Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.58, p.741-754. March/April, 2026.
DOI: https://doi.org/10.5935/jetia.v12i58.2986

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE

OPEN ACCESS

MULTI-SCALE ATTENTION-GUIDED CNN-BILSTM FRAMEWORK FOR EMOTION RECOGNITION IN MULTIMODAL VIDEO DATA

J. Biju[*1], Lavanya K[2],J. Raja[3],M. Kiruthiga Devi[4], Payala Krishnanjaneyulu[5] and N. Kanya[6]

[1]Assistant Professor, Division of Data Science and Cyber Security, Karunya Institute of Technology and Sciences, Coimbatore, India.
[2]Assistant professor of Artificial Intelligence & Data Science, S, Akshaya college of engineering and technology, Kinathukadavu, CBE, India.
[3]Associate. professor, Computer Science and Engineering, School of Computing, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India.
[4]Assistant Professor (Sr.G), Department of Computer Science and Engineering, SRM Institute of Science and Technology, Vadapalani, India.
[5] Assistant professor, Koneru Lakshmaiah Education Foundation Bowrampet Hyderabad, India.
[6]Professor of Information Technology, Dr.M.G.R. Educational and Research Institute, Chennai, India.

[1]https://orcid.org/0009-0009-1628-7871, [2]https://orcid.org/0000-0002-1116-5738, [3]https:/orcid.org/0000-0003-2183-8585
[4]https://orcid.org/0000-0003-2949-6102,[5]https://orcid.org/0000-0003-4767-0178, [6]https://orcid.org/0000-0001-8549-6444

Email:*jbijuinfo@gmail.com, lavanya030891@gmail.com, drrajaj@veltech.edu.in, kiruthim6@srmist.edu.in, krishna54888@gmail.com, kanya.v@drmgrdu.ac.in

ARTICLE INFO

Article History
Received: November 21, 2025
Reviewed: December 30, 2025
Accepted: March 10, 2026
Published: April 30, 2026

Keywords:
Bi-LSTM
CNN
multimodal emotion recognition,
AI technologies

ABSTRACT

A mental state, emotion is connected to human behaviour, thoughts, and the degree of positive or negative experiences. Human emotion does not yet have a precise definition. By allowing AI systems to precisely comprehend and sympathetically react to human emotions, this discovery has the potential to completely transform human-machine interaction and open the door for increasingly sophisticated and emotionally intelligent computers. The main research problem is creating models that accurately read emotions from multimodal data; this calls for big, diverse datasets for video data to capture complex emotional cues and fine-tuned CNNs for audio data to identify minor speech changes. This study introduces a novel multimodal emotion detection method that seamlessly combines voice and video modalities to correctly infer emotional states. The attention-based CNN-Bi-LSTM model handles the video component and provides deep semantic understanding through its bidirectional layers. An attention-based fusion process is used to blend the results of both modalities, balancing their respective contributions. Here, the suggested methodology is thoroughly tested using two different datasets: the YouTube and Carnegie Mellon University SAVEE datasets.The results show higher efficacy compared to current frameworks. This comprehensive technology enables accurate emotion recognition and contributes to a number of noteworthy developments in the industry.

I. INTRODUCTION

Numerous fields, including computer science, psychology, and neuroscience, have made considerable strides in their understanding of human emotion over the last 20 years. Human emotion is defined subjectively in the literature [1]. Human emotion is explained by both cognitive and physiological neuroscientific ideas (Figure 1). According to the physiological viewpoint, emotion originates from physiological reactions. The difference in physiological reaction could be due to the external incident. Emotions are linked to thoughts and other mental processes. People get scared when they see a snake, for example, since they think they are in danger [2]. Therefore, human emotion is defined as a feeling of pleasure or discontent that arises from a change in physiological and cognitive processes. A human emotion monitoring system is a crucial part of an effective human-computer interface. People interact with computers in many ways using desktops, laptops, mobile devices, kiosks, and other platforms. Computer agents with emotional intelligence will monitor user behavior.

It will sound a warning before the user can compose a hostile email, an angry comment, or any other negative response. It will also assist the user in selecting music based on their emotional state. An emotionally intelligent computer agent will assist the user in learning systems like pilot training, driving car learning, and others [3]. The emotion-aware agent monitors the driver's emotional state while they are operating a vehicle. It notifies other people when the driver is upset.

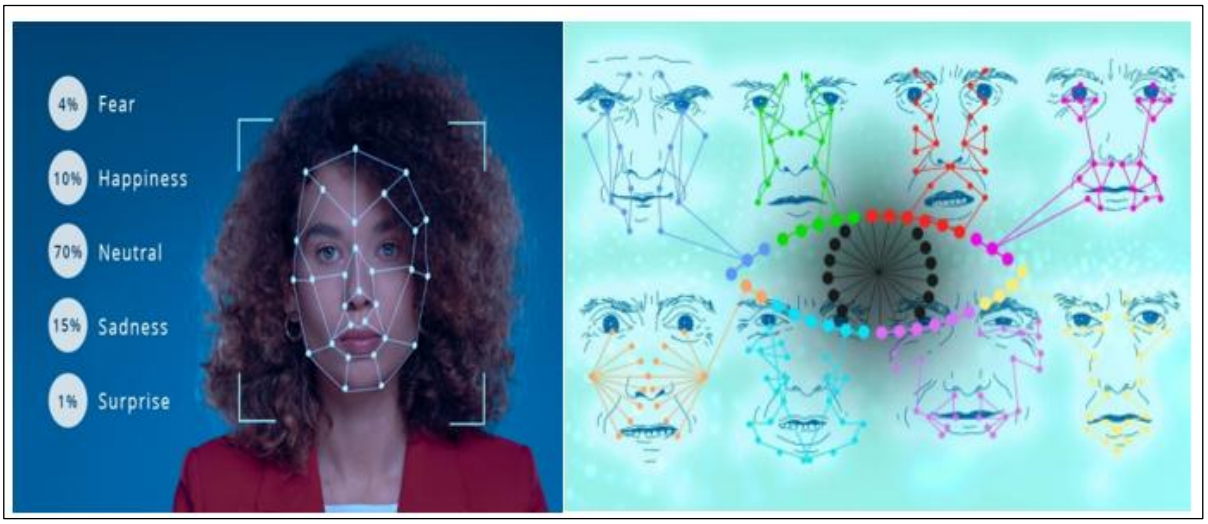

Figure 1: Emotion recognition by Software.
Source: Authors, (2026).

The emotion-aware agent thus attempts to safeguard the user. In affective computing, multimodal emotion recognition is a relatively new field of study. Physical, optical, and aural clues are all combined to create multimodal signals. The audio and video signals demonstrate how people express themselves through their voices and facial expressions when a particular emotion arises [4]. A particular emotion triggers physiological signs that indicate the body's symptoms, including heart rate, body temperature, breathing rate, muscle current, and brain electrical activity. Researchers are urged to develop a multimodal emotion identification system in order to differentiate between human emotions since multimodal emotion recognition is a key pattern recognition problem [5].

To more precisely ascertain people's moods, it combines data from their speech, body language, and facial expressions. In this study, we use human speech and facial expressions to recognize emotions in several ways. Here, the fusion process balances the audio and video contributions using an attention-based technique. It assigns ratings of attention, giving the most relevant modality priority according to the situation. We used an FC layer to standardize the feature dimensions for fair attention allocation [6]. To ensure thorough emotion analysis, the gathered features were subsequently run through a SoftMax layer to categorize the underlying emotion. The goals of the suggested projects are:

- Disclosed a novel multimodal emotion-detection technique that outperformed the accuracy of the industry standard models. Future research on emotion detection will be made easier by this innovative method;
- Textual input was efficiently encoded using a CNN-BILSTM-based model that had been previously trained;
- For a bimodal system that prioritizes audio and video, fuse the feature vectors, dynamically prioritizes modalities, and uses multistage processing to improve classification accuracy, an attention-based fusion method was used.
- Extensive experiments were conducted on two benchmark datasets, YouTube and SAVEE, and the results clearly showed how much better our technique is at recognizing emotions.

This is how the rest of the document is organized: While the survey's existing emotion detection scheme is presented in Section 2, An overview of the suggested architecture is given in Section 3. A comparison of the findings from the numerous papers that are part of this taxonomy is given in Section 4. Section 5 brings the process to a close.

II. LITERATURE REVIEW

As the most fundamental and natural form of human communication, emotions are important in daily life. The newest problem in affective computing and human-computer interaction (HCI) is emotion recognition [7]. Although a lot of study has been done in this area over the last 20 years, issues with accuracy and real-time analysis persist. Emotion recognition is useful in a wide range of fields, such as marketing, e-tutoring, security, medical diagnostics, and intelligent automotive systems. This survey will outline the areas that most need emotion detection in addition to analyzing the literature on human emotion identification from speech, facial photos, and brain signals [8]. According to [9] employed facial cues for facial features that CNN subsequently adopted in order to enhance performance.

To identify emotions, extracted the faces in key frames from the entry FACE and SAVEE databases using the Viola-Jones face recognition technique. According to [10] suggested a CNN for dynamic emotion recognition that uses fuzzy fusion in two stages. A two-step fuzzy fusion method was developed by fusing fuzzy broad learning with canonical correlation analysis. isual characteristics that were connected to the locations of the 2D marker coordinates were employed by [11]. Principal Component Analysis was then applied to the Gaussian Classifier. The impact of altering the characteristics and techniques on the eNTERFACE'05 database was examined by [12]. In the visual channel, face tracking and geometric elements for facial emotion identification were taken into consideration. Next, it was discussed how to classify using a multiclass SVM.

In order to extract high-dimensional Gabor coefficients from facial images and obtain detailed texture and structural information for facial feature extraction, by [13] used Gabor filters with five scales and eight orientations. In recent years, CNN has been used to extract facial features from still images of individuals. In order to determine which model would perform best for tasks like facial analysis, emotion identification, or pattern classification, Venkataraman et al. examined the performance of CNNs, DNNs, LSTMs, and HMMs. In [14] used a fuzzy ARTMAP Neural Network (FAMNN)-based SAVEE database to identify emotions. Retrieved facial expressions, Patterns of Oriented Edge Magnitude (POEM) features, and Local Phase Quantization (LPQ) features from the BAUM-1s database.

Livingstone et al. tested the Cohen's and Kappa values using the RAVDESS database. According to [15] employed the Hidden Markov Model (HMM) as the classifier model for the video series. According to [16] employed hybrid feature representation to recognize facial expressions from a photo frame. They combined different degrees of SIFT features from a CNN model with deep learning. For [17] proposed a feature extraction method for facial emotion recognition by fusing facial depth and facial texture. Facial expressions in still photos are recognized using LSTM, CNN, Alexnet, and other techniques. Furthermore, one of the most crucial methods for removing facial features from dynamic photos is 3D CNN. A CNN-CTS LSTM network and local enhanced motion history image-based video face expression identification system was proposed by [18].

According to [19] recovered the features for speech emotion identification using MFCC and relative special features. Pitch, intensity, the first four formants, and their band characteristics are among the 20 audio qualities that discovered. In order to extract speech components, by [20] employed an autocorrelation technique that splits the input voice signals according to shift intervals. To enhance the performance, by [21] coupled pitch and intensity with the first 13 MFCC elements. In conclusion, speech emotion components can be divided into three categories: spectral, prosody, and voice quality. Prosodic features include things like pitch period, speech rate, amplitude energy, and others; voice quality factors include things like formant frequency and glottis parameters. The term spectrum-based features refer to the linear predictor coefficient, or MFCC.

Based on prosodic characteristics, pitch frequency, MSPS, and MFCC were used in chapter 3; however, it was demonstrated that MFCC was superior in detecting emotions. A method called hybrid feature fusion blends characteristics from several modalities to increase the precision of sentiment analysis. Hybrid feature fusion is able to capture more subtle and detailed characteristics of sentiment than can be achieved with a single modality by combining data from multiple modalities. The currently used text sentiment analysis approaches are unable to handle uncertain assertions, such as reviews that exhibit bipolar behaviour. Automating sentiment recognition in natural language processing is far more challenging than it is for humans.

Instead of sending in their reviews in writing, end customers would want to provide audio snippets that express their thoughts. However, due of the massive feature dimensionality of the existing SER, the most difficult problem is the high computing cost. Because people frequently express their ideas on video, the reviewer's facial expression is considered a critical indication for sentiment analysis. These days, managing the generated features from several modalities requires multimodal systems to employ strong feature optimization and feature fusion approaches. Among the drawbacks of the earlier models are weak convergence and a clear tendency to become caught in local optima.

III. MATERIALS AND METHODS

One of the newest technologies is emotional recognition, which has been used in the medical field to determine a patient's emotion for treatment in recent years. Many technologies now include facial recognition technology into their apps to enable efficient connection with people. Because it makes human interactions more realistic, fields have been focusing on integrating these systems' capacity to modify their reactions and behavior based on human emotions. Facial expressions, body motions, and machine learning applications are the main methods used in computer science for emotional recognition [22]. The study's emotional recognition technology has been enhanced by the hybrid model, which blends the Attention-Guided CNN-Bi-LSTM models (Figure 2).

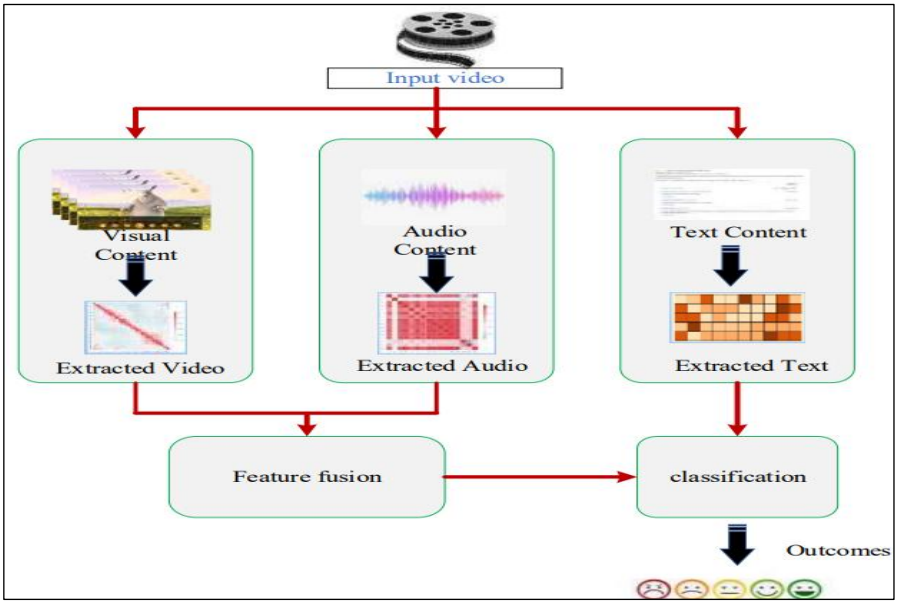

Figure 2: Proposed work.
Source: Authors, (2026).

III.1 DATASET

The SAVEE and YouTube datasets are used in the experiment utilizing the suggested methodology. There are 47 video clips in the YouTube collection, which includes product evaluations and political commentary. Of these datasets, twenty contain movies with female speakers, whereas the other 27 contain videos with male speakers (Morency et al., 2011). The speakers range in age from 14 to 60. Each video sample lasts two to five minutes and contains 360480 Mp4 pixels. Four male artists' 480 words from the SAVEE dataset represent a range of emotions. The visual media lab used state-of-the-art audiovisual technologies to construct the audio files in the SAVEE dataset [23]. Figure 3 shows sample frames from the SAVEE and YouTube datasets.

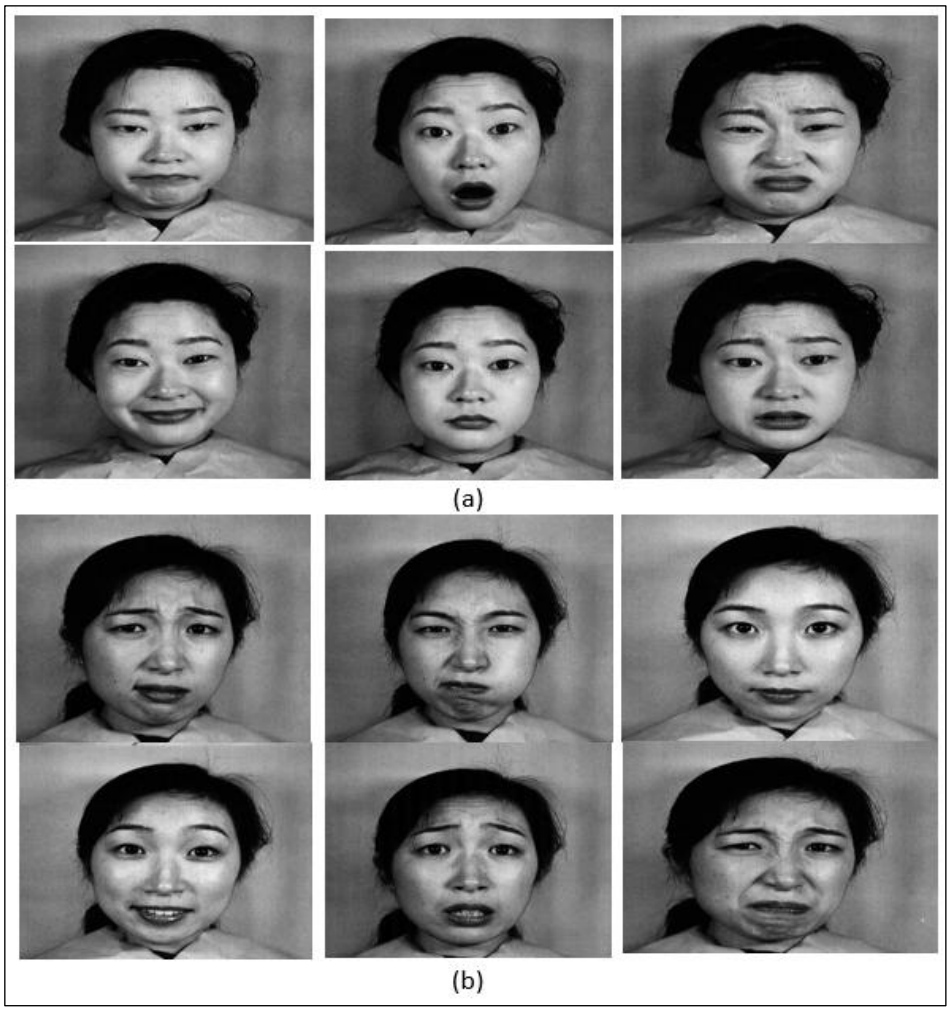

Figure 3: Sample Information from the YouTube and SAVEE datasets.
Source; Authors, (2026).

III.2 PRE-PROCESSING

Facial expression and emotion recognition are critical components of applications that must assess veracity, detect fraud, and gauge credibility. Unfortunately, most studies demonstrate little accuracy in recognizing emotions, mostly due to low-quality pictures. The application of artificial intelligence, especially deep learning techniques, and significant pre-processing are increasing the accuracy of computational predictions [24]. In order to identify emotions, our work blends deep learning techniques with preprocessing exercises. Our method aims to increase the accuracy of emotion recognition by comparing and using four deep learning models for image pre-processing [25]. The illumination and reflectance of the object or objects in the scene are multiplied to determine the intensity at any pixel in a picture, or the amount of light reflected by a point on the object.

$$I(x,y) = L(x,y) * R(x,y) \tag{1}$$

Where I, L, and R stand for the image, scene illumination, and scene reflectance, respectively. In order to convert the multiplicative components into additive components, this filtering technique moves to the log domain employing,

$$lnI(x,y) = ln(L(x,y)R(x,y)) \tag{2}$$

$$lnI(x,y) = ln(L(x,y) + R(x,y)) \tag{3}$$

In the log domain, the high-frequency reflectance component is preserved while the low-frequency light component is removed using a high-pass filter.

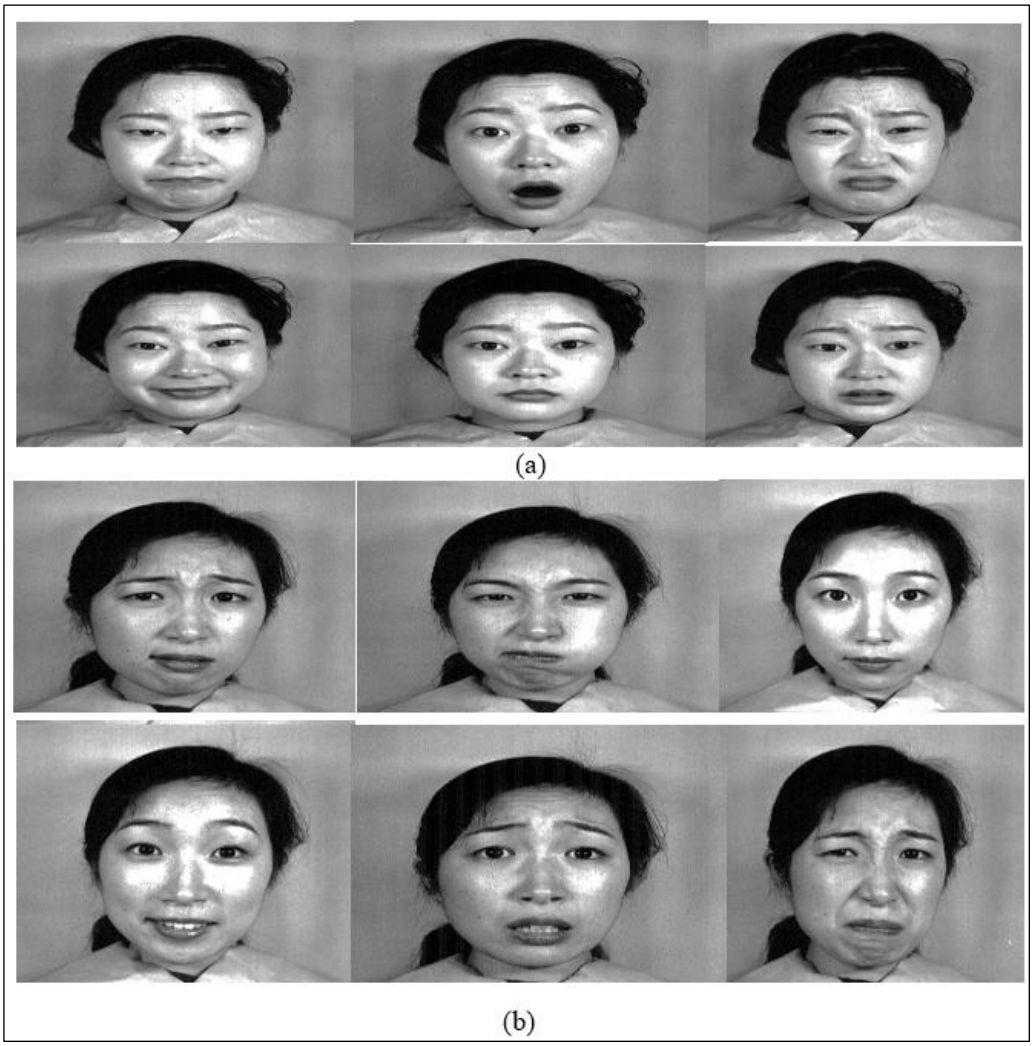

Figure 4: Pre-processing results.
Source; Authors, (2026).

A common method for handling nonuniform illumination in figure 4 is homomorphic filtering, a frequency domain filtering technique that enhances contrast from the object's reflecting qualities while compressing brightness from lighting circumstances.

III.3 FUTURE EXTRACTION AND FEATURE FUSION

The facial traits are extracted using a geometric feature-based approach by locating the characteristic feature points on the cropped face pictures. These attributes can be extracted manually or automatically using AAM tools [26]. A technique for equitably distributing gains or profits among several participants according to each person's unique contribution to the final result is the Shapley value. This idea, which comes from cooperative game theory, guarantees a fair and just distribution. The Shapley value is defined as follows:

$$\emptyset(x_i) = \Sigma_{s\subseteq\{1,2,\ldots,k\}\backslash\{i\}} \frac{|S|!(K-|S|-1)!}{K!} [f_x S\cup\{i\}) - f_x(S)], \quad (4)$$

Where K stands for the total number of participants. Equation (2)'s bracket section demonstrates that the entity's contribution can be characterized as a marginal contribution, or the difference between the group-S members' profit (f_x(S)) and the group members' and entity-i's profit: $[f_x S\cup\{i\})$. Friedman's gain is based on the same principle, but it can be altered by group members, which leads to inconsistencies. The Shapley value, which is then determined once more for each possible combination, is the average of the marginal contributions of all possible combinations. Only the Shapley value approach to profit allocation satisfies the four requirements of linearity, symmetry, efficiency, and null player. Utilizing the concept of the Shapley value, the SHapley Additive exPlanation (SHAP) illustrates the patient's outcome-j: $f(x^{(i)}$ as the sum of each features-i's contribution $\emptyset_i(x_i^{(j)})$.

$$\emptyset_0 = \frac{1}{N} \sum_{j=1}^{N} f(x^{(i)}) \quad (5)$$

$$\emptyset_i(x_i^{(j)}) = \emptyset(x_i^{(j)}) - \frac{1}{N}\sum_{K=1}^{N}\emptyset(x_i^{(K)}) \quad (6)$$

$$f(x^{(i)} = \emptyset_0 + \sum_{i=1}^{K}\emptyset_i\,(x_i^{(j)}) \quad (7)$$

Where N is the number of patients. We derived $\forall i, E(\emptyset(x_i)) = \frac{I}{N}\sum_{j=1}^{N}\emptyset_i(x_i^{(j)}) = 0$ from Eq. (7). The relationship between a feature x_i and SHAP value in GLM $\emptyset_{GLM}$ is given as follows [8]:

$$\emptyset_{GLM}(x_i^{(j)}) = a_i x_i^{(j)} - E(a_i x_i) = a_i x_i^{(j)}\, a_i E(x_i) \quad (8)$$

Feature x_i has a proportional relation with its SHAP value and the proportionality factor is given by coefficient a_i . This result is consistent with how GLM is currently interpreted. A SHAP dependence graphic illustrates the relationship between the characteristic and its impact on the SHAP-measured result. The logistic regression model converts SHAP values to log-odds for binary prediction. The linear relationship provided by Equation (6) is displayed in the SHAP dependence graphic for GLM.

$$IMP(X_i) = Var(\emptyset(x_i)) = \frac{1}{N}\sum_{j=1}^{N}[\,\emptyset(x_i^{(j)}) - E(\emptyset(x_i))]^2 \quad (9)$$

$$= \frac{1}{N}\sum_{j=1}^{N}[\,\emptyset(x_i^{(j)})]^2 - \{E(\emptyset(x_i))\}^2 \quad (10)$$

$$= \frac{1}{N}\sum_{j=1}^{N}[\emptyset(x_i^{(j)})]^2 \quad (11)$$

When all attributes are normalized, the variable significance of GLM in Eq. (11), using our idea, is as follows:

$$IMP_{GLM}(X_i) = \frac{1}{N}\sum_{j=1}^{N}[a_i\, x_i^{(j)} - a_i E(x_i)]^2 \quad (12)$$

$$= \frac{a_i^2}{N}\sum_{j=1}^{N}[x_i^{(j)} - E(x_i)]^2 \quad (13)$$

$$= |a_i|^2 .Var(x_i) = |\beta_i|^2 \quad (14)$$

Both the feature importance concept and the ranking based on absolute beta coefficients ($|\beta_i|$) are in line with your definition and the feature packing method that follows. This implies a unified method for prioritizing and evaluating features.

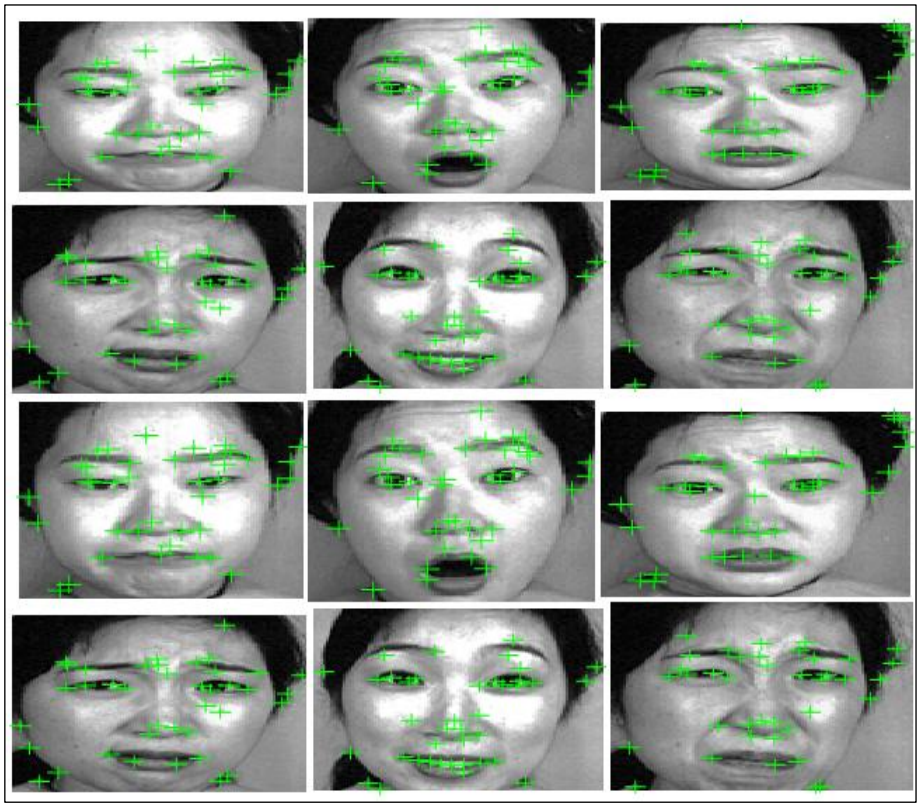

Figure 5: 39 feature points extracted.
Source: Authors, (2026).

The areas of the face with the most noticeable expressions, like the lips, nose, eyes, and eyebrows, contain 39 feature points, as shown in Figure 5. The quantity and location of the feature points are changed from. The locations of these feature points formed the feature vector. The feature vector recovered has 78 dimensions since 39 feature points are marked, and each feature point contains "x" and "y" coordinates. Because linked factors can complicate interpretation, it is usual practice to eliminate them from the data and build the predictor instead. However, this approach may result in greater reconstructive computation time and a poorer forecast accuracy. Therefore, without having to construct the predictor, we wish to combine these variables into a single grouped variable. Furthermore, feature packing has no impact whatsoever on prediction accuracy and does not necessitate model rebuilding. Therefore, packing qualities with many variables or similar interpretations is a highly helpful strategy.

III.4 CLASSIFICATION

The classification is done using the NN classifier, which is trained using the feature vectors of all emotions for the frontal position only. Following the mapping from a non-frontal posture to a frontal pose in the pose normalization stage using MSVR, the output $a*$ feature vector contains the position of the non-frontal pose's feature points onto the frontal pose. This is then input into NN, which classifies the emotion according on its label. If the head position estimation output is frontal posture, the feature vector is provided directly to the classifier without going through MSVR. To improve the emotion classification model's performance, the most pertinent audio clips and the features taken from the peak facial expression frame are combined [27].

The early feature level fusion is performed by concatenating the retrieved face action units of the peak visual frame with the most relevant audio segment characteristics. Fi is the length of the fused feature vector, where i = 1, 2..., n, and n = 1871. It has 1759 characteristics to describe the facial expressions of peak frames and 112 audio features (GTCCs features along their first and second order derivatives, spectral and time domain features). To prevent overfitting, the early halting technique was applied. The model's weights were changed if the validation loss was less than any minimal loss that had been documented before. When the training process was complete, the best model was ultimately retained.

III.4.1 Cnn-Bi-Lstm

In spectrograms, emotional expressions can occasionally appear as unique energy patterns, especially in auditory media like speech. These patterns can be identified using spectrogram analysis because they have unique energy distributions across a range of frequencies. Particularly effective at emphasizing data points with sudden changes or high energy levels are convolutional neural networks (CNNs). Given that CNNs are adept at recording emotions that emerge rapidly and intensely, this property is helpful for emotion analysis. In circumstances where emotions may abruptly surge or fall, this targeted analysis ensures a comprehensive awareness of the emotional environment and helps the model distinguish between strong and subtle emotional cues.

This method is very good at revealing little features in the frequency and temporal domains. With artificial neurons that can respond to a selection of neighbouring units, CNNs are one kind of feedforward neural network. It works best when processing images on a large scale. Utilized CNNs to identify handwritten numbers in the past century, but the restricted processing capacity of computers prevented them from being widely deployed. The benefits of CNNs have gradually become clear, though, and deep neural network training has gotten easier as processing power keeps increasing. As previously said, convolutional, pooling, and fully connected layers make up a CNN's network topology. Extracting local attributes from the input data is the primary function of the convolutional layer.

It is distinguished by weight sharing and sparse connection. In order to reduce the computational load and dimensionality of the feature maps while maintaining crucial feature information, the pooling layer is typically positioned after the convolutional layer. The fully connected layer aggregates the features extracted by the convolutional and pooling layers, applies linear transformations through weights and biases, and then performs nonlinear transformations by passing them through an activation function to produce. These functions have the benefits of minimizing overfitting and preventing gradient vanishing. The equation represents the function of ReLU.

$$Y = max(0, x) \tag{15}$$

The gradient is 0 for negative values even if the gradient-vanishing problem is fixed by the ReLU function, which could leave some neurons untrained and reduce the expressive power of the network. Leaky ReLU was created by Maas et al. to solve this issue. It smoothes and expresses the activation function by employing a modest slope for negative inputs rather than ReLU's zero slope. It has therefore been widely used in practical contexts. Equation (16) gives the definition of Leaky ReLU:

$$y = \begin{cases} x, & x \geq 0 \\ ax, & x < 0 \end{cases} \tag{16}$$

The linear classifier is often used in the traditional method to perform one layer of an operation; for example, SVM has only one weight set (Equation 2). CNN and other deep learning algorithms, however, carry out multiple levels of operation during classification. CNN's input is a two-dimensional array.

$$S = W \times x_i + b \tag{17}$$

Where b is bias, W is the weight matrix, and S is the classification score.

LSTM-based models are ideal for working with time series data, including audio data, because they are made to handle variable-length sequence data. Sequence data is handled sequentially by LSTM and concurrently backwards and forwards by Bi-LSTM. Bi-LSTM can therefore effectively evaluate both historical and prospective data simultaneously. This could lead to increased accuracy in pattern detection.

The Transformer model's attention mechanism weights various data points to draw attention to important features of the incoming data. This method is effective for recognizing emotions since certain speech patterns may be more important than others. Multi-head attention is utilized to capture more information than single-head focus.

$$z = \frac{x - \mu}{\sigma} \tag{18}$$

The forget gate layer's equation is provided as,

$$f_t = \sigma(W_f[h_{t-1}, x_t], b_f \tag{19}$$

The input gate layer comes next, where the additional features are used to retrain the recall state data.

$$i_t = \sigma(W_f[h_{t-1}, x_t], b_i \tag{20}$$

The output of the forget gate layer is multiplied by the cell state vector of the preceding LSTM cell (c_{t-1}). After being multiplied by the hidden state vector of the stage, the input gate layer's output is combined with the outcome of a "tanh" function. Together, they produce the cell state vector for the next LSTM cell.

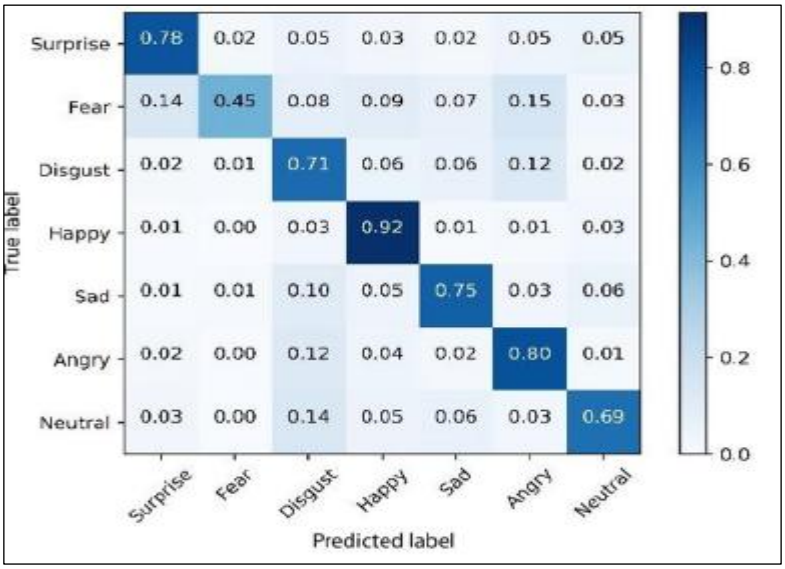

Figure 6: Confusion matrix results.
Source: Authors, (2026).

The prior concealed state vector is then multiplied by this new cell state following its passage via a "tanh" function (h_{t-1}). Our suggested model's whole architecture is depicted by

$$y = W_d X_t + B_d \tag{21}$$

Where,
y = output of the output layer, which is the target image's face emotion
X_t = the CNN-Bi LSTM model-learned vector
W_d B_d = bias and weight of the CNN-Bi LSTM model

Equation (28) defines the output or dense layer to which the fixed or variable length vector representation as the Bi-LSTM's output is fed in order to classify facial images into the different emotion classes of happy, sad, angry, disgusted, neutral, fear, and surprise. The multi-head attention mechanism can simultaneously identify many important patterns or qualities by employing numerous "heads" to calculate attention from different data viewpoints (Figure 6). Furthermore, the transformer structure makes parallel processing possible, which speeds up learning and simplifies the management of massive volumes of data. The combination of these two architectures enables the model to concurrently take into account a large number of patterns and characteristics in audio data since the transformer is skilled at learning the intricate hierarchical structures of data and the Bi-LSTM is good at identifying the temporal qualities of audio data.

IV. RESULT AND DISCUSSION

The CNN-Bi-LSTM model was implemented using TensorFlow and the Anaconda platform. An NVIDIA GTX 1650 Ti graphics card, 16 GB of RAM, and an Intel i7-10750H 2.6 GHz processor were used for the training and validation procedures. The CNN-Bi-LSTM model was trained using a batch size of two and a learning rate of 0.00005 through trial and error throughout the training phase. 168 of the 210 video examples are used for CNN-Bi-LSTM training, and 42 are validated using 5-fold cross validation.

Table 1: Proposed Model Performance Evaluation for YouTube Dataset.

Modality	Accuracy (%)	Sensitivity (%)	Specificity (%)	F-score (%)	MCC (%)
Audio	85.80	94.33	81.04	85.65	79.15
Text	93.25	94.30	96.52	91.42	90.75
Video	93.15	93.53	95.56	92.91	91.81
Video & Text	91.96	92.53	92.53	90.93	89.26
Audio & Text	93.16	95.54	91.96	90.39	85.71
Video & Audio	94.34	95.74	92.20	90.04	91.60
All Modalities	97.70	95.53	98.19	95.65	94.97

Source: Authors, (2026).

The CNN-Bi-LSTM model's effectiveness is assessed using a variety of modalities in Table 1. To provide the best fusion analysis results, the proposed model combines text, visual, and auditory modalities based on the observation. The YouTube dataset produces findings that are more closely related to both combined and individual sequences, with the best. Figure 7 displays the various modalities' performance on the YouTube dataset.

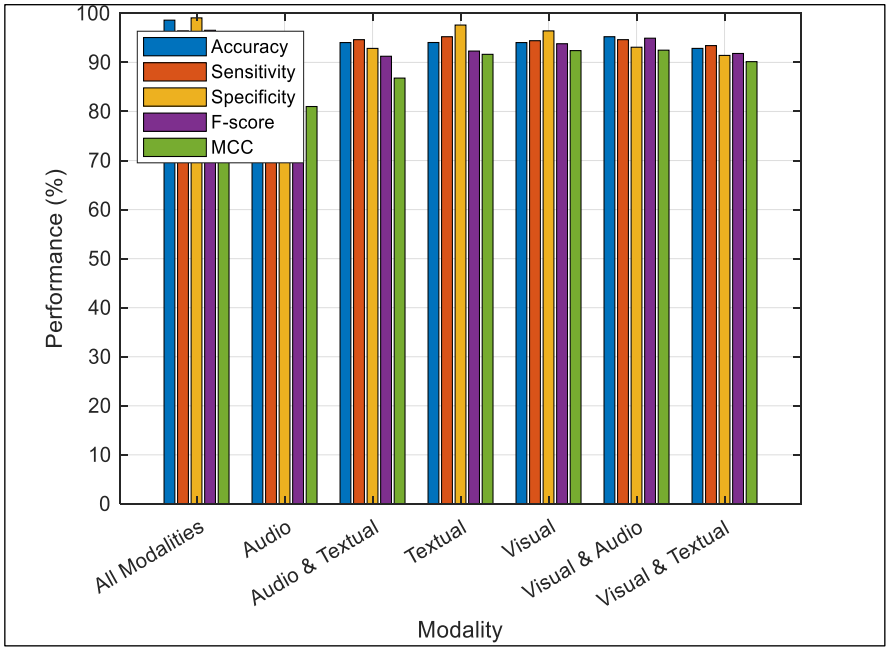

Figure 7: SHAP features Performance Measures on YouTube Dataset.
Source: Authors, (2026).

Furthermore, all three modalities are used to assess the proposed SHAP technique on the SAVEE dataset in MSA. The performance of all combined modalities on the SAVEE dataset is shown in Table 2.

Table 2: Performance Evaluation based on SAVEE Dataset.

Modality	Accuracy (%)	Sensitivity (%)	Specificity (%)	F-score (%)	MCC (%)
Audio	84.81	89.51	92.51	92.58	91.10
Textual	93.71	99.51	98.00	97.05	96.54
Visual	95.54	97.43	96.20	91.50	90.32
Visual & Textual	90.19	89.51	84.23	87.86	86.01
Audio & Textual	90.77	96.04	92.57	84.18	81.32
Visual & Audio	93.16	96.23	88.40	91.76	91.03
All Modalities	99.71	97.91	98.23	100.00	98.60

Source: Authors, (2026).

Combining the modalities according to MCC (99.71%), F-score (97.91%), Specificity (98.23%), Sensitivity (100%), and Accuracy (98.60%) yields the optimal response. Figure 8 displays the performance of the different modalities on the SAVEE dataset. To fuse all modalities using feature level fusion, the feature vectors from each modality are collected and used as the input for the classification phase. The integration of heterogeneous feature vectors, or feature level fusion, is the issue that the model that is currently being described attempts to solve.

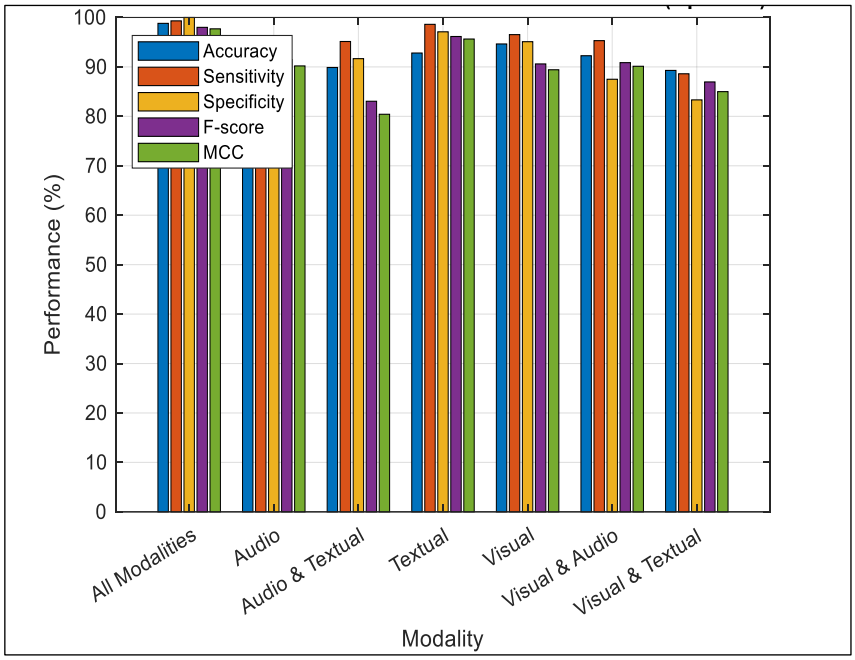

Figure 8: SAVEE Dataset Performance Measures.
Source: Authors, (2026).

Figure 9's output graphic breaks down high-dimensional feature vectors into two halves so you can see how similar two photos are to one another. More distant points imply larger differences, while closer points show comparable feature patterns. Understanding feature distribution, identifying outliers, and getting ready for tasks like clustering or classification are all made easier with the help of this visualization.

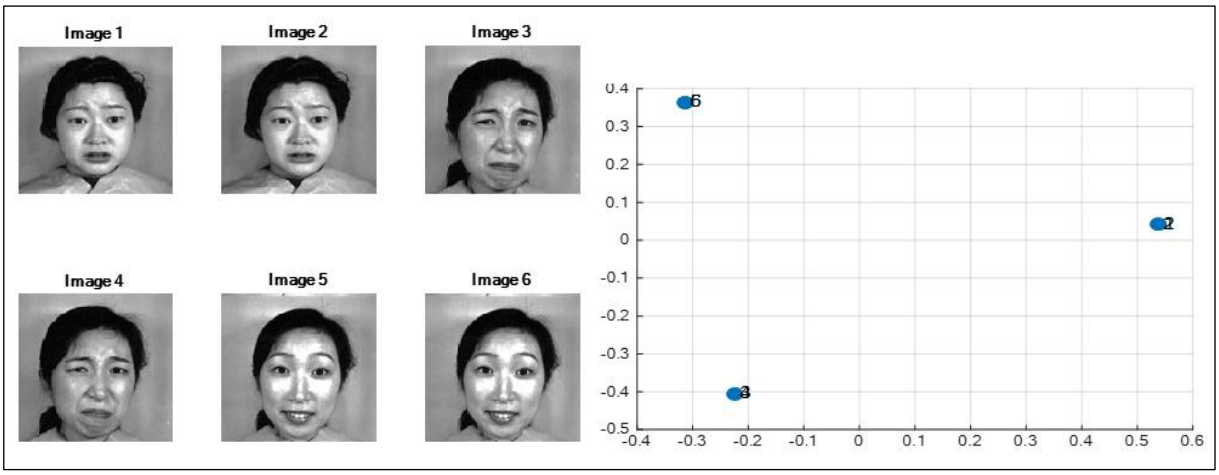

Figure 9: Feature fusion.
Source: Authors, (2026).

Following the collection of various features, the judicial feature standards were chosen using the relief strategy and subsequently submitted to the CNN-Bi-LSTM algorithm. Because the CNN-Bi-LSTM model uses a bi-directional long- and short-term memory neural network, it performs worse than the CNN-LSTM model in every temporal data processing error.

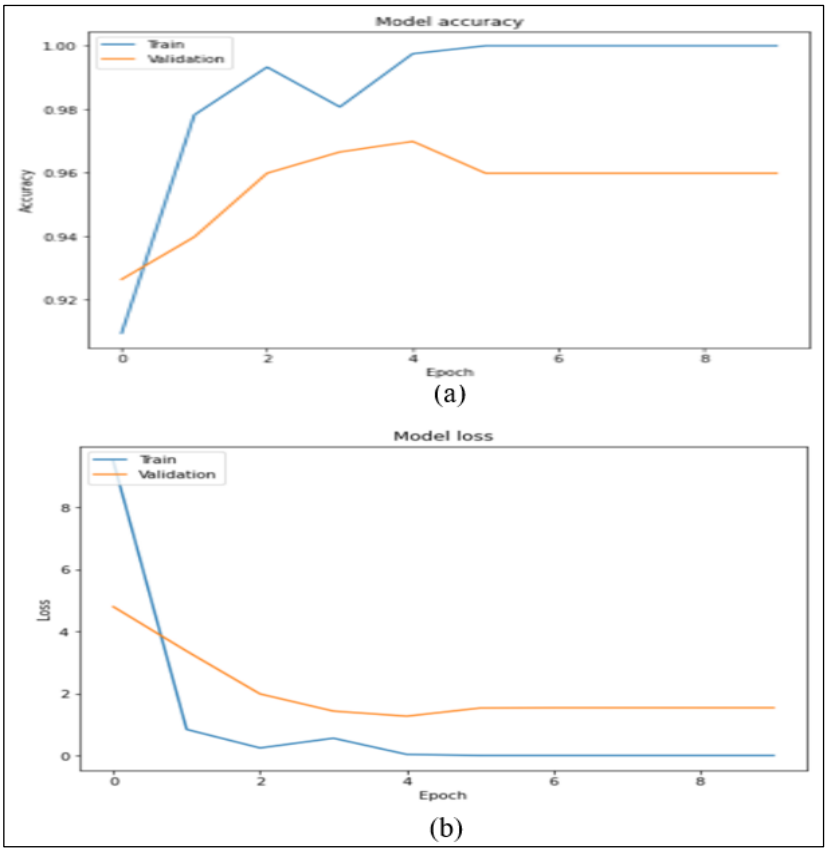

Figure 10: CNN-Bi-LSTM Accuracy and Loss plot.
Source: Authors, (2026).

First, of the four models, the CNN and LSTM models have the lowest prediction accuracies and the shortest average calculation time due to their comparatively simple architectures. Out of the six models, the enhanced CNN-Bi-LSTM-attention model has the lowest prediction error and the highest prediction accuracy (Figure 10). The CNN-Bi-LSTM-attention model has a more potent temporal data-processing capability than the CNN-Bi-LSTM model because it incorporates an attention mechanism that can handle some temporal features that the network missed and weigh the channels, allowing the model to focus on important temporal features. It can produce forecast accuracy with the least amount of error and has a high degree of generalizability (Table 3). CNN-LSTM-attention has a greater total error and a lower prediction accuracy because it cannot handle time-series data in both directions and cannot extract enough long-term features (Figure 11).

Table 3: Proposed dataset Comparative Analysis.

Metric	YouTube Dataset	SAVEE Dataset
Accuracy	98.60	98.20
Sensitivity	97.74	98.74
Specificity	97.14	97.14
Precision	96.77	97.77
NPV	98.33	99.33
FPR	2.86	2.86
FNR	5.26	5.26
F1 Score	95.74	95.74
MCC	91.99	91.99
Balanced Accuracy	98.94	98.94
Youden's Index	0.92	0.92

Source: Authors, (2026).

The CNN-Bi-LSTM hybrid model's ability to identify emotions is seen in Table 4.8, which compares classification models on two datasets. For better emotion recognition, this combined strategy makes use of the advantages of CNNs (feature extraction) and Bi-LSTMs (sequence modelling). The CNN-Bi-LSTM model showed enhanced sensitivity and accuracy on the YouTube dataset and SAVEE, which were utilized to assess its performance. The relief-random forest model is more accurate (94%) and sensitive (91%), as shown in Table 4.8. The YouTube data's sensitivity.

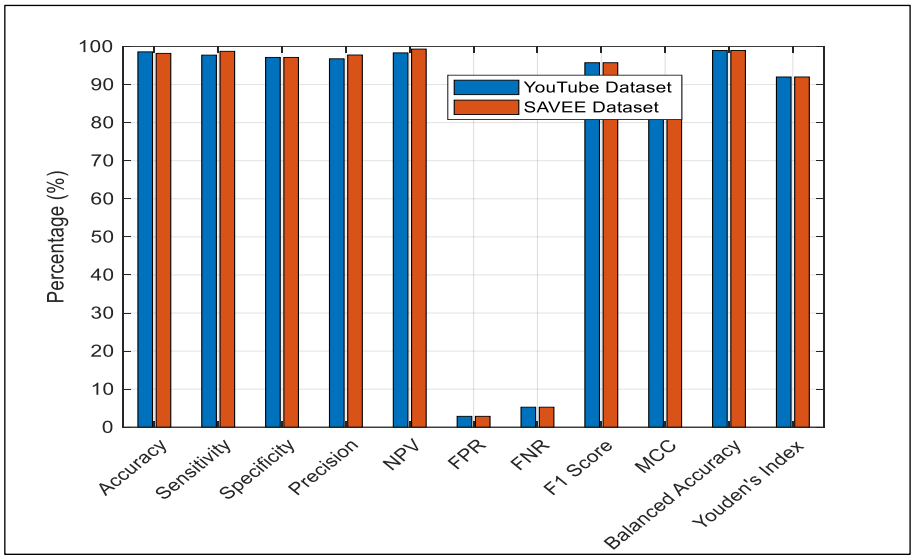

Figure 11: Proposed dataset Comparative Analysis.
Source: Authors, (2026).

The CNN and LSTM models have the fastest calculation times when it comes to model computational efficiency. They are the least precise, nevertheless, and have the worst computational error. The CNN-LSTM model has the second-highest computation time after the CNN and LSTM models because to its comparatively simple structure. The CNN-Bi-LSTM-attention model performs better than the CNN-Bi-LSTM-attention and CNN-LSTM-attention models in terms of computational efficiency and model prediction accuracy.

Table 4: Proposed dataset Comparative Analysis.

Models	Dataset	Accuracy (%)	Sensitivity (%)
MKL-CNN [28]	YouTube	88.60	87.36
INCA-SoftMax Classifier [29]	SAVEE	95	81
Relief-random Forest [30]	YouTube	94.01	91
CNN-Bi-LSTM model	**YouTube**	**98.60**	**97.74**
	SAVEE	**98.20**	**98.74**

Source: Authors, (2026).

Out of the six models tested, the CNN-Bi-LSTM-attention model performed the best, achieving the highest prediction accuracy and the lowest prediction error for both inbound and outbound predictions.

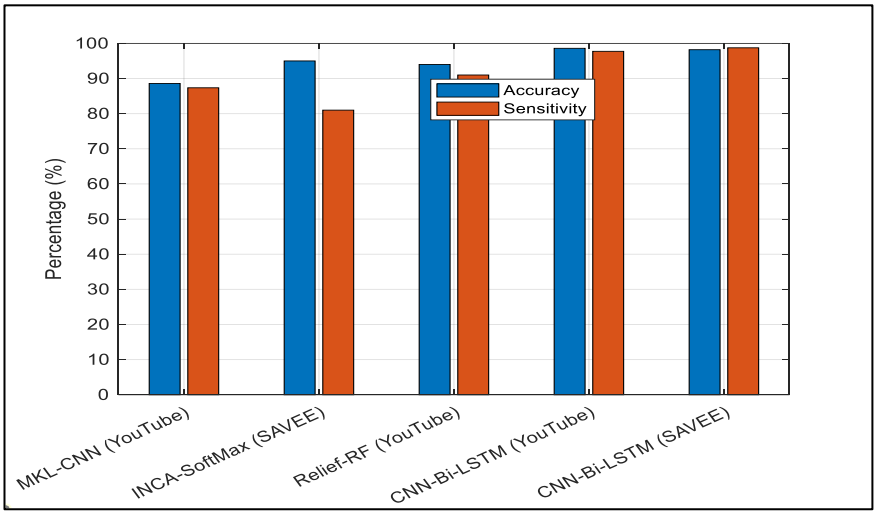

Figure 12: Proposed and Existing Methods Comparative Analysis.
Source: Authors, (2026).

Figure 12 and Table 4 demonstrate the great prediction efficiency of the CNN-Bi-LSTM-attention model. The accuracy of the departing and arriving inventory prediction is higher than that of the other models. Therefore, out of all the models currently in use for emotion recognition and experimental data, the CNN-Bi-LSTM-attention model is the most successful.

V. CONCLUSION

In order to create a reliable and effective emotion detection system, this study investigated a multimodal strategy that integrated text, speech, and video data. By utilizing a variety of emotional cues, it may be possible to increase accuracy and dependability. A comprehensive analysis of multimodal inputs is made possible by the architecture's integration of CNNs (for audio) and CNN-Bi-LSTM (for video), which effectively recognizes emotions by identifying subtle patterns in both audio and visual data. Future studies should concentrate on refining the model to lower processing requirements, enabling real-time processing, and guaranteeing deploy ability on devices with constrained computing capability in order to overcome the constraints of our multimodal emotion-recognition system and enhance its capabilities. Creating CNN-Bi-LSTM versions and distilling models are two techniques that could maintain good accuracy while consuming fewer resources.

Other modalities, such facial expressions and physiological clues, can also be used to improve the ability to identify emotions. This expansion would require the development of SHAP-based fusion techniques that can handle the diversity and complexity of data from several sources. It is essential to lessen dependence on huge annotated datasets. Methods such as few-shot learning, transfer learning, and synthetic data generation can increase robustness, boost data efficiency, and allow for efficient use in scenarios with limited data. As a result, this work provides a variety of potential avenues for further research. To make the system viable for real-time applications, more research is therefore required to create lightweight architectures or use model quantization and pruning strategies. Furthermore, by taking cultural differences in emotional manifestations into consideration, we hope to create broadly applicable models.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: J. Biju, Lavanya K, J. Raja, M. Kiruthiga Devi, Payala Krishnanjaneyulu and N. Kanya.
Methodology: Kiruthiga Devi, Payala Krishnanjaneyulu and N. Kanya.
Investigation: J. Biju, Lavanya K, J. Raja, M. Kiruthiga Devi, Payala Krishnanjaneyulu and N. Kanya.
Discussion of results: J. Biju.
Writing – Original Draft: Payala Krishnanjaneyulu and N. Kanya.
Writing – Review and Editing: J. Biju and Lavanya K.
Resources: J. Biju, Lavanya K, J. Raja, M. Kiruthiga Devi, Payala Krishnanjaneyulu and N. Kanya.
Supervision: J. Biju.
Approval of the final text: J. Biju, Lavanya K, J. Raja, M. Kiruthiga Devi, Payala Krishnanjaneyulu and N. Kanya.

VII. REFERENCES

[1] Zhao, Wenyu, Min Xia, Liguo Weng, Kai Hu, Haifeng Lin, Youke Zhang, and Ziheng Liu. "SPNet: Dual-Branch Network with Spatial Supplementary Information for Building and Water Segmentation of Remote Sensing Images." Remote Sensing 16, no. 17 (2024): 3161.

[2] Ursuleanu, Tudor Florin, Andreea Roxana Luca, Liliana Gheorghe, Roxana Grigorovici, Stefan Iancu, Maria Hlusneac, Cristina Preda, and Alexandru Grigorovici. "Deep learning application for analyzing of constituents and their correlations in the interpretations of medical images." Diagnostics 11, no. 8 (2021): 1373.

[3] Mohanty, Manas Ranjan, Pradeep Kumar Mallick, and Debahuti Mishra. "Bald eagle-optimized transformer networks with temporal-spatial mid-level features for pancreatic tumor classification." Biomedical Physics & Engineering Express 11, no. 3 (2025): 035019.

[4] Nerella, Subhash, Sabyasachi Bandyopadhyay, Jiaqing Zhang, Miguel Contreras, Scott Siegel, Aysegul Bumin, Brandon Silva et al. "Transformers in healthcare: A survey." arXiv preprint arXiv:2307.00067 (2023).

[5] Su, Pin-Chen, and Mau-Tsuen Yang. "Integrating Depth-Based and Deep Learning Techniques for Real-Time Video Matting without Green Screens." Electronics 13, no. 16 (2024): 3182.

[6] Yu, Ying, Zhen Cai, Duoqian Miao, Jin Qian, and Hong Tang. "An interactive network based on transformer for multimodal crowd counting." Applied Intelligence 53, no. 19 (2023): 22602-22614.

[7] Ouzar, Yassine, Frédéric Bousefsaf, Djamaleddine Djeldjli, and Choubeila Maaoui. "Video-based multimodal spontaneous emotion recognition using facial expressions and physiological signals." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2460-2469. 2022.

[8] Ezzameli, Kaouther, and Hela Mahersia. "Emotion recognition from unimodal to multimodal analysis: A review." Information Fusion 99 (2023): 101847.

[9] Wang, Zhongmin, Xiaoxiao Zhou, Wenlang Wang, and Chen Liang. "Emotion recognition using multimodal deep learning in multiple psychophysiological signals and video." International Journal of Machine Learning and Cybernetics 11, no. 4 (2020): 923-934.

[10] Bhattacharya, Prasanta, Raj Kumar Gupta, and Yinping Yang. "Exploring the contextual factors affecting multimodal emotion recognition in videos." IEEE Transactions on Affective Computing 14, no. 2 (2021): 1547-1557.

[11] De Silva, Liyanage C., Tsutomu Miyasato, and Ryohei Nakatsu. "Facial emotion recognition using multi-modal information." In Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications (Cat., vol. 1, pp. 397-401. IEEE, 1997.

[12] Ranganathan, Hiranmayi, Shayok Chakraborty, and Sethuraman Panchanathan. "Multimodal emotion recognition using deep learning architectures." In 2016 IEEE winter conference on applications of computer vision (WACV), pp. 1-9. IEEE, 2016.

[13] Kahou, Samira Ebrahimi, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean et al. "Emonets: Multimodal deep learning approaches for emotion recognition in video." Journal on Multimodal User Interfaces 10 (2016): 99-111.

[14] Poria, Soujanya, Iti Chaturvedi, Erik Cambria, and Amir Hussain. "Convolutional MKL based multimodal emotion recognition and sentiment analysis." In 2016 IEEE 16th international conference on data mining (ICDM), pp. 439-448. IEEE, 2016.

[15] Soleymani, Mohammad, Maja Pantic, and Thierry Pun. "Multimodal emotion recognition in response to videos." IEEE transactions on affective computing 3, no. 2 (2011): 211-223.

[16] Cui, Can, Haichun Yang, Yaohong Wang, Shilin Zhao, Zuhayr Asad, Lori A. Coburn, Keith T. Wilson, Bennett A. Landman, and Yuankai Huo. "Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review." Progress in Biomedical Engineering 5, no. 2 (2023): 022001.

[17] Preethi, S. J. "Analyzing lower half facial gestures for lip reading applications: Survey on vision techniques." Computer Vision and Image Understanding 233 (2023): 103738.

[18] Chen, Hongyan, Zhiwei Li, and Xinjie Xiao. "An Unsupervised Anomaly Detection Method for Railway Fasteners Based on Knowledge-Distilled Generative Adversarial Networks." Applied Sciences 15, no. 11 (2025): 5933.

[19] Jia, Wenchao, Bin Gong, Yaoke Shi, and Zhipeng Luo. "An efficient lightweight gearbox fault diagnosis method based on parameter optimization and adaptive attention enhancement across working conditions." Measurement Science and Technology 36, no. 5 (2025): 056109.

[20] Song, Siyang, Yupeng Huo, Shiqing Tang, Jiaee Cheong, Rui Gao, Michel Valstar, and Hatice Gunes. "Automatic Depression Assessment using Machine Learning: A Comprehensive Survey." arXiv preprint arXiv:2506.18915 (2025).

[21] Yang, Yang, Zhilei Wu, Yuexiang Yang, Shuangshuang Lian, Fengjie Guo, and Zhiwei Wang. "A survey of information extraction based on deep learning." Applied Sciences 12, no. 19 (2022): 9691.

[22] Wang, Min, Hongbin Chen, Dingcai Shen, Baolei Li, and Shiyu Hu. "RSRNeT: a novel multi-modal network framework for named entity recognition and relation extraction." PeerJ Computer Science 10 (2024): e1856.

[23] Deshpande, Kedar, Manjit Singh Sodhi, Nidhi Raniyer, and Madhav Rao. "A Time-Distributed CNN-LSTM with Attention Model for Speech Based Emotion Recognition." In Proceedings of the 2024 7th International Conference on Digital Medicine and Image Processing, pp. 67-71. 2024.

[24] Feng, Tianzhi, Chennan Wu, Yi Niu, Fu Li, Yang Li, Boxun Fu, Zhifu Zhao, and Xiaotian Wang. "Adaptive Progressive Attention Graph Neural Network for EEG Emotion Recognition." arXiv preprint arXiv:2501.14246 (2025).

[25] J. Anand, V. Thamilarasi, A. Rayal, H. K. Gupta, K. Jyothi and P. Vishwakarma, "Fuzzy Logic-Based Deep Learning for Human-Machine Interaction and Gesture Recognition in Uncertain and Noisy Environments," 2025 First International Conference on Advances in Computer Science, Electrical, Electronics, and Communication Technologies (CE2CT), Bhimtal, Nainital, India, 2025, pp. 435-440, doi: 10.1109/CE2CT64011.2025.10939160.

[26] Kundu, Niloy Kumar, Sarah Kobir, Md Rayhan Ahmed, Tahmina Aktar, and Niloya Roy. "Enhanced Speech Emotion Recognition with Efficient Channel Attention Guided Deep CNN-BiLSTM Framework." arXiv preprint arXiv:2412.10011 (2024).

[27] Jia, Ning, Chunjun Zheng, and Wei Sun. "A multimodal emotion recognition model integrating speech, video and MoCAP." Multimedia Tools and Applications 81, no. 22 (2022): 32265-32286.

[28] Bahreini, Kiavash, Rob Nadolski, and Wim Westera. "Towards multimodal emotion recognition in e-learning environments." Interactive Learning Environments 24, no. 3 (2016): 590-605.

[29] Xue, Jianfei, Zhaojie Luo, Koji Eguchi, Tetsuya Takiguchi, and Tsukasa Omoto. "A Bayesian nonparametric multimodal data modeling framework for video emotion recognition." In 2017 IEEE International Conference on Multimedia and Expo (ICME), pp. 601-606. IEEE, 2017.

[30] Ouzar, Yassine, Frédéric Bousefsaf, Djamaleddine Djeldjli, and Choubeila Maaoui. "Video-based multimodal spontaneous emotion recognition using facial expressions and physiological signals." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2460-2469. 2022.