International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Multimodal Emotion Recognition: Integrating Audio and Visual Features Using Enhanced Deep Learning Techniques

Archna Kirar[1], Sumeet Gill[2], Vikas Jangra[3],Binny Sharma[4]

Submitted:02/11/2024 **Accepted:**15/12/2024 **Published:**25/12/2024

Abstract: Emotion recognition is a pivotal area in human-computer interaction, crucial for enhancing system responsiveness and adaptability. The expression of human emotion depends on various verbal and non-verbal. Emotion recognition is thus well suited as a multimodal rather than single-modal learning problem. This study introduces Multimodal that integrates speech (Audio) and facial features to recognize three primary emotions: happiness, sadness, and surprise from a video dataset (MELD). In audio feature extraction, an autoencoder is used, which improves the model's capacity to identify subtle emotional subtleties from speech signals. Concurrently, ResNet is used to extract image features by transfer learning, using pre trained weights to identify intricate visual patterns from summary pictures. The Improved Zebra Algorithm (IZA) is used in feature selection to maximize discriminative feature subsets. Our suggested Bi- Directional LSTM with self-attention mechanism is evaluated by comparison with two baseline models, namely Bi Directional LSTM and Convolutional Neural Network (CNN). Our method achieves state-of-art results on MELD. More specifically, the highest accuracy was obtained by the Bi-LSTM-self attention model with 89.83%, followed by 85.15% by the Bi-LSTM, and 86.87% by the CNN respectively. These findings demonstrate the efficiency of the Bi-LSTM- SA model on multimodal emotion recognition.

Keywords and phrases: *Multimodal Emotion Recognition, Bi-Directional LSTM with self-attention mechanism, Bi directional LSTM, Autoencoder, ResNet, CNN*.

[1]*Department of Mathematics, M.D. University, Rohtak-124001, Haryana, India (E-mail: archna.rs.maths@mdurohtak.ac.in)*

ORCID: 0009-0006-9617-6730

[2]*Department of Mathematics, M.D. University, Rohtak-124001, Haryana, India (E-mail: drsumeetgill@mdurohtak.ac.in)*

ORCID: 0000-0001-6471-1192

[3]*Department of Mathematics, M.D. University, Rohtak-124001, Haryana, India (E-mail: vikasjangra96.rs.maths@mdurohtak.ac.in)*

ORCID: 0009-0000-5324-8358

[4]*Department of Mathematics, M.D. University, Rohtak-124001, Haryana, India (E-mail: binny.rs.maths@mdurohtak.ac.in)*

ORCID: 0000-0002-8023-2076

1 Introduction

Emotions play a crucial role in human communication, being conveyed through voice, gestures, facial expressions, and other channels that carry significant affective information [1][2][3]. However, Human-Computer Interaction (HCI) still lacks certain emotional components needed for truly human-centered communication. Affective computing addresses this gap by enabling systems to recognize emotions and generate appropriate responses, enhancing interaction effectiveness. This field has diverse applications, including healthcare [4], education [5], entertainment [6], and autonomous vehicles [7].

Early emotion recognition studies primarily focused on unimodal approaches, such as speech emotion recognition [8][9], text-based emotion recognition [10][11], and facial expression analysis [12][13]. However, single-modal systems often suffer from data incompleteness and noise, limiting their accuracy. To overcome these challenges, researchers have

increasingly adopted Multimodal Emotion Recognition (MER), which integrates multiple modalities to leverage complementary information and improve recognition accuracy. MER utilizes mutual information to measure relationships between different modalities, enabling the extraction of more discriminative features. As a result, MER has attracted significant research interest, as it enhances emotional judgments by combining complementary modalities [14][15]. With the rapid advancements in deep learning, MER based on deep learning has emerged as a crucial research area, focusing on designing effective network architectures [16].

Emotion recognition using both facial expressions and speech is gaining prominence due to the complementary nature of these modalities, enhancing emotion analysis in video-based applications [17]. There have been multiple machine learning (ML) and deep learning (DL) methods applied to audio-video emotion recognition (ER) with much progress made on hybrid models fusing different learning schemes.

Traditional Approaches

Early approaches were heavily based on statistical models like Gaussian Mixture Models (GMM) [18] and Support Vector Machines (SVM) [19]. Feature selection methods such as PCA were also used to enhance the classification performance [20]. While useful to a certain extent, these methods had difficulty handling complexity and variability in real emotional expressions.

Transition to Deep Learning

With the development of deep learning technology, neural network-based frameworks such as Convolutional Neural Networks (CNNs) [21], Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks improved recognition accuracy substantially [22]. Hybrid models combining machine learning with deep learning, such as CNN-SVM models, showed better results [23,24].

Some of these studies investigated on CNN for facial expression analysis and SVM for speech emotion recognition. Feature and decision-level fusion methods were applied in [25], and 77.4% accuracy (SAVEE) and 69.3% (RML) were reached. Likewise, CNN-LSTM model for audio and 3D-CNN for video achieved 74.3% accuracy on eNTERFACE'05 [22]. SAVEE indicating the accuracy of 63.1% with pre-trained ResNet together with CNNs [26]. LSTM- RNNs models for speech and CNN based classifiers for images combined through weighted sum functions had some promising results [27].

Advanced Hybrid Models

For higher performance, more complex hybrid models combined various neural network architectures. TCN were utilized with fusion of CNN and SVM [28]. The authors in [29] combined features using Deep Belief Networks (DBNs), and were able to attain 80.36% and 85.95% on RML and eNTERFACE05, respectively. One model, which jointly used CNN to analyze Mel-spectrogram of speech and emotional recognition of facial video, fused by using both Extreme Learning Machines (ELMs) and SVM, achieved an accuracy of 84.6% on the eNTERFACE'05 [30].

The hybrid models SVM + CNN and RF + CNN had achieved 100%, 99.72%, and 98.73% recognition rates for the SAVEE, RML, and eNTERFACE05 datasets indicating the superior of the hybrid classifier systems [24]. A recent work that integrated 2D-CNN for audio, 3D-CNN for video, DBN for feature fusion and SVM for classification achieved accuracies of 82% and 84.5% on RML and eNTERFACE05 [31]. A hybrid of rule-based and machine learning model reached 90.83% on RML and 86.67% on eNTERFACE05 [20].

Recent Developments

Another line of work developed cross-attention modules for CNN-TCN models [28] and attention mechanisms for temporal emotion recognition, leading to better multimodal learning effectiveness [32]. The work in [24] combined traditional and deep learning methods using prosodic features, MFCCs and FBEs. The method employed a multi-class SVM for emotion classification for speech and a CNN classifier for facial expression recognition and subsequently combined the produced decisions through an extra classification layer. When compared to other models, Random Forest classifier performed the best by obtaining the state- of-the-art results on different datasets.

In [31], to improve the recognition performances, a hybrid approach that concatenated 2D-CNN with 3D-CNN for speech and visual data and used Deep Bayesian Network (DBN) for fusion was proposed. Another model used Temporal Convolution Networks (TCN) for detecting facial features and 2D-CNN for speech together using cross-attention mechanism for final classification [28].

Furthermore, attention-based multimodal architectures have also been studied to capture the temporal relationship between facial and speech information. In [32], a Multi-modal Attention network was constructed to mine the time variance of emotion segments and facilitate classification accuracy.

In this research, we explore the fusion of facial and speech features for emotion recognition using advanced deep learning architectures. We make use of transfer learning (ResNet-50) for more efficient and accurate image feature extraction. Moreover, the Enhanced Zebra algorithm is used to select feature and decrease dimensionality and to enhance model behavior. For classification, we utilize a Bidirectional LSTM (BiLSTM) combined with a self- attention mechanism which captures temporal relationships and emphasizes most important input sequences for achieving higher accuracy and interpretability.

Problem Statement

Despite recent advances, emotion recognition remains a challenging task due to the presence of noisy and unimodal data, the struggle of fusing multimodal cues and the limited generalization ability to various environments. Unimodal strategies such as only focusing on the facial expression or the speech, however, cannot always capture the full complexity of human emotions. To mitigate such limitations, in this work we propose a multi-modality approach, which combines the ResNet50 for facial features and the BiLSTM empowered with self- attention for the prosodic aspects of speech.

Our contributions are as follows:

- Multimodal Fusion Strategy: We present a deep learning based multimodal model that effectively fuses facial and speech information for emotion recognition.
- Optimized Feature Selection: To further improve model efficiency, the improved zebra algorithm is adopted to particularly optimize the feature selection, which reduce data redundancy.
- Advanced Classification Model: We implement a BiLSTM-based classification model with a self-attention mechanism to capture temporal dependencies and improve interpretability.
- Comprehensive Evaluation: We have compared our studies with baselines models, demonstrating the effectiveness of our proposed model.

The paper is structured as follows. The introduction and review of literature is presented in Section 1. The background of the methods used in this research is explained in Section 2. The methodology, including details on the dataset, feature extraction, and model development, is outlined in Section 3. Section 4 presents the results and performance analysis of the proposed approach. Finally, the conclusion and future scope are discussed in Section 5.

2 BACKGROUNDS

2.1 Bidirectional LSTM

The spatial relationship between different facial regions is crucial in recognizing facial expressions, as they are composed of various movements of brows and lips. Nevertheless, convolutional filters struggle to capture this relationship as they only apply to specific image regions. Thus, it is crucial to investigate the spatial dependencies within facial expression images to enhance facial expression recognition performance. As a result, we have chosen to utilize the LSTM method. This approach treats each row or column in the feature maps as a directed sequence. The created sequence is then arranged independently, either top to bottom or left to right. Every element in the sequence corresponds to conv5 3 receptive field, an important area in the original image sample.

The left-to-right (o→ t) and top-to-bottom (o↓ t) LSTM responses of the two spatial sequences are concatenated and totaled together. The following equations provide a summary of the process

$$o_t^{\rightarrow} = \sigma\left(W_o^{\rightarrow} x_t^{\rightarrow} + U_o^{\rightarrow} h_{t-1}^{\rightarrow} + b_o^{\downarrow}\right) \quad (1)$$

$$o_t^{\downarrow} = \sigma\left(W_0^{\downarrow} x_t^{\downarrow} + U_0^{\downarrow} h_{t-1}^{\downarrow} + b_o^{\rightarrow}\right) \quad (2)$$

$$o_t = o_t^{\rightarrow} + o_t^{\downarrow} \quad (3)$$

2.1 Self-attention mechanism

The low-level input sequence is transformed into higher-level and more abstract representations via the self-attention mechanism using multi-head scaled dot-product attention. A feature sequence X = {x1, x2,…, xT} ∈ $R^{T \times d}$, where $x_i \in R^d$ is the frame-level feature at step t and T is the maximum time step, is fed into the self-attention layer.

In order to perform multi-head scaled dot-product attention on sequence X, relevant queries Q, keys K, and values V must be created. In order to accomplish this goal, we use multiple linear projection layers, which are computed as follows, to apply X to h:

$$Q_i = XW_i^Q \quad (4)$$

$$K_i = XW_i^K \quad (5)$$

$$V_i = XW_i^V \quad (6)$$

where Q_i, K_i and $V_i \in R^{d \times (d/h)}$, i ranges from 1 to h and h is the number of heads.

We use the following equation to execute the scaled dot-product attention for each head's query Q_i, key K_i, and value V_i:

$$Head_i = Soft\,max\left(Q_i K_i^T / \sqrt{d_k}\right) V_i$$

Where the scale factor is denoted by dk = d/h and head ∈ RT ×(d/h). The outcomes of each head are then combined and projected linearly to produce

$$R = Concat\,(Head_1, \ldots\ldots, Head_h) W_O \quad (8)$$

The projection matrix, denoted as WO ∈ $R^{(d \times d)}$, is being referred to. Based on reference [24], To get the final encoded sequence S, we add a residual connection and layer normalization.

$$S = layerNorm(X + R) \quad (9)$$

2.2 Autoencoder

The goal of a regular autoencoder (AE) is to minimize error when reconstructing the input data into the output. An encoder component compresses the network's input into lower-dimensional variables, also known as codes or latent variables. A decoder component then reconstructs the latent variables into their representation (such as an image, text, or speech) at the network's output. This is how a regular AE network is put together. In Figure 1, this network is displayed.

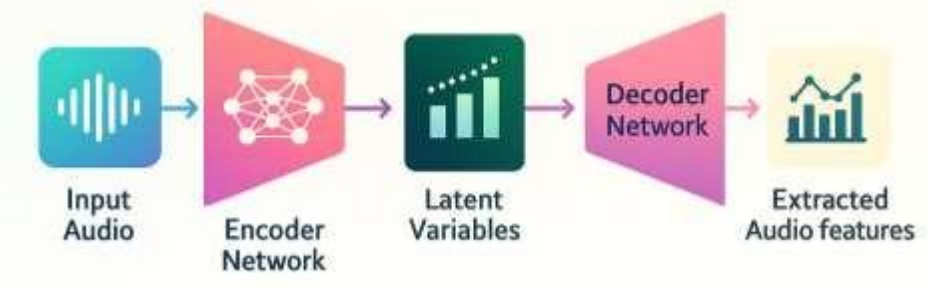

Figure1: Architecture for Autoencoder

The network cannot learn any meaningful representations if the regular AE is built to be flawlessly able to duplicate its input. Alternatively, the standard AE network has a restriction. The constraint is situated at the latent variable's dimension, which is smaller than the input dimension. As a result, the encoder picks up some crucial aspects from the input, and the decoder attempts to figure out how to use those features to reconstruct the output. As seen in Equation 10, where L is the MSE loss function, Xˆ is the reconstructed picture, X is the input image, and N is the total quantity of training data, they are often trained collectively using mean- squared error (MSE) loss between the reconstruction and the input. If there is a difference between the input and the output, the network is penalized by the loss. Because of the pixel-wise MSE loss, the reconstructed image is therefore blurrier than the original. Undercomplete autoencoder is an illustration of this type of regular AE network [33].

$$L = \frac{1}{N}\sum_{i=0}^{N}\left(\overline{X} - X\right)^2 \quad (10)$$

2.1 Transfer learning with ResNet-50

One of ResNet's convolutional neural network variations [9] is ResNet-50, which has 50 layers. It has one MaxPool, one Average Pool, and forty-eight Convolution layers. ResNet-50's design is described in completely in Figure 2. ResNet is based on the deep residual learning framework [9]. It resolves the vanishing gradient issue even in cases where neural networks are quite deep. Resnet-50 has more than 23 million trainable parameters despite only 50 layers, significantly fewer than existing architectures.

Though there is still room for debate, the simplest explanation of its performance is to review residual blocks and how they function. Let's look at a neural network block with x input. Where the true distribution H(x) is what one has to know. The difference (or residual) between these can be represented as follows:

$$R(x) = Output - Input = H(x) - x \quad (11)$$

We end up rearranging it.

$$H(x) = R(x) + x \quad (12)$$

.

The residual block attempts to discover H (x), the true output.

The layers are learning the residual, or R(x), as can be seen from taking a deeper look at the figure since x results in an identity connection. A typical network's layers learn the true output (H(x)), but the layers in a residual network learn the residual (R(x)). It is also shown that learning the residual of the input and output together is less complicated than learning the input by itself. The identity residual model allows for the reuse of these activation functions from earlier levels because it bypasses them and doesn't complicate the design.

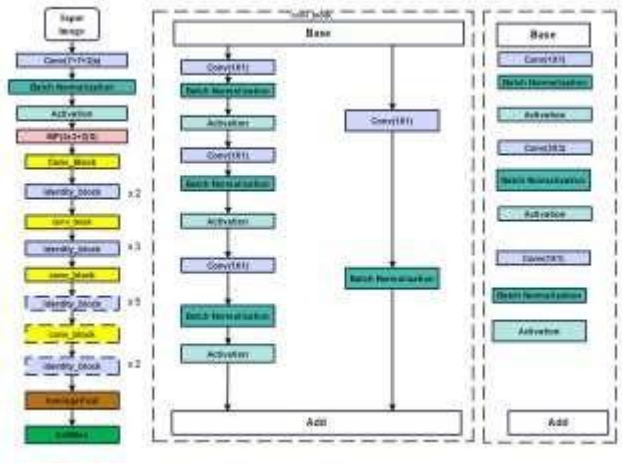

Figure 2: ResNet 50 Architecture

3. METHODOLOGY

This section describes the full methods for emotion recognition based on audio and image features derived from video dataset.

3.1 Dataset

This study used Multimodal Emotion Lines Dataset (MELD), which includes not only textual dialogues, but also their corresponding visual and audio counterparts. MELD contains around 1400 conversations and 13,000 utterances from the television show Friends. The discussions featured a number of speakers. All seven emotions—Anger, Disgust, Sadness, Joy, Neutral, Surprise, and Fear—have been assigned labels to each dialogue exchange, but in this article, we focused on three primary emotions. MELD was selected due to its rich multimodal

characteristics, which enable a comprehensive evaluation of emotion recognition based on audio and facial expressions [34].

3.2 Feature Extraction

In this step, acoustic and visual features were extracted independently and then combined to form a unique feature representation.

3.2.1 Extracting Audio Feature: We first extracted audio tracks from videos files through the common FFmpeg library for processing audio or video media files. The audio was saved as wav format due to its high fidelity and compatibility with deep learning frameworks.

The audio went through pre-processing to enhance the quality of the data, such as noise reduction (spectral subtraction) and normalization to keep the amplitude consistent. After pre- processing, a deep feature extraction process was applied using an Autoencoder. The pre- processed audio signals were encoded into a lower-dimensional representation before being reconstructed, with deep features being extracted from the bottleneck layer. These features encoded important sound characteristics that contributed to emotion recognition.

3.2.2 Visual Feature Extraction: Complementing the audio processing, important visual features were extracted from the videos as summary images. This was done by key frame extraction operations like frame averaging that produces composite images summarizing the video visual content. The summary images preserved the essential facial information for emotion recognition.

ResNet-50 was employed for feature extraction because it is a better learner of high- level spatial representations. We employed transfer learning with pre-trained ResNet-50 to obtain deep facial features to make it easy to well represent emotional expressions.

3.3 Feature Concatenation

Following extraction, the audio and image features were concatenated to create a joint feature vector. This cross-transformation exploited the complementary information of the two modalities, which makes the model for the emotion recognition task robust.

3.4 Feature Selection Using Improved Zebra Algorithm (IZA)

To remove redundant features and enhance classification accuracy, feature selection was conducted with IZA. This is a feature selection approach which selects the most informative features and discards irrelevant ones, thus, simplifying the computational complexity. The IZA begins with the creation of an initial population of possible feature subsets. A fitness function is employed to evaluate each subset in terms of classification accuracy. The approach repeatedly hones the population by exploiting and exploring simultaneously to balance the novelty of discovered feature subsets with the improving nature of currently identified ones. This iterative process goes on until an optimal set of features have been achieved, leading to choose subset of features(dimensions) that are considerably small but informative.

3.5 Train-Test Split

Following feature selection, the dataset was divided into testing and training sets. Training data contains about 80% of the data which is utilized to train the model, the remaining 20% is used as testing set, this test data is used to evaluates how well the model predicts on new data. It ensures its real-world effectiveness and generalization capability. The flow chart of the modal is given by Figure 3.

Figure 3: Flow Chart

3.6 Model Building and classification

The mixed Emotion Recognition Model was developed with Bi-LSTM (Bi-Directional Long Short-Term Memory) network together with Self-Attention Mechanism. This architecture was chosen for its effectiveness in capturing sequential dependencies in speech while selectively attending to the most important features in the input sequence.

The procedure of model training was pipeline mannered. In the first stage, we input the multimodal features from the IZA dataset into the Bi-LSTM layer, which learned the temporal dependencies in the speech and facial signals. The self-attention mechanism was then applied to enhance focus on the most informative segments of the input. Finally, the features were forwarded to the fully connected layer for classifying the emotions among the selected categories.

This hybrid model has achieved a robust modeling temporal cue in speech and vision cues in facial expression and achieved better performance labeling emotions.

3.7 Model Evaluation

For evaluating the proposed emotion recognition model, the accuracy, precision, recall, F1-score and confusion Matrices were used to measure the generalization and performance. These measures guarantee a complete assessment of the classification performance over various emotion types.

- **Accuracy:** The percentage of the correctly predicted samples from the total samples is called the accuracy and computed as:

$$\text{Accuracy} = \frac{T_p+T_n}{T_p+T_n+F_p+F_n}$$

- **Precision :** Precision (also known as Positive Predictive Value) measures the ratio of correctly predicted positive samples to the total number of predicted positives:

$$\text{Precision} = \frac{T_p}{T_p+F_p}$$

- **Recall (Sensitivity**): The ability of the model to identify the actual positive samples is measured by recall (also known as sensitivity or true positive rate):

$$\text{Recall} = \frac{T_p}{T_p+F_n}$$

- **F1-Score:** F1-score is the harmonic mean of precision and recall, which is useful for balancing Precision and Recall, especially in the imbalanced dataset situation:

$$\text{F1score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where T_p = True positives, T_n = True negatives, F_p = False positives and F_n= False negatives.

- **Confusion Matrix Analysis:** A confusion matrix was created to include the ratio that shows the number of correct and wrong classifications between any two types of emotions. It enabled the analysis of the patterns of misclassifications and the directions in which the model needed to be improved.

3.8 Experimental setup

The experimental analysis for the proposed technique involved conducting the experiments on a local PC operating on Windows 11 OS with 8GG RAM and Core i5 processor. The code was created and executed using a jupyter notebook. Pre-processing of the models and their execution were carried out using various machine learning packages, such as numpy, pandas, sklearn, and more.

4. RESULTS AND DISCUSSION

4.1 Results

Results of this study show that various deep learning models for Multimodal Emotion Recognition. The models were compared using accuracy, precision, recall, F1-score, and confusion matrices to check which emotion was best predicted in three emotions category: Sad, Happy and Surprise. Comparing the results shows the superiority of the method of adding self- attention mechanism to Bi-LSTM as for classification.

Interpretation of Findings

The Bi-Directional LSTM with Self-Attention Mechanism obtained best accuracy of 89.83% which 4.68% more than the standard Bi-Directional LSTM which had an accuracy of 85.15% and 2.96% more than CNN model which had accuracy of 86.87% (as shown in Table 1). The self-attention mechanism improved the efficacy of the model in concentrating on the relevant features, resulting in better classification performance. We note also that this gain is also observed on precision, recall and F1-score where the Bi-Directional LSTM with Self-Attention Mechanism had a superior recall, which lead to lower misclassification rates.

Table 1: Results

Models	Accuracy	Precision	Recall	F1-Score
Bi-directional-LSTM with self-attention mechanism	0.89834	0.89833	0.898	0.898
Bi-directional-LSTM	0.85151	0.85165	0.85151	0.85152
CNN	0.8687	0.86877	0.8687	0.86867

The confusion matrices (as shown in Figures 4 -6) give additional insight in model success:

- The Bi-Directional LSTM with Attention Mechanism (as we see in Figure 4) is able to accurately classify the 963 Sad, 1000 Happy, and 1067 Surprise in almost every cases. Happy and Sad were the most often misclassified AUs, which is not surprising because it is known that emotion prediction with speech is particularly difficult for Happy and Sad [7].
- Bi-Directional LSTM (as we see in figure 5) was able to accurately recognise 983 Sad, 981 Happy, and 908 Surprise instances, but got more confused with Happy and Sad emotions. This indicates that though the synchronous word based Bi-LSTM captures sequential dependency well, the model lacks the attentions, which leads to a bit of misclassification.
- CNN (as we see in Figure 6) had a good performance, accurately classifying for CNN, which relied on spatial feature extraction without sequence modelling, these mistakes are understandable.

The confusion matrices also reveal that the self-attention mechanism has been effective in reducing the coefficient of the misclassification, especially in subtle differences of emotions. The CNN model had more balanced errors across categories but was less effective than Bi- LSTM models in distinguishing temporal patterns in speech-based data.

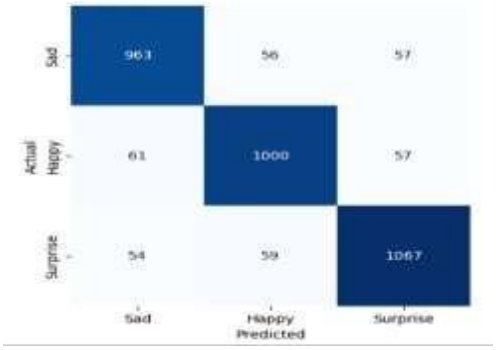

Figure 4: Bi-Directional-LSTM with self-attention-mechanism Confusion matrix

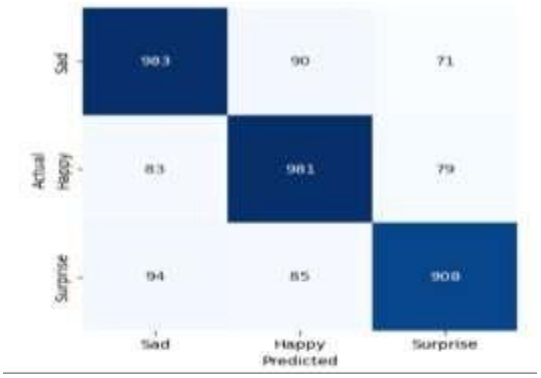

Figure5: Bi-Directional-LSTM Confusion matrix

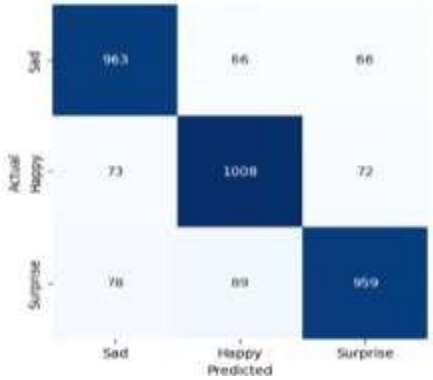

Figure 6: CNN Confusion-Matrix

To investigate the model performance in depth, a bar graph (as shown in Figure 7) was drawn to compare the precision, recall, F1-score, and accuracy for all three models. The main points illustrated by the bar graph:

- The Bi-Directional LSTM with Self-Attention Mechanism has significant improvements in recall and F1-score, which suggests to have more generalization and less misclassification cases.
- The CNN model performs well in precision but struggles slightly with recall, likely due to its reliance on spatial features without sequential modeling.
- Bi-Directional LSTM model has comparable performance on all the metrics but they lack the improved feature selection mechanism of the attention-based version.

This visual illustration provides additional evidence of why the use of self-attention mechanisms in deep learning models is better suited for multi-modal emotion recognition.

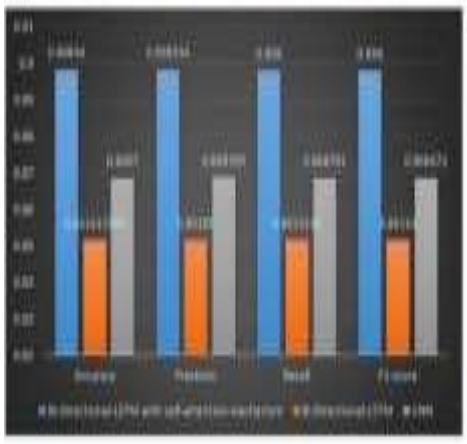

Figure 7: Comparison of Proposed Models

3.1 Discussion

Our findings demonstrate that the proposed multimodal approach-using ResNet50 for visual feature extraction and BiLSTM based attention for speech can achieve a large improvement in emotion recognition accuracy. The model yielded an accuracy of 89.83% that outperformed both similar unimodal systems as well as previous multimodal literature. By employing self-attention, our model was allowed to attend to important temporal pattern in the speech data and this justified its effectiveness in sequential data processing.

When compared to prior studies, our results show a clear improvement. For instance, J. He (2025) reported an accuracy of 73.56%, while Alsaadawi and Daş (2024) achieved 81%.

Table 2 highlights this comparison.

Table 2: Comparison with provious work

J. He (2025) [35]	73.56%
Alsaadawı, H. F. T., & Daş, R. (2024) [36]	81%
Our Proposed Model	**89.83%**

The superior performance of our model may result from the fusion deep visual features and improved temporal dynamics of BiLSTM with self-attention. Whereas previous works were based on single modality features, our method takes the full advantage of multimodal information, which demonstrates the good ability of multimodal feature fusion. Nevertheless, the model performance might still depend on background noise or speaker variability indicating further room for improvement.

This work is part of and makes a new contribution to emotion aware AI and human-computer interaction. Verification that self-attention is useful for BiLSTM-based structures for speech emotion recognition allows progress to be made toward systems capable of more precise, real- time multimodal interaction.

Conclusion

In this paper, we proposed a multimodal for emotion recognition, which utilized MFCC based audio features and visual features generated by CNN and finally input into the Bi-Directional LSTM with a self-attention mechanism. The focus layer increases the importance of the features and decreases the noise to reduce noise and improve performance.

We found that the proposed model attained better results in terms of accuracy, precision, recall, and F1-score in comparison to both baseline CNN and Bi-LSTM models after its performance was evaluated based on various test metrics and confusion matrices. These findings demonstrate the robustness of attention based end-to-end deep models for emotion recognition.

Apart from the performance improvements, this work leads to emotionally sensitive human-computer interaction. Potential future improvements include incorporating other modalities such as physiological signals and text, and facilitating cross-cultural generalization to enable real-world deployment.

References

[1] P. V. Rouast, M. Adam, and R. Chiong, Deep learning for human affect recognition: Insights and new developments, IEEE Transactions on Affective Computing, 12 (2019), 524-543, DOI: 10.1109/TAFFC.2018.2890471.

[2] T. Baltruˇ saitis, C. Ahuja, and L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence, 41 (2019), 423–443, DOI: 10.1109/TPAMI.2018.2798607.

[3] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, A survey of affect recognition methods: Audio, visual, and spontaneous expressions, IEEE transactions on pattern analysis and machine intelligence, 31 (2009), 39– 58, DOI: 10.1145/1322192.1322216.

[4] C.-H. Wu, Y.-M. Huang, and J.-P. Hwang, Review of affective com puting in education/learning: Trends and challenges, British Journal of Educational Technology, 47 (2016), 1304–1323, DOI: 10.1111/bjet.12324.

[5] B. G. Lee, T. W. Chong and B. Kim, Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers, IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, (2017), DOI: 10.1016/j.eswa.2017.01.040.

[6] S. Cosentino, E. I. Randria, J.-Y. Lin, T. Pellegrini, S. Sessa, and A. Takanishi, Group emotion recognition strategies for entertainment robots, International Conference on Intelligent Robots and Systems (IROS), (2018), 813–818, DOI: 10.1016/j.eswa.2017.01.040.

[7] L. Y. Mano, B. S. Faic¸al, L. H. Nakamura, P. H. Gomes, G. L. Libralon, R. I. Meneguete, P. Geraldo Filho,

[8] G. T. Giancristofaro, G. Pessin, B. Krishnamachari et al., Exploiting iot technologies for enhancing health smart homes through patient identification and emotion recognition, Computer Communications, 89 (2016), 178–190, DOI: 10.1016/j.comcom.2016.03.010.

[9] Y. Zong, H. Lian, H. Chang, C. Lu, and C. Tang, Adapting Multiple Distributions for Bridging Emotions from Different Speech Corpora, Entropy, 24 (2022), 1–14, DOI: 10.3390/e24091250.

[10] H. Fu, Z. Zhuang, Y. Wang, C. Huang, and W. Duan, Cross-Corpus Speech Emotion Recognition Based on Multi-Task Learning and Subdomain Adaptation, Entropy, 25 (2023), 1–10, DOI: 10.3390/e25010124.

[11] S. Shaheen, W. El-Hajj, H. Hajj, and S. Elbassuoni, Emotion Recognition from Text Based on Automatically Generated Rules, International Conference on Data Mining Workshop, (2014), 383–392, DOI: 10.1109/ICDMW.2014.80.

[12] C. H. Wu, Z. J. Chuang, and Y. C. Lin, Emotion recognition from text using semantic labels and separable mixture models, ACM Transactions on Asian Language Information Processing., 5 (2006), 165–182, DOI: 10.1145/1165255.1165259.

[13] S. Li and W. Deng, Deep Facial Expression Recognition: A Survey, IEEE Transactions on Affective Computing, 13 (2022), 1195–1215, DOI: 10.1109/TAFFC.2020.2981446.

[14] H. Yang, L. Xie, H. Pan, C. Li, Z. Wang, and J. Zhong, Multimodal Attention Dynamic Fusion Network for Facial Micro-Expression Recognition, Entropy, 25 (2023), DOI: 10.3390/e25091246.

[15] J. Zeng, T. Liu, and J. Zhou, Tag-assisted Multimodal Sentiment Analysis under Uncertain Missing Modalities, Association for Computing Machinery, 1 (2022), DOI: 10.1145/3477495.3532064.

[16] Y. Li, Y. Wang, and Z. Cui, Decoupled Multimodal Distilling for Emotion Recognition, Proc. IEEE Comput. Soc. Conf. Computer Vision Pattern Recognition, (2023), 6631–6640, DOI: 10.1109/CVPR52729.2023.00641.

[17] S. E. Kahou et al., Combining modality specific deep neural networks for emotion recognition in video, ICMI 2013 International Conference Multimodal Interaction, (2013), 543–550, DOI: 10.1145/2522848.2531745.021.03.058.

[18] S. Lee, D. K. Han, H. Ko, Multimodal emotion recognition fusion analysis adapting bert with heterogeneous feature unification, IEEE Access, 9 (2021), 94557–94572, DOI: 10.1109/ACCESS.2021.3092735.

[19] S. Dobrisek, R. Gajsek, F. Mihelic, N. Pavesic,

V. Struc, Towards efficient multi-modal emotion recognition, Int J Adv Robotic Syst, 10 (2013), 1–10, DOI: 10.5772/54002.

[20] Z. Sara, A. Zahid, C. E. Cigdem, Multimodal emotion recognition based on peak frame selection from video, Signal Image Video Process., 10 (2016), 827–843, DOI: 10.1007/s11760-015-0822-0.

[21] S. K. Phooi, A. Li-Minn, O. C. Shing, A combined rule-based & machine learning audio-visual emotion recognition approach, IEEE Trans Affect Comput, 9 (2018), 3–13, DOI: 10.1109/TAFFC.2016.2588488.

[22] A. Shrestha, A. Mahmood, Review of deep learning algorithms and architectures, IEEE Access, 7 (2019), 53040–53065, DOI: 10.1109/ACCESS.2019.2912200.

[23] Z. Farhoudi, S. Setayeshib, Fusion of deep learning features with mixture of brain emotional learning for audio-visual emotion recognition, Speech Commun, 127 (2020), 92–123, DOI: 10.1016/j.specom.2020.12.001.

[24] E. Avots, T. Sapinski, M. Bachmann, D. Kaminska, Audiovisual emotion recognition in wild, Mach Vis Appl, 30 (2019), 975–985, DOI: 10.1007/s00138-018-0960-9.

[25] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, G. Anbarjafari, Audio-visual emotion recognition in video clips, IEEE Trans Affect Comput, 10 (2019), 60–75, DOI: 10.1109/TAFFC.2017.2713783.

[26] E. Avots, T. Sapin, M. Bachmann, D. Kamin, Audiovisual emotion recognition in wild, Mach Vis Appl, 30 (2019), 975–985, DOI: 10.1007/s00138-018-0960-9.

[27] T. Hussain, W. Wang, N. Bouaynaya, H. Fathallah-Shaykh, L. Mihaylova, Deep learning for audio visual emotion recognition, 2022 25th International Conference on Information Fusion (FUSION), (2022), 1–8, DOI: 10.23919/FUSION49751.2022.9841342.

[28] W. Ding, M. Xu, D. Huang, W. Lin, M. Dong, X. Yu1, H. Li, Audio and face video emotion recognition in the wild using deep neural networks and small datasets, 2016 18th ACM International Conference on Multimodal Interaction (ICMI), (2016), 505–513, DOI: 10.1145/2993148.2997637.

[29] G. P. Rajasekar, W. C. Melo, N. Ullah, H. Aslam, O. Zeeshan, T. Denorme, M. Pedersoli, A. Koerich, S. Bacon, P. Cardinal, E. Granger, A Joint Cross-Attention Model for Audio-Visual Fusion in Dimensional Emotion Recognition, arXiv, (2022), DOI: 10.48550/ARXIV.2203.14779.

[30] S. Zhang, S. Zhang, T. Huang, Learning affective features with a hybrid deep model for audio-visual emotion recognition, IEEE Trans Circ Syst Video Technol, 28 (2018), 3030–3043, DOI: 10.1109/TCSVT.2017.2719043.

[31] M. S. Hossain, G. Muhammad, Emotion recognition using deep learning approach from audio-visual emotional big data, Inf Fusion, 49 (2019), 69–78, DOI: 10.1016/j.inffus.2018.09.008.

[32] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, A. Košir, Audio-visual emotion fusion (avef): A deep efficient weighted approach, Inf Fusion, 46 (2019), 184–192, DOI: 10.1016/j.inffus.2018.06.003.

[33] E. Ghaleb, J. Niehues, S. Asteriadis, Joint modelling of audio-visual cues using attention mechanisms for emotion recognition, Multimed Tools Appl, 82 (2023), 11239–11264, DOI: 10.1007/s11042-022-13557-w.

[34] S. Zhao et al., A two-stage 3D CNN based learning method for spontaneous micro-expression recognition, Neurocomputing, 448 (2021), 276–289, DOI: 10.1016/j.neucom.2021.03.058.

[35] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, Meld: A multimodal multi-party dataset for emotion recognition in conversations, arXiv preprint arXiv:1810.02508, (2018), DOI: 10.48550/arXiv.1810.02508.

[36] J. He, A Multimodal Approach for Emotion

Recognition in Conversations Using the MELD Dataset, 2025 Asia-Europe Conference on Cybersecurity, Internet of Things and Soft Computing (CITSC), (2025), 54-58, DOI: 10.1109/CITSC64390.2025.00016.

[37] [H. F. T. Alsaadawı and R. Daş, Multimodal Emotion Recognition Using Bi-LG-GCN for MELD Dataset, Balkan Journal of Electrical and Computer Engineering, 12 (2024), 36-46, DOI: 10.17694/bajece.1372107