ENHANCING EMOTION RECOGNITION WITH MULTIMODEL APPROACH USING DEEP NEURAL NETWORKS

Dr. Komal Anadkat[1], Ayush Solanki[2], Dhruva Patel[3], Vraj Thakkar[4]

•

[1]Assistant Professor, Information Technology department, G.E.C, Gandhinagar, India.
komalanadkat@gecg28.ac.in
[2]Student, Information Technology, G.E.C, Gandhinagar, India.
ayush17solanki@gecg28.ac.in
[3]Student, Computer Engineering, G.E.C, Gandhinagar, India.
dhruvapatel1210@gmail.com
[4]Student, Information and Communication Technology, DAIICT, Gandhinagar,
India. vrajthakkar03@gmail.com

Abstract

Recognizing and extracting different emotions, and then validating those emotions have become important for enhancing human-computer interaction. Emotions play a crucial role in social interactions, facilitating rational decision-making and perception. Previously researched emotion recognition models have typically focused on a single input type like images, text, or audio, where each model can identify the emotion of a person through a single source like facial expressions, voice, social media posts, etc. However, these uni-model approaches are limited because they rely on just one type of data, which often misses the full range of emotional cues. To overcome these limitations, multi-model emotion recognition techniques are proposed which are useful for detecting emotions through a person's facial expressions, speech, social media status, and then EEG data. Model fusion techniques have been applied to detect the most accurate emotion for a particular person through fusion of all the models. A recognition rate-based weighting approach is proposed for model fusion, wherein models are assigned weights proportional to their individual recognition rates. This approach enhances overall performance by combining the outputs of various models with higher emphasis on those with better accuracy. The decision fusion-based multi- model emotion recognition model is proposed which achieved a maximum of 87%. accuracy using a bi-model approach and 92% accuracy with a tri-model approach. The weighted decision fusion approach assigns more weight to the model which is more accurate and achieved 93% accuracy. The proposed recognition rate-based weighting approach for fusion has provided significant results, achieving approximately 93% accuracy with 0.900 and 0.904 Cohen kappa and Mathew score respectively using facial expression, speech, and social media text modalities on combined dataset. The proposed model achieved 63% accuracy on a real-world collected dataset without considering EEG data and improved to 73% if EEG is also considered.

Keywords: *Multimodel Fusion, Emotion recognition, Deep Learning,EEG*

I. Introduction

The absence of emotional cues in individual models, coupled with their susceptibility to external influences, often results in reduced accuracy for emotion recognition. Human communication and emotional expression are inherently multimodal, involving the concurrent use of textual, auditory, and visual cues. The primary advantages of multi-model emotion recognition (MER) include a reduction in the total error rate of classification and enhanced overall model accuracy. Additionally, MER is less vulnerable than the single model to external factors, so it is quite robust, and it also addresses missing modality problems. Among the different fusion techniques, decision- level fusion is applied here, where input from every modality is modeled separately and the final uni-model affect recognition is integrated. It permits the use of the best classifier for each model since different predictors have greater flexibility and can better represent each modality. In situations where one or more modalities are absent, it makes prediction easier and even permits training in the absence of parallel data. Multimodal emotion recognition techniques are attractive for a variety of reasons. First of all, voice, body, and face are all viewed holistically in real life, where human emotion recognition occurs in a multi-model environment. When attempting to teach a computer to reproduce elements of human emotional intelligence, it seems appropriate to teach them to utilize the same approach. Secondly, integrating multiple-affective signals enriches the data collection. The effect of uncertainty in the raw data will be lessened when other modalities are combined to infer mood. Finally, the ability to identify emotions becomes more flexible with richer data gathering, especially in cases where one or more source signals are absent. Put differently, the information from the remaining modality can serve as a supplement for the emotion categorization job when one modality contains limited emotional information. Multimodel fusion, the process of combining data from various modalities to produce a single effect classification result, is required in multi-model emotion identification techniques. In terms of multi-model fusion, the literature focuses on two different kinds of fusion techniques: decision level fusion and fusion-level fusion. We shall outline the main concepts and general principles of the two multi-model fusion techniques in the subsection that follows. These approaches are critical for combining cues from multiple modalities to generate more robust predictions. Numerous experts in the field have examined this subject in detail, leading to various classifications of fusion techniques. However, earlier surveys adhere to the subsequent classification.

I. Early Fusion

Before passing the joint representation through a model, feature-level fusion, also known as early fusion, concatenates the features from various modalities. Finding the most effective approach to concatenate features that can improve emotion identification performance is the aim of feature level fusion. A straightforward concatenation of the modalities has been applied for feature-level fusion in a number of successful applications. The primary benefit is the ease with which correlation between modalities may be utilized. However, syncing features from distinct modalities can be challenging and computationally expensive because they sometimes have different forms. As a result, feature-level fusion’s benefits could occasionally be restricted.

II. Late Fusion

Decision-level fusion, also known as late fusion, is a fusion strategy in which the outputs from each classifier are combined after independent classifiers for each modality are used and trained. The primary benefit of decision-level fusion is that decisions have a common format, making it easier to fuse them together. The ultimate prediction is derived from the combination of two uni-model classifiers. The synchronization problems encountered during early fusion are thereby avoided. Additionally, applying the best classifiers appropriate for each modality is made possible by decision-level fusion, giving the classification step greater flexibility. The following categories comprise the most common late fusion techniques for emotion recognition. The maximum of all posterior probabilities is chosen using the maximum rule.

- Maximum rule: selects the maximum of all posterior probabilities.

- Sum rule: sums probabilities from each classifier and then picks the class with the highest value.
- Product rule: multiplies probabilities between classifiers and then chooses the class with the largest value.
- Weight criterion: results in a linear combination of the classifier's output, where the constants are confidence rates of the predictors.
- Rule-based: selects a dominant modality for each class.
- Model-based: employs a machine-learning algorithm to fuse the output of the classifiers.

Consider a situation where four classes (A, B, C, and D) need to be classified using data from two modalities. Due to the late fusion strategy, each modality is trained using two distinct classifiers. Thus, it is anticipated that the maximum rule fusion system will receive two prediction vectors as input. The system will return the maximum value for each class, as shown in the figure, and choose the class with the greatest value as the winner. Keep in mind that the output could be normalized at the end to make the probability of the whole equal.

III. Hybrid Fusion

By merging the outputs of the early fusion process with the individual uni-model predictors, this strategy aims to combine the best aspects of both fusion techniques. This method only makes sense, though, when more than two modalities are being used. In a scenario where there are three modalities—audio, video, and MRI (Magnetic Resonance Imaging) —for example, the features from the audio and video could be concatenated and used to train a classifier (early fusion), and the MRI features could be used to train another predictor, which would then be used to fuse the output from both classifiers (late fusion).

IV. Cohen's Kappa and Matthews Correlation Coefficient (MCC)

A common statistic for evaluating the degree of agreement between two raters is Cohen's kappa. It can also be applied to evaluate a classification model's performance. Similar to accuracy, Cohen's kappa assesses the agreement between the target and anticipated class, but it additionally accounts for the random probability of receiving the predictions. Cohen Kappa has been adopted by the machine learning community as a means of comparing classifier performance.
It is calculated with the following formula:

$$Kp \ = \ (P0 - Pe)/\,(1 - Pe) \tag{1}$$

The measure of agreement between the model predictions and the actual class values as if they happened by chance is called Pe, and P0 is the model's overall accuracy. By eliminating the potential for agreement between the classifier and a random guess, Cohen's kappa calculates the proportion of predictions the classifier makes that are not consistent with a random guess. A high score is only obtained if the prediction performed well in each of the four confusion matrix categories (TP, FN, TN, and FP), proportionately to the size of both positive and negative elements in the dataset. Matthews Correlation Coefficient (MCC) is a statistical rate that ranges from -1 to 1. With a few modifications to the equation, the performance metric—which was initially designed for binary classification—has been extended to the multi-class scenario.

$$MCC \ = \ \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}} \tag{2}$$

II. Related Work

Recognition and extracting various emotions and then validating those emotions have become important for improving the overall human computer interaction. Emotions play an important role in social interactions and facility rational decision making and perception. To achieve specific objectives of the research work, the researcher has referred various research articles and papers regarding emotion recognition system. It was found that there are broadly four approaches used for recognizing human emotions. The first is by using facial expressions, the second is using speech samples of different people, and the third approach is using social media text and last on is using EEG signals. This highlights the critical role of multimodal fusion in improving emotion recognition accuracy. Extensive research has been con- ducted in this domain. Chao et al. [1] aimed to predict continuous values of emotional dimensions, such as arousal and valence, using audio, visual, and physiological data. They employed an LSTM-RNN (long short-term memory recurrent neural network) on the RECOLA dataset. Samira E. Kahou [2] used DBNs (Deep Belief Networks) and CNNs (Convolutional Neural Networks) with K-Means and Relational Auto-Encoders to detect emotions in videos, where individual audio and video models were trained, followed by a decision fusion method for emotion classification. An ensemble approach was adopted by [3], where the output features of a CNN were combined with those of a ResNet and then fed into an LSTM network. Sahay et al. [4] proposed a Relational Tensor Network architecture that modeled inter-modal interactions within a segment, as well as interactions between segments in a video. Hazarika, Poria, et al. [5] developed a framework that utilized a multimodal approach, incorporating visual, audio, and textual features having a GRU to model past utterances of each speaker. These utterances were then integrated by utilizing attention-based hops to capture inter-speaker dependencies. Hassan [6] introduced an unsupervised deep belief network (DBN) for extracting deep-level features from fused sensor signals such as Electro-Dermal Activity (EDA), Photoplethysmogram (PPG), and Zygomaticus Electromyography (zEMG). Five fundamental emotions were classified using the feature-fusion vector created by combining the statistical characteristics of EDA, PPG, zEMG, and DBN features. These emotions were classified as Happy, Relaxed, Disgust, Sad, and Neutral. The use of a pre-trained 'BERT-like' architecture for self-supervised learning to rep- resent both language and text modalities in the recognition of multimodal language emotions was investigated by Siriwardhana et al. [7]. They demonstrated that a simple fusion mechanism (Shallow-Fusion) could simplify the overall architecture while enhancing the effectiveness of complex fusion methods. Priyasad et al. [8] presented a deep-learning approach for encoding emotion characteristics. They used band-pass filtering methods, neural networks, and a SincNet layer to extract acoustic properties from unprocessed audio. The band-pass filter output was then fed into a DCNN. A bidirectional recurrent neural network generated a set of representations at the N-gram level first, and then another recurrent neural network using cross-attention generated a set of representations before merging them into a final score. Mittal [9] used cues from several co-occurring modalities—such as audio, text, and face—to improve each modality's robustness against sensor noise. Their MER model unveiled a brand-new, data-driven multiplicative fusion technique that discovered which cues are more trust- worthy and which ones should be suppressed on a sample-by-sample basis. Lastly, Njoku [10] examined how well deep learning-based models performed for multimodal emotion recognition and data fusion.

III. Proposed Approach

I. Decision Fusion Approach

Figure 1 shows the proposed approach of multimodel emotion recognition. With n being the number of emotion categories (n = 3), let W be a linear transformation square matrix of order n. Different weight scenarios result from different values for W. W is an identity matrix of order n, meaning that there is no weight. Several approaches are available for classifying objects when the decision fusion method is applied. Figure 2 displays the confusion matrix for the multi-model test dataset's facial expression, voice, social media text, and EEG uni-models. Then, a total of six bi-model emotion detection models have been tested utilizing a straightforward decision fusion approach. The accuracy report for each bi-model emotion recognition is displayed in Table 2. It is evident that, in the absence of EEG, the bi-model approach attains 86% to 88% accuracy and good values for Cohen kappa and Matthews scores. Four tri-modal emotion recognition models have been evaluated, once more combining the decisions of three models using decision fusion. Medical equipment and a controlled laboratory setting can be used to gather EEG readings. If the user wishes to assess their emotional condition without visiting a doctor, they can do so by utilizing social media text, speech, and facial expressions. The model can attain 92% accuracy without taking EEG into account. The emotion expression ordering of the trained uni-, bi-, tri-, and multi-model emotion models is displayed in Table 3.

Table 1: *Multi-model test Dataset Description*

	Image Dataset	**Speech dataset**	**Social media text dataset**	**EEG dataset**
	FER2013[25]	Ravdess[26]	sankha1998 [27]	Jordanbird[28]
Total Data	20019	575	2039	2132
X_train	16040	460	1366	1705
X_test	3979	115	673	427

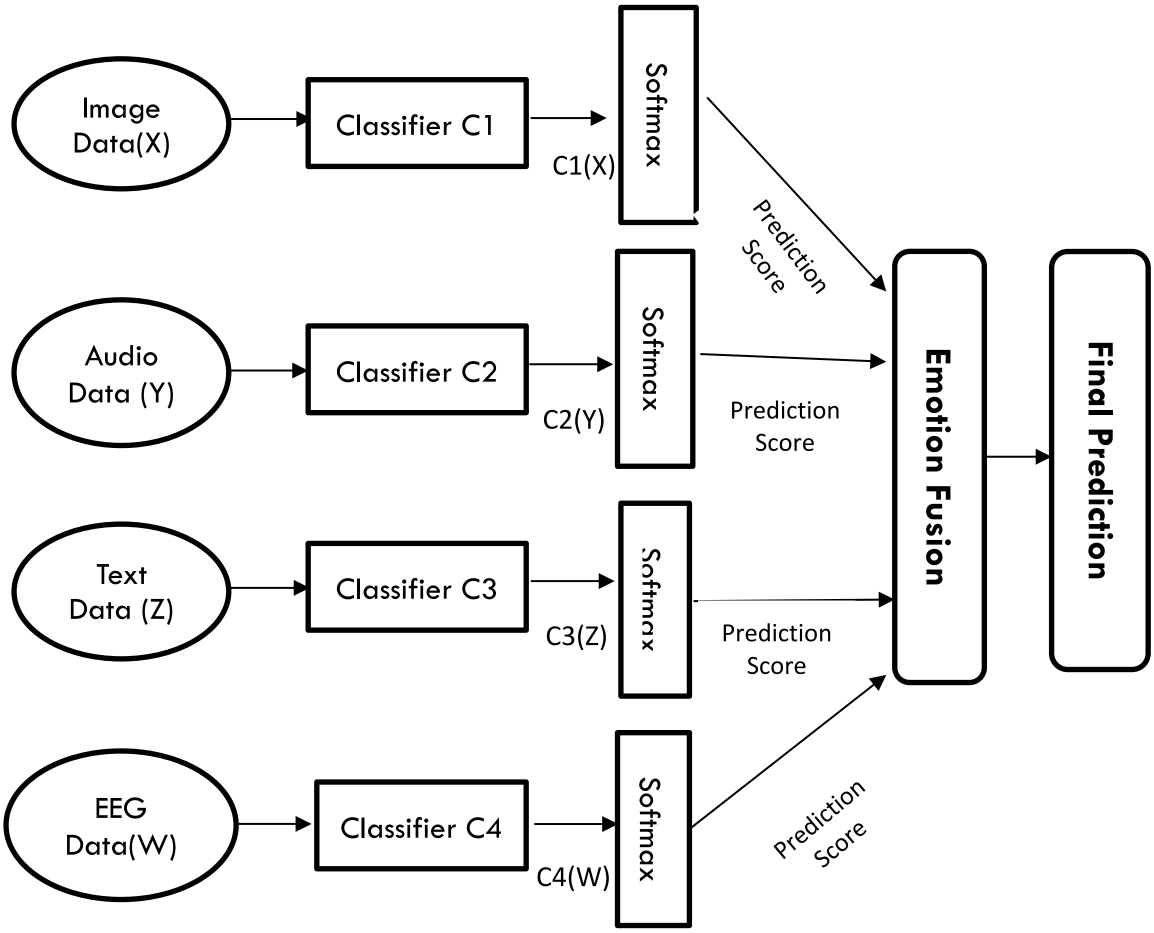

***Figure 1**: Proposed Multi model Fusion Architecture*

Table 2: *Accuracy report of various bi-model emotion recognition*

Bi-model Emotion Recognition (Without EEG)						
	Happy	Sad	Angry	Avg Accuracy	Cohen-Kappa Score	Matthews Score
I+A	0.90	0.86	0.81	0.86	0.783	0.783
I+S	0.78	1.00	0.84	0.87	0.783	0.791
A+S	0.86	0.96	0.83	0.88	0.817	0.820
Bi-model Emotion Recognition (With EEG)						
	Happy	Sad	Angry	Avg accuracy	Cohen-Kappa Score	Matthews Score
E+I	1.00	1.00	1.00	1.00	0.100	0.100
E+A	1.00	0.97	1.00	0.99	0.983	0.984
E+S	1.00	0.97	0.97	0.98	0.967	0.967

Table 3: *Emotion expression ordering of each Emotion model*

	Emotion expression ordering
Image(I)	Angry > Happy > Sad
Audio (A)	Sad > Angry > Happy
Social media text (S)	Sad > Angry > Happy
EEG (E)	Happy = Angry > Sad
I + A	Happy > Sad > Angry
I + S	Sad > Angry > Happy
I + E	Happy = Sad = Angry
A + S	Sad > Happy > Angry
A + E	Happy = Angry > Sad
E + S	Happy > Angry = Sad
I + A + S	Sad > Angry = Happy
I + A + E	Happy = Sad > Angry
A + S + E	Happy > Sad > Angry
I + E + S	Sad > Angry = Happy
I + A + S + E	Happy = Sad > Angry

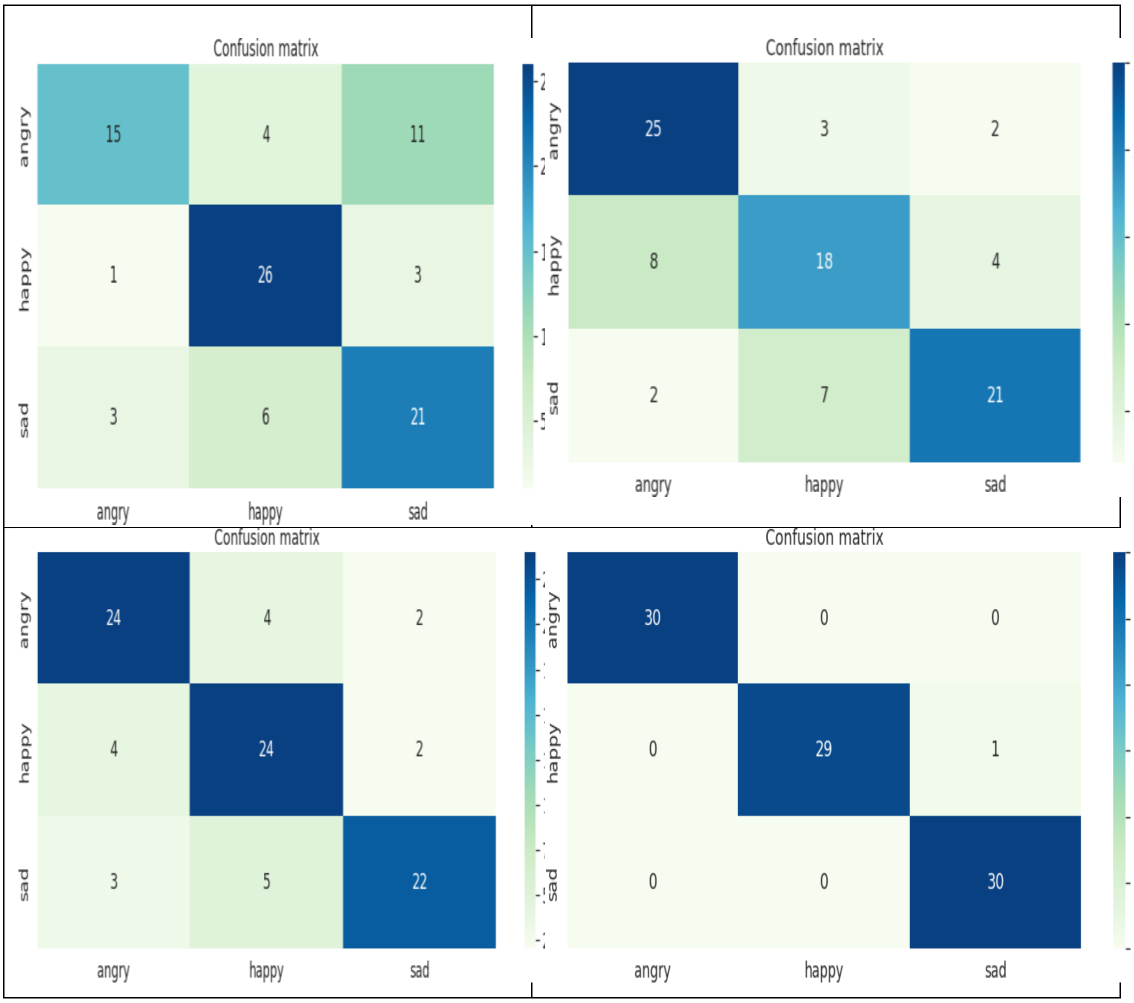

***Figure 2**: Confusion matrix of facial expression, speech, social media text, and EEG model*

II.Weighted averaging Approach

The other approach is to assign different weights to different uni-modals, called Weighted-decision fusion. In decision fusion, equal weights are assigned to each model but if we know that a model is performing better, we can assign higher weightage to that model. In the weighted averaging approach, the accuracy is not assigned as a weight but the prediction of the models which perform better will be multiplied by 2 and the prediction of other models will be multiplied by 1. Since in this particular case, facial expression and speech emotion recognition models are not the best model, they will be assigned lower weights whereas the social media text model is performing better so higher weights will be assigned to that model. Now, basically, these weights will be multiplied by individual predictions and then their mean will be taken. So social media text predictions will multiply by a factor of 2 and others will be multiplied by a factor of 1. Then calculating a weighted average from these models; the model's performance determined the weights.

III. Rank averaging Approach

In this method, the model with the lowest performance is assigned rank 1. Accordingly, the model with rank 1 performs the worst, the model with rank 2 is the next best, and the model with rank 3 is the best. Following the ranking of each of these models, weights are derived from their ranks. Essentially, each rank will be split by the overall value when these ranks have been added up. In this case, the least performance model is the facial expression model, we will divide it by the sum of 1+2+3 = 6. So the weight for social media text models comes down to 1/6 = 0.16. So, all the predicted values of the speech emotion model and social media text model will get multiplied by 0.33 (1/3) and 0.5 (1/2) respectively. And then all these values will be summed up and the final outcome will be taken.

Then Recognition rate as weight Approach is used where Wi (1 ≤ i < n) is the weight of the ith category in W, a diagonal matrix of order n, and not all Wi are equal to each other.

$$W = \begin{matrix} W_1 & & \\ & W_2 & \\ & & W_n \end{matrix} \tag{3}$$

First, for every emotion model, recognition results are derived from four distinct uni-model classifiers. The recognition rates of each classifier (R_{i1}, . . . , R_{im}) are used as a weight matrix W_i. The following is how the linear data fusion concept yields the classifier result

$$C = \sum_{i=1}^{n} C_i W_i \tag{4}$$

The following recognition outcome was achieved using a max-win method [15]:

$$\max_{j=1} m \left\{ \sum_{i=1}^{n} C_{ij} \, R_{ij} \right\} = \sum_{i=1}^{n} C_{ik} \, R_{ik} \tag{5}$$

Where k is the most likely category emotion label. As shown in Table 5, the recognition rate as a weighted approach achieves 93% accuracy and 0.900 and 0.904 values of Cohen kappa and Matthews score respectively. Table 4 shows the accuracy and other measures for all weighted techniques. Figure 3 shows the confusion matrix of the fusion of models. Figure 4 shows the chart of class-wise accuracy v/s different emotion models.

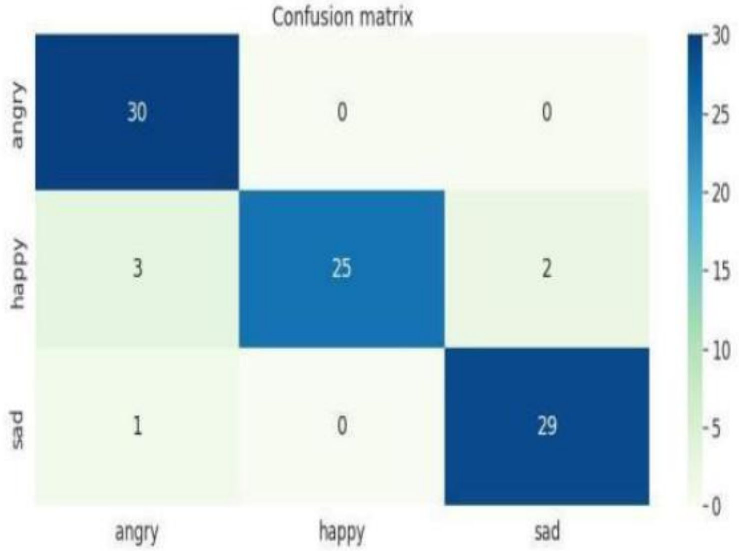

***Figure 3**: confusion matrix of multi-model weighted decision fusion (I+A+S)*

Table 4: *Class-wise precision, recall, f1-score and support of weighted decision fusion*

	Precision	Recall	F1- score	Support
Angry	0.88	1.00	0.94	30
Happy	1.00	0.83	0.91	30
Sad	0.94	0.97	0.95	30
Accuracy			0.93	90
Macro avg	0.94	0.93	0.93	90
Weighted avg	0.94	0.93	0.93	90

Table 5: *Performance of different weighted approaches*

Technique name	Accuracy	Weights(I,A,S)	Cohen-Kappa Score	Matthews Score
Weighted Averaging	92%	(1,1,2)	0.889	0.900
Rank Averaging	90%	(0.16,0.33,0.5)	0.887	0.852
Recognition rate as Weights	93%	(0.70,0.71,0.78)	0.900	0.904

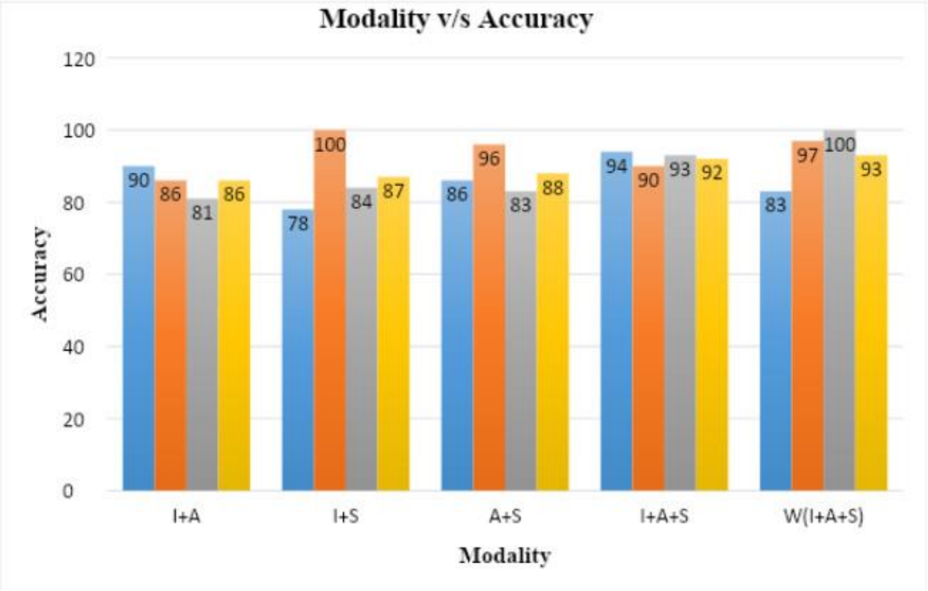

Figure 4: *Chart of accuracy v/s different emotion modalities*

IV. Testing Model with real world collected dataset

As shown in Table 5, from all the fusion approaches, the recognition rate as weights approach gives more accuracy. To test the robustness of the proposed approach, the model is again tested with the real-world collected dataset. The real-world dataset is collected from seven different actors and a total of 168 samples were collected. In this dataset, a total 42 samples are collected from each category. For the facial expression model, two happy, two sad, and two angry samples from each actor are collected. So for the final testing of models, the same number of samples should be selected from each category and from each model. Here, 10 samples are randomly selected from each emotion category from each emotion model, and a total of 30 samples from each emotion model. Table 6 shows the class-wise accuracy of all the uni-model, bi-model, and multi-model combinations. Figure 5 shows the confusion matrix result of the model on a real-world dataset. Figure 6 shows the chart of class-wise accuracy v/s different emotion models when tested on real-world collected datasets. The primary obstacle encountered while utilizing real-world data that has been gathered is the shift in the input or independent variable's data distribution. Although the model's output and correlations may still be technically accurate, the model's predictions have become less accurate due to changes in input data or demography. The major causes of data drift here are Data Quality and Integrity Issues, Demographic Shifts, or Changes in Human Behavior. The Real-time data has been collected using a simple mobile device so the resolution of images, the format of images, the quality of audio samples, and the noise cancellation quality of headphones are not adequate. The Ravdess dataset has been collected in a closed environment, with high-quality devices and the actors used neutral North American accents and they are professional actors. The actors of the Real-time collected dataset are not professional and the accents are totally different. The FER2013 dataset contains only the "pixels" column in .csv format and the images are 48*48 pixels. The real time images are totally in different formats.

The multi-modal emotion recognition model aims to integrate emotions detected from various individual models, resulting in the accurate identification of human emotions. The uni-modal emotion recognition approach faces challenges such as missing modalities and lower accuracy. To address these issues, the data from different modalities needs to be fused and transformed into a consolidated format. This consolidation enables the integration of decisions from various modalities.

Table 6 : Class-wise accuracy of all modalities data when tested model using real world data

Emotion model	Happy	Sad	Angry	Average accuracy
Image(I)	0.30	0.40	0.50	0.40
Audio (A)	0.20	0.50	0.40	0.37
Social media text (S)	0.60	0.50	0.40	0.50
A+S	0.60	0.50	0.60	0.57
I+A	0.50	0.50	0.60	0.53
I+S	0.60	0.40	0.70	0.57
I+A+S	0.60	0.50	0.70	0.60
W(I+A+S)	0.60	0.50	0.80	0.63
EEG(E)	0.70	0.60	0.70	0.67
I+A+S+E	0.70	0.70	0.80	0.73

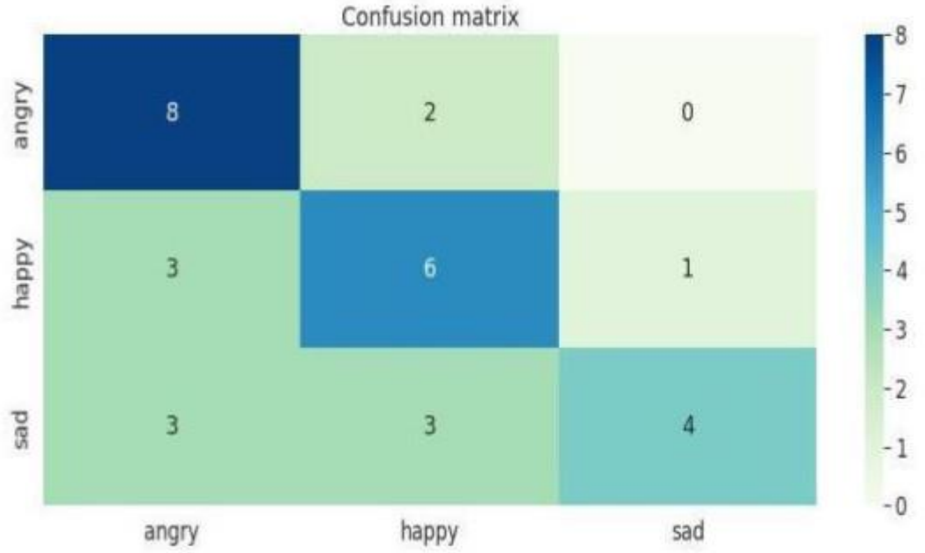

Figure 5: Confusion matrix of Proposed Decision fusion model on collected data (I+A+S)

Table 7: Class-wise precision, recall, f1-score, and support of real-world data of decision fusion (I+A+S)

	Precision	Recall	F1- score	Support
Angry	0.57	0.70	0.67	10
Happy	0.55	0.60	0.57	10
Sad	0.80	0.50	0.53	10
Accuracy			0.60	30
Macro avg	0.64	0.60	0.59	30
Weighted avg	0.64	0.60	0.59	30

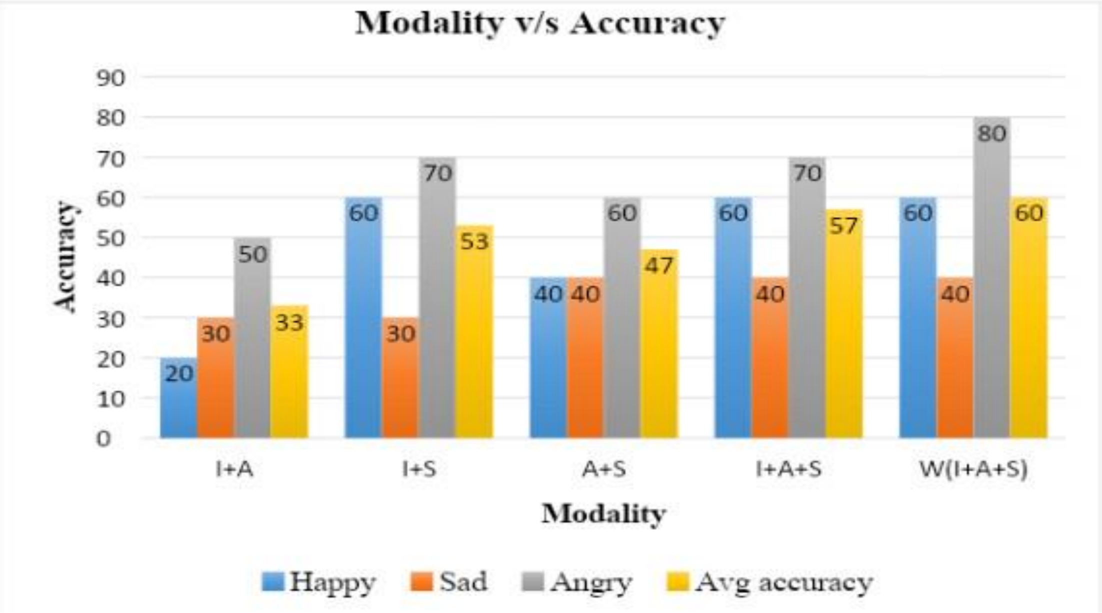

Figure 6: Chart of accuracy v/s different emotion modalities for real-world data

Table 8: Comparison of different existing and proposed multi-model emotion recognition approaches

Ref No	Year	Modality				Fusion	Dataset	Accuracy (in %)
		Image	Speech	Text	EEG			
[1]	2015	√	√		√	Feature level	Recola	66.70
[2]	2015	√	√			Decision Level	Fertfd	47.67
[3]	2017	√	√			Decision Level	Recola	76.00
[4]	2018	√	√	√		Decision Level	Cmu-osei	49.10
[5]	2018	√	√	√		Feature Level	IEM Oca	76.60
[6]	2019				√	Feature Level	Deap	89.53
[7]	2020	√	√	√		Hybrid	Iem Oca	73.98
[8]	2020		√	√		Feature Level	Iem Oca	80.51
[9]	2020	√	√	√		Feature Level	Iem Oca	82.70
[10]	2022	√	√		√	HL, FL, DL	Ravdes	78.75
Proposed Architecture		√	√	√	√	Weighted Decision level	Combined Customized Dataset	93.00
		√	√	√	√		Real world Collected Dataset	67.00

V. Conclusion and Future Work

In conclusion, integrating information across multiple modalities and time holds the potential for enhancing emotion recognition and outcome prediction. The proposed recognition rate based weighting approach for fusion uses the recognition rates of each model as weights. has provided significant results, achieving approximately 93% accuracy with a combined collected dataset with facial expression, speech, and social media text modalities. To test the proposed models, a real-world dataset is collected from seven subjects, encompassing facial expressions, speech, social media text, and EEG signals for three emotions. The collected data is pre processed and formatted for validation. The weighted decision fusion model attained 63% accuracy on the collected real-world dataset with facial expression, speech, and social media text modalities. Challenges with the real-time dataset include lower image resolution, varied image formats, audio quality, and headphone noise cancellation due to the use of a simple mobile device for data collection.

References

[1] Chao, L., Tao, J., Yang, M., Li, Y., Wen, Z (2015).: Long short term memory recurrent neural network based multimodal dimensional emotion recognition. Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge .

[2] Kahou, S.E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, K.,Jean, S., Froumenty, P., Dauphin, Y., Boulanger-Lewandowski, N., Chandias Ferrari, R., Mirza, M., Warde-Farley, D., Courville, A., Vincent, P., Memisevic, R., Pal, C., Bengio, Y.(2015):Emonets: Multimodal deep learning approaches for

emotion recognition in video. Journal on Multimodal User Interfaces 10(2), 99–111.

[3] Tzirakis, P., Trigeorgis, G., Nicolaou, M.A., Schuller, B.W., Zafeiriou, S (2017).: End-to-end multimodal emotion recognition using deep neural networks. IEEE Journal of selected topics in Signal Processing 11(8), 1301–1309 .

[4] Sahay, S., Kumar, S.H., Xia, R., Huang, J., Nachman, L(2018).: Multimodal relational tensor network for sentiment and emotion classification. Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML).

[5] Hazarika, D., Poria, S., Zadeh, A., Cambria, E., Morency, L.-P., Zimmermann, R (2018).: Conversational memory network for emotion recognition in dyadic dialogue videos.Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1.

[6] Hassan, M.M., Alam, M.G.R., Uddin, M.Z., Huda, S., Almogren, A., Fortino, G (2019).:,Human emotion recognition using deep belief network architecture. Information Fusion 51, 10–18 .

[7] Siriwardhana, S., Reis, A., Weerasekera, R., Nanayakkara, S.: Jointly fine-tuning 'bertlike' self supervised models to improve multimodal speech emotion recognition. Interspeech 2020.

[8] Priyasad, S., Fernando, T., Denman, S., Sridharan, S., Fookes, C.: Attention driven fusion for multimodal emotion recognition. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) .

[9] Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., Manocha, D. (2020): M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. Proceedings of the AAAI Conference on Artificial Intelligence 34(02), 1359–1367.

[10] Njoku, J., Caliwag, A.C., Lim, W., Kim, S., Hwang, H.-J., Jeong, J.-W.: Deep learning based data fusion methods for multimodal emotion recognition. The Journal of Korean Institute of Communications and Information Sciences 47(1), 79–87.

[11] Verma, R.: Fer2013. Kaggle (2018). [Online].

[12] Livingstone, S.R.: Ravdess emotional speech audio. Kaggle (2019). [Online].

[13] Mondal, S.S.: Emotion data for whatsapp status. Kaggle (2020). [Online].

[14] Bird, J.: Eeg brainwave dataset: Feeling emotions. Kaggle (2018). [Online].

[15] Jadav, J. , Chauhan, U. : Personalized features-based stress detection with hyperparameter tuning using genetic algorithm.Reliability Theoy and Application,volume 19 ,2024.