis (DCCA). EEG and physiological features are fused using DCCA and fused features are classified using traditional machine learning classifiers. The method was able to achieve an average accuracy of 67.50 for two emotions of arousal and liking. Context information was not considered for emotion recognition in this work. DCCA can be used for add personalization in feature fusion. The approach is simple and can be extended for wide emotion coverage by modifying the training dataset. The approach is useful for controlled environment and many challenges exist in adapting it to real world scenarios.

Li et al. [49] proposed a multimodal emotion classification system using multi channel EEG. Discrete Wavelet Transform (DWT) features are extracted from each channel of EEG separately and features are classified to emotion using K-nearest neighbor classifier. The method was able to achieve an

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7464
***Article Received**: 21 May 2023 **Revised:** 14 July 2023 **Accepted:** 02 August 2023*

average accuracy of 87% for two different emotions of valence and arousal. Since a fixed window is used for DWT feature extraction, the approach lacks sensitivity to micro expressions. Personalization can be easily added by training with a more elaborate dataset. Context information is not considered for emotion classification. The measurements are made in controlled environment and have many challenges in extending to real world environment. The approach can be extended for more emotion coverage by modifying the training dataset.

Cai et al. [50] used facial and speech features for multimodal emotion recognition. Speech features are extracted using convolutional neural network combining with LSTM. Facial features are extracted using convolutional neural network with multiple small scale kernel convolutions. Feature fusion of speech and facial features is done and the heterogeneous features are used for emotion classification. The approach was able to achieve an average accuracy of 70.24% for four emotion classes of angry, excite, neutral and sad. The approach can be extended for micro expressions due to its use of multiple small scale kernel. The approach lacked context awareness. LSTM was used only for feature extraction. The approach can be easily extended for personalization and larger emotion coverage by modifying the dataset.

The summary of the surveyed solutions are presented in Table 3 and the summary of critical analysis on five requirement factors are presented in Table 4. From the results, it can be seen long term context awareness is the most important issue in addressing the next generation human computer interactions. It is followed by micro expression detection and its localization in macro expressions. This is very important to detect the true emotion in a context.

Table 3 Survey summary

Solution	Features	Classifier	Remarks
Zheng et al [37]	EEG and eye movements	Deep neural networks	Classified three emotions – happy,sad, fear with accuracy of 72.39%. Approach did not address EEG noises due to movement.
Nguyen et al [38]	Spatio temporal features in audio and video streams.	3D CNN	Tested for six different emotions of anger,disgust, fear,happiness,sadness and surprise. Achieved an average accuracy of 89.39. Experimented only with test videos and not benchmarked against real time datasets. Also computation time was not measured. But 3DCNN have higher computation time
Tzirakis et al [39]	Extract CNN features from speech	Resnet + LSTM	Tested for two class of arousal and valence wit average 61.2 accuracy.
Zhang et al [40]	Handcrafted Physiological features from EEG + EMG + GSR + RES	SVM	Tested for two class of valence and arousal with average accuracy of 63.1%
Chen et al [41]	Acoustic and textural features using deep learning	SVM	Tested for two class of valence and arousal with average accuracy of 67.2%
Cimtay et al [42]	Facial cues, skin response and brain wave	Decision tree	Tested for three class of sad, neutral and happy. Average accuracy is 74.2 %
Zhou et al [43]	Deep learning Audio and visual features	Softmax classifier	Tested for four emotional categories happy, sad, angry and neutral. Average accuracy is 63.09%
Wu et al [44]	Three EEG features : strength, clustering coefficient, and eigenvector centrality and eye movement features	Deep Canonical Correlation Analysis Model	Tested for five emotions of five emotions of disgust, fear, sadness, happiness, and neutrality with average accuracy of 83% . System is not robust against noise
Dai et al [45]	Audio and visual features	Weakly supervised multi task learning	Tested for four class of neutral, happy,sad, angry with average accuracy of 59.7%.
Lee et al [46]	Audio,text and visual cues	BERT	Tested for four class of happy,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7464
***Article Received**: 21 May 2023 **Revised:** 14 July 2023 **Accepted:** 02 August 2023*

			angry, sad and neutral with average accuracy of 72%. But they have not proposed any method to remove incongruent cues.
Noroozi et al [47]	MFCC, Filter Bank Energies, prosodic features, Deep learning features from face.	Random Forest	Tested for six basic emotions with a average accuracy of 95.64%. But the experimentation was done only on eNTERFACE'05 dataset.
Zhang et al [48]	Deep learning features from Text, audio and visual information.	Deep canonical correlation analysis	Tested for 3 class of arousal, valence and dominance with average accuracy of 89%. Significant frame selection was not addressed.
Li et al [49]	DFT features from Multi channel EEG in different frequency bands	KNN	Tested for 2 class of valence and arousal for different number of channels. Maximum accuracy of 92% is achievable for 32 channels. As number of channel increased accuracy increases. But the test in done in ideal situations in absence of noise.
Cia et al [50]	CNN+ LSTM for speech features , Multiple small-scale kernel convolution block for visual features	Deep neural network based fusion	Tested for angry, excite, neutral and sad with average accuracy of 70%. Tested only in controlled situations.
Buitelaar et al [58]	Semantic features from text, MFCC Audio features and CNN Facial features	SVM	Tested for arousal and valence with accuracy of 54%.
Nath et al [59]	band power, a frequency-domain feature, from the EEG signals	LSTM	Tested for Valence and arousal with average accuracy of 93% Was tested only against DEAP dataset
Guo et al [60]	Time domain and wavelet features from EEG	SVM and HMM	Tested for valence and arousal with average accuracy of 61%.
Pandey et al [61]	VMD features from EEG signals	Deep neural network	Tested for valence and arousal with average accuracy of 61% against DEAP dataset.
Zheng et al [62]	Entropy features from EEG	Discriminative Graph regularized Extreme Learning Machine	Tested for valence and arousal with average accuracy of 69.67% for DEAP dataset and 91% for SEED dataset
Appriou et al [63]	EEG signals in single band	Filter Bank FgMDM	Tested for valence and arousal with average accuracy of 60% for DEAP dataset.

Table 4 Adaptability to requirements

Solution	R1	R2	R3	R4	R5
Zheng et al [37]	×	×	×	×	×
Nguyen et al [38]	×	×	×	×	×
Tzirakis et al [39]	×	×	×	×	×
Zhang et al [40]	×	×	×	×	×
Chen et al [41]	×	√	×	×	×
Cimtay et al [42]	×	×	×	×	×
Zhou et al [43]	×	√	×	×	×
Wu et al [44]	×	×	×	√	×
Dai et al [45]	×	√	√	×	√
Lee et al [46]	×	√	√	√	√
Noroozi et al [47]	×	×	×	×	√
Zhang et al [48]	×	×	√	√	√
Li et al [49]	×	×	×	√	√
Cia et al [50]	×	√	√	√	√

IV. CRITICAL ANALYSIS ON DATASETS

Zheng et [37], Wu et al [44] used SEED-IV dataset [51]. The data was collected from 15 participants with data collected from participants for three days at three different times. The data was collected in a controlled environment with participants aware of data collection. The data was collected by showing films with

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7464
***Article Received**: 21 May 2023 **Revised:** 14 July 2023 **Accepted:** 02 August 2023*

higher emotional contents to participants and their response in terms of EEG and eye movements were collected. But the emotions triggered in person were different between first watching and subsequent watching many times. These factors were not considered. Also there can be differences in emotion levels by each person. These factors too were not considered in the dataset. Context of how the user historically responded to these types of emotional contents too was not considered in testing. The dataset was also limited to four emotion class of four emotions: happy, sad, fear and neutral.

Nguyen et al [38], Noroozi et al [47] used eNTERFACE audio-visual database[52]. Speech and video modalities were collected from 44 participants. A total of 1166 video sequences were collected. The data collected six basic emotions of the participants. The dataset has no provisions for micro expressions and more emotion class coverage. Though the dataset considered gender differences, it does not have more user annotations like age, race etc for testing personalization.

Tzirakis et al [40] used REmote COLlaborative and Affective (RECOLA) database. The dataset has four modalities of speech, video, electro dermal activity and ECG. The multimodal was collected for duration of nine and half hours from forty six participants. The collected data were annotated every five minutes. Data collection was done when participants were in video conference. The dataset also considered demographics. The subjects were distributed to French, German and Italian across gender and age. The data is comprehensive for testing personalization but it covered only two emotion class of valence and arousal. Context and user personality were not considered in the dataset.

Zhang et al [40], Zhang et al [48], Li et al [49] used DEAP[53] and DECAFE [54] datasets. The DEAP dataset was collected from thirty two participants by making the participants to watch the video clips. As participants watched the videos, their facial expression and physiological responses were collected. DECAF data was collected in similar lines by making thirty participants to watch thirty six video clips. As the participants watched video clips, their physiological responses and near infrared facial capture were obtained. These signals are tagged. As in SEED-IV dataset, these datasets too did not consider the emotion trigger difference between first watching and subsequent watching many times. These factors were not considered. Also there can be differences in emotion levels by each person. These factors too were not considered in the dataset. Context of how the user historically responded to these types of emotional contents too was not considered in testing.

Chen et al [41], Zhou et al [43], Dai et al [45] used IEMOCAP dataset [55] for their experimentation. The participants were actors. The data was collected when the actors performed scenarios in scripts. Twelve hour recording were made. The data collected included video, speech and text transcriptions. The dataset is annotated with emotion label. Labeling was done against four emotions of angry, sad, neutral and happy. Though the dataset is diverse and detailed, it lacks context information and micro expressions. The dataset is limited to six basic emotions.

Cimtay et al [42] used Loughborough University Multimodal Emotion Database-2 (LUMED-2) [56] for experimentation. The audio visual stimuli were collected from 13 participants by making them watch emotion triggering video sequences. Stimuli were collected for about nine minutes. Emotion triggering video sequences were collected form web sources. The dataset covered only three basic emotions and it lacked any personalization categories. The observations were done in controlled environment and there were no facilities to test robustness for real world scenarios.

Lee et al [46] used CMU-MOSI dataset [57] for experimentation. It is a collection of 2199 opinion video clips with annotation for each video in range of -3 to +3. The audio and video features were annotated per milliseconds. The dataset is more useful for micro expression due to per milliseconds annotations. Due to it features of subjectivity and sentiment intensity, the dataset can be used for testing personalization.

The summary of analysis of dataset suitability to test the requirements of next generation human computer interactions is listed in Table 5. From the results, it can be seen that CMU-MOSI appears to be most promising to test the requirements of next generation human computer interactions compared to other datasets. But CMU-MOSI does not have context information. Thus there is a necessity to create datasets similar to CMU-MOSI augmenting with environment context on conversation and user past personality profiles. Augmenting user past personality profiles in the data and using it emotion recognition, will improves the accuracy of emotion recognition in real world scenarios.

Table 5 Dataset adaptability for testing requirements

Dataset	R1	R2	R3	R4	R5
SEED-IV	×	×	×	×	×
eNTERFACE	×	×	×	×	×
RECOLA	×	√	√	√	×
DECAFE	×	×	×	×	×
DEAP	×	×	×	×	×
IEMOCAP	×	√	×	×	×
LUMED-2	×	√	×	×	×
CMU-MOSI	×	√	√	√	√

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7464
***Article Received**: 21 May 2023 **Revised:** 14 July 2023 **Accepted:** 02 August 2023*

V. DISCUSSION

Most of the current multimodal emotion recognition solution fit into model as shown in Figure 3.

In this model, the features are extracted in isolation from each modality and joint feature encoding from multiple modalities is done using spatio temporal correlation. Either feature level fusion followed by emotion classification or classification for each modality followed by decision level fusion is done to provide the final emotion classification result. But this model needs to be extended to address the requirements of next generation human computer interactions.

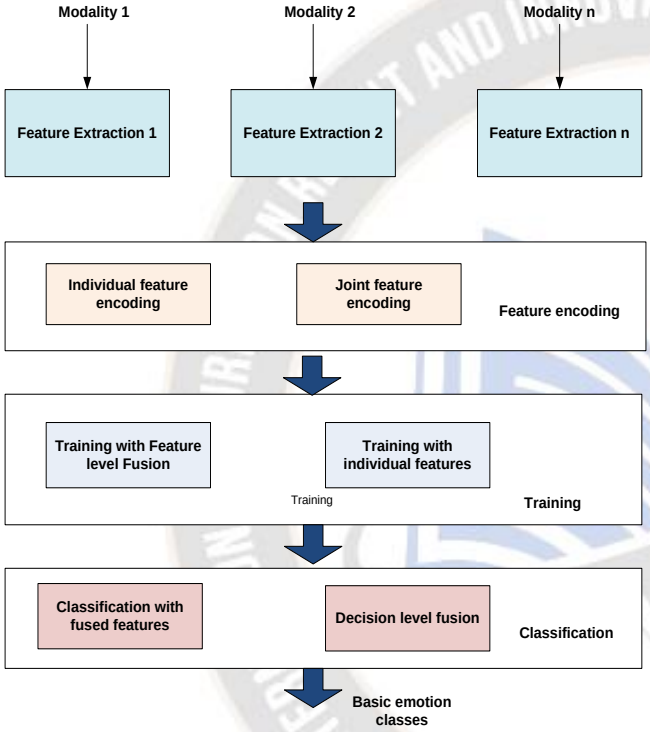

Figure 3 Generalized architecture of existing multi domain emotion recognition models

A possible extension model recommended by this study is given in Figure 4. In this model, the input modalities are pre-processed to remove noises, impedance variability and non static characteristics in the inputs. This is done to make the classification to be robust for real world scenarios. This is followed by windowing analysis to identify outliers occurring as discontinuity and this discontinuity duration should be used to mark the window boundaries. A variable windowing must be done on input modalities. This differs from current approaches of fixed window without consideration for the dynamics in input modalities. This windowing analysis helps to detect micro expressions. In process of training, adversarial learning must be accommodated to personalize the classification model. The current solutions are more on generic classification model. The classification must also accommodate context information. The context will be both short term: learnt from continuous observation of input modalities and long term based on knowledge fused about environment and the observer.

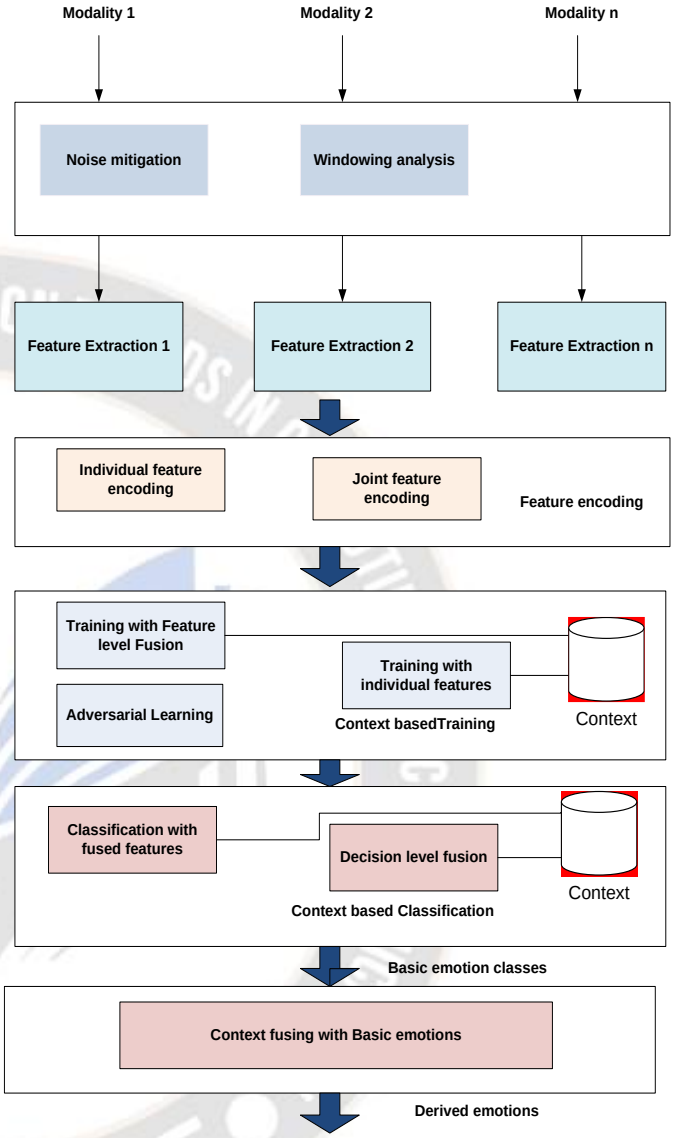

Figure 4 Recommended architecture for multimodal emotion recognition

By this way context dependent emotion classification can be realized which is accurate and more suited for next generation human computer interactions. To increase the emotion coverage classes' two strategies can be followed: first is training the model with datasets having multiple coverage classes or correlating the basic emotions with context information to predict derived emotions as in Plutchik wheel of emotions.

VI. CONCLUSION

A critical analysis on multimodal emotion recognition systems both in terms of methodology and dataset is presented in this work. The survey identified five critical requirements on multimodal emotion recognition systems by next generation human computer interactions. Critical analysis was done

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7464
***Article Received**: 21 May 2023 **Revised:** 14 July 2023 **Accepted:** 02 August 2023*

against these requirements. The survey found a large gap in existing methodologies and datasets in addressing these requirements and presented recommendations addressing the same.

REFERENCES

[1] Mehrabian, A., Ferris, S.R.: Inference of attitudes from nonverbal communication in two channels. J. Consult. Psychol. 31(3), 248 (1967)

[2] Knapp ML, Hall JA. Nonverbal Communication in Human Interaction. 6th. Belmont, CA: Wadsworth; 2005

[3] Mood Ring Monitors Your State of Mind, Chicago Tribune, 8 October 1975, at C1: Ring Buyers Warm Up to Quartz Jewelry That Is Said to Reflect Their Emotions. The Wall Street Journal, 14 October 1975, at p. 16; and "A Ring Around the Mood Market", The Washington Post, 24 November 1975, at B9

[4] Picard, R.W., Vyzas, E., Healey, J.: Toward machine emotional intelligence: analysis of affective physiological state. IEEE Trans. Pattern Anal. Mach. Intell. 23(10), 1175–1191 (2001)

[5] Garcia-Garcia, Jose & Penichet, Victor & Lozano, María. (2017). Emotion detection: a technology review. 1-8. 10.1145/3123818.3123852.

[6] Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: a survey. Ain Shams Eng. J. **5**(4), 1093–1113 (2014)

[7] Ravi, K., Ravi, V.: A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowl.-Based Syst. **89**, 14–46 (2015)

[8] Zucco, C., Calabrese, B., Agapito, G., Guzzi, P.H., Cannataro, M.: Sentiment analysis for mining texts and social networks data: Methods and tools. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, pp. e1333 (2019)

[9] Ekman, P., Wallace, V.: Unmasking the Face. Malor Book, Cambridge (2003)

[10] Friesen, W.V., Ekman, P.: Emfacs-7: Emotional facial action coding system. Unpublished manuscript, University of California at San Francisco

[11] Mozziconacci, S.J.L.: Modeling emotion and attitude in speech by means of perceptually based parameter values. User Modeling and User-Adapted Interaction **11**(4), 297–326 (2001)

[12] Murray, I.R., Arnott, J.L.: Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. J. Acoust. Soc. Am. **93**(2), 1097–1108 (1993)

[13] Erika H Siegel, Molly K Sands, Wim Van den Noortgate, Paul Condon, Yale Chang, Jennifer Dy, Karen S Quigley, and Lisa Feldman Barrett. 2018. Emotion fingerprints or emotion populations? A meta-analytic investigation of autonomic features of emotion categories. Psychological bulletin 144, 4 (2018), 343

[14] Sylvia D Kreibig. 2010. Autonomic nervous system activity in emotion: A review. Biological psychology 84, 3 (2010), 394–421.

[15] Shu, L., et al.: A review of emotion recognition using physiological signals. Sensors (Basel) 18(7), 2074 ,2018

[16] Kessous, L., Castellano, G. and Caridakis, G. 2010. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. Journal on Multimodal User Interfaces, 3, 33-48.

[17] Paleari, M., Benmokhtar, R. and Huet, B. 2009. Evidence theory-based multimodal emotion recognition. In Proceedings of Proceedings of the 15th International Multimedia Modeling Conference (MMM '09) (Chongqing, China, January 6-8, 2009).Springer-Verlag, Berlin, Heidelberg, 435-446.

[18] Wöllmer, M., Metallinou, A., Eyben, F., Schuller, B. and Narayanan, S.S. 2010. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling. In Proceedings of Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010) (Makuhari, Japan, September 26-30, 2010). 2362-2365.

[19] Lin, J., Wu, C. and Wei, W. 2012. Error Weighted SemiCoupled Hidden Markov Model for Audio-Visual Emotion Recognition. IEEE Transactions on Multimedia, 14, 142 - 156.

[20] Jiang, D., Cui, Y., Zhang, X., Fan, P., Ganzalez, I. and Sahli, H. 2011. Audio visual emotion recognition based on triplestream dynamic bayesian network models. In Proceedings of Fourth International Conference on Affective Computing and Intelligent Interaction (Memphis TN, October 9-12, 2011).Springer-Verlag, Berlin Heidelberg, 609-618

[21] Wei, W., Jia, Q. X., Feng, Y. L., Chen, G., and Chu, M. (2020). Multi-modal facial expression feature based on deep-neural networks. J. Multimod. User Interfaces 14, 17–23

[22] Zhang, J. H., Yin, Z., Cheng, P., and Nichele, S. (2020). Emotion recognition using multi-modal data and machine learning techniques: a tutorial and review. Inform. Fus. 59, 103–126.

[23] Mou, W., Gunes, H., and Patras, I. (2019). Alone versus in-a-group: a multi-modal framework for automatic affect recognition. ACM Trans. Multimed. Comput. Commun. Appl. 15, 1–23.

[24] Zhao, S. C., Gholaminejad, A., Ding, G., Gao, Y., Han, J. G., and Keutzer, K. (2019). Personalized emotion recognition by personality-aware high-order learning of physiological signals. ACM Trans. Multimed. Comput. Commun. Appl. 15, 1–18

[25] Huddar, M. G., Sannakki, S. S., and Rajpurohit, V. S. (2020). Multi-level context extraction and attention-based contextual inter-modal fusion for multimodal sentiment analysis and emotion classification. Int. J. Multimed. Inform. Retriev. 9, 103–112.

[26] Lovejit, S., Sarbjeet, S., and Naveen, A. (2019). Improved TOPSIS method for peak frame selection in audio-video human emotion recognition. Multimed. Tools Appl. 78, 6277–6308

[27] Poria, Soujanya & Cambria, Erik & Bajpai, Rajiv & Hussain, Amir. (2017). A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion. Information Fusion. 37. 10.1016/j.inffus.2017.02.003.

[28] X. Gu, Y. Shen and J. Xu, "Multimodal Emotion Recognition in Deep Learning:a Survey," in 2021 International Conference on Culture-oriented Science & Technology (ICCST), Beijing, China, 2021 pp. 77-82.

[29] Sharma, G., & Dhall, A. (2021). A survey on automatic multimodal emotion recognition in the wild. In G. Phillips-Wren, A. Esposito, & L. C. Jain (Eds.), *Advances in Data Science: Methodologies and Applications* (pp. 35-64). (Intelligent Systems Reference Library; Vol. 189). Springer

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7464
***Article Received**: 21 May 2023 **Revised:** 14 July 2023 **Accepted:** 02 August 2023*

[30] Zhang, Tao; Tan, Zhenhua (2021): Deep Emotion Recognition using Facial, Speech and Textual Cues: A Survey. TechRxiv

[31] Devaram, Sarada. (2020). Empathic Chatbot: Emotional Intelligence for Mental Health Well-being. 10.13140/RG.2.2.16077.46564.

[32] Kapil Sharma, Rajiv Khosla, Yogesh Kumar. (2023). Application of Morgan and Krejcie & Chi-Square Test with Operational Errands Approach for Measuring Customers' Attitude & Perceived Risks Towards Online Buying. International Journal of Intelligent Systems and Applications in Engineering, 11(3s), 280–285. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2685

[33] Frank, M., Herbasz, M., Sinuk, K., Keller, A., and Nolan, C. (2009). "I see how you feel: training laypeople and professionals to recognize fleeting emotions," in The Annual Meeting of the International Communication Association (New York City, NY: Sheraton New York.

[34] Ekman, P., and Friesen, W. V. (1969). Nonverbal leakage and clues to deception. Psychiatry 32, 88–106.

[35] Stephan, Hamann, , , Turhan, and Canli, "Individual differences in emotion processing," Current Opinion in Neurobiology, 2004

[36] U. Hess, C. Blaison, and K. Kafetsios, "Judging facial emotion expressions in context: The influence of culture and self-construal orientation," Journal of Nonverbal Behavior, vol. 40, no. 1, pp. 55– 64, 2016.

[37] Wikipedia. (March 27th, 2019). "Robert Plutchik" from: https://es.wikipedia.org/wiki/Robert_Plutchik

[38] W. -L. Zheng, W. Liu, Y. Lu, B. -L. Lu and A. Cichocki, "EmotionMeter: A Multimodal Framework for Recognizing Human Emotions," in IEEE Transactions on Cybernetics, vol. 49, no. 3, pp. 1110-1122, March 2019

[39] D. Nguyen, K. Nguyen, S. Sridharan, A. Ghasemi, D. Dean, and C. Fookes, "Deep spatiotemporal features for multimodal emotion recognition," in Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on, pages 1215–1223. IEEE, 2017

[40] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks," in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1301-1309, Dec. 2017,

[41] Zhang, Xiaowei, Jinyong Liu, JianShen, Shaojie Li, KechenHou, Bin Hu, Jin Gao, and Tong Zhang. "Emotion recognition from multimodal physiological signals using a regularized deep fusion of kernel machine." IEEE transactions on cybernetics (2020).

[42] B. Chen, Q. Cao, M. Hou, Z. Zhang, G. Lu and D. Zhang, "Multimodal Emotion Recognition With Temporal and Semantic Consistency," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3592-3603, 2021,

[43] Y. Cimtay, E. Ekmekcioglu and S. Caglar-Ozhan, "Cross-Subject Multimodal Emotion Recognition Based on Hybrid Fusion," in IEEE Access, vol. 8, pp. 168865-168878, 2020

[44] Zhou, Hengshun, Jun Du, Yuanyuan Zhang, Qing Wang, Qing-Feng Liu, and Chin-Hui Lee. "Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition." IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021)

[45] Wu, Xun, Wei-Long Zheng, and Bao-Liang Lu. "Investigating EEG-based functional connectivity patterns for multimodal emotion recognition." arXiv preprint arXiv:2004.01973 (2020).

[46] Dai, Wenliang, Samuel Cahyawijaya, Yejin Bang, and Pascale Fung. "Weakly supervised Multi-task Learning for Multimodal Affect Recognition." arXiv preprint arXiv:2104.11560 (2021).

[47] Lee, Sanghyun, David K. Han, and HanseokKo. "Multimodal Emotion Recognition Fusion Analysis Adapting BERT With Heterogeneous Feature Unification." IEEE Access 9 (2021)

[48] Noroozi, Fatemeh, Marina Marjanovic, Angelina Njegus, Sergio Escalera, and GholamrezaAnbarjafari. "Audio-visual emotion recognition in video clips." IEEE Transactions on Affective Computing 10, no. 1 (2017): 60-75

[49] K. Zhang, Y. Li, J. Wang, Z. Wang and X. Li, "Feature Fusion for Multimodal Emotion Recognition Based on Deep Canonical Correlation Analysis," in IEEE Signal Processing Letters, vol. 28, pp. 1898-1902, 2021

[50] Li, Mi, HongpeiXu, Xingwang Liu, and Shengfu Lu. "Emotion recognition from multichannel EEG signals using K-nearest neighbor classification." Technology and health care 26, no. S1 (2018): 509-519

[51] L. Cai, J. Dong and M. Wei, "Multi-Modal Emotion Recognition From Speech and Facial Expression Based on Deep Learning," *2020 Chinese Automation Congress (CAC)*, 2020, pp. 5726-5729

[52] https://bcmi.sjtu.edu.cn/~seed/seed-iv.html

[53] O. Martin, I. Kotsia, B. Macq, and I. Pitas. The eNTERFACE'05 audio-visual emotion database. In Proceedings of the 22Nd International Conference on Data Engineering Workshops, ICDEW '06, pages 8–, Washington, DC, USA, 2006. IEEE Computer Society

[54] S. Koelstra et al., "DEAP: A database for emotion analysis; using physiological signals," IEEE Trans. Affect. Comput., vol. 3, no. 1, pp. 18–31, Jan.–Mar. 2012

[55] Prof. Shweta Jain. (2017). Design and Analysis of Low Power Hybrid Braun Multiplier using Ladner Fischer Adder. International Journal of New Practices in Management and Engineering, 6(03), 07 - 12. https://doi.org/10.17762/ijnpme.v6i03.59

[56] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, "DECAF: MEG-based multimodal database for decoding affective physiological responses," IEEE Trans. Affect. Comput., vol. 6, no. 3, pp. 209–222, Jul.–Sep. 2015

[57] https://sail.usc.edu/iemocap/

[58] LUMED-2 Dataset. [Online]. Available: https:// figshare.com/articles/dataset/Loughborough_University_Multimodal_ Emotion_Dataset_-_2/12644033

[59] http://multicomp.cs.cmu.edu/resources/cmu-mosi-dataset/

[60] Buitelaar, Paul, Ian D. Wood, SapnaNegi, MihaelArcan, John P. McCrae, Andrejs Abele, Cecile Robin et al. "Mixedemotions: An open-source toolbox for multimodal emotion analysis." IEEE Transactions on Multimedia 20, no. 9 (2018): 2454-2465.

[61] Nath, Debarshi, Mrigank Singh, DivyashikhaSethia, DikshaKalra, and S. Indu. "An efficient approach to EEG-based emotion recognition using LSTM network."In 2020 16th IEEE international colloquium on signal processing & its applications (CSPA), pp. 88-92.IEEE, 2020.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7464
***Article Received**: 21 May 2023 **Revised:** 14 July 2023 **Accepted:** 02 August 2023*

[62] Guo, Kairui, Henry Candra, Hairong Yu, Huiqi Li, Hung T. Nguyen, and Steven W. Su. "EEG-based emotion classification using innovative features and combined SVM and HMM classifier."In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 489-492.IEEE, 2017.

[63] Omondi, P., Rosenberg, D., Almeida, G., Soo-min, K., & Kato, Y. A Comparative Analysis of Deep Learning Models for Image Classification. Kuwait Journal of Machine Learning, 1(3). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/128

[64] Pandey, Pallavi, and K. R. Seeja. "Emotional state recognition with eeg signals using subject independent approach." In Data Science and Big Data Analytics, pp. 117-1 Springer, Singapore, 2019.

[65] W. Zheng, J. Zhu and B. Lu, "Identifying Stable Patterns over Time for Emotion Recognition from EEG," in IEEE Transactions on Affective Computing, vol. 10, no. 3, pp. 417-429, 1 July-Sept. 2019, doi: 10.1109/TAFFC.2017.2712143.

[66] Appriou, Aurélien, AndrzejCichocki, and Fabien Lotte. "Modern machine-learning algorithms: for classifying cognitive and affective states from electroencephalography signals." IEEE Systems, Man, and Cybernetics Magazine 6, no. 3 (2020): 29-38.

[67] H. Chanyal, R. K. . Yadav, and D. K. J. Saini, “Classification of Medicinal Plants Leaves Using Deep Learning Technique: A Review”, International Journal of Intelligent Systems and Applications in Engineering, vol. 10, no. 4, pp. 78–87, Dec. 2022

[68] Sharma, P. ., R. K. Yadav, and D. J. B.Saini. “A Survey on the State of Art Approaches for Disease Detection in Plants”. International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 11, Nov. 2022, pp. 14-21, doi:10.17762/ijritcc.v10i11.5774.

[69] Oktavia, Nur Yusuf, Adhi Dharma Wibawa, EviSentiana Pane, and MauridhiHeryPurnomo. "Human emotion classification based on EEG signals using Naïve bayes method." In 2019 International Seminar on Application for Technology of Information and Communication (iSemantic), pp. 319-324. IEEE, 2019