A Deep Learning Framework with Optimizations for Facial Expression and Emotion Recognition from Videos

Original Scientific Paper

Ranjit Kumar Nukathati*
Department of Computer Science and Engineering, JNTUA,
Anantapur, Andhra Pradesh, India
e-mail: ranjitnukathati@gmail.com

Uday Bhaskar Nagella
Government College (A),
Anantapur, Andhra Pradesh, India
e-mail: udaynagella@gmail.com

AP Siva Kumar
Department of Computer Science and Engineering, JNTUA,
Anantapur, Andhra Pradesh, India
e-mail: sivakumar.ap@gmail.com

*Corresponding author

Abstract *– Human emotion recognition has many real-time applications in healthcare and psychology domains. Due to the widespread usage of smartphones, large volumes of video content are being produced. A video can have both audio and video frames in the form of images. With the advancements in Artificial Intelligence (AI), there has been significant improvement in the development of computer vision applications.Accuracy in recognizing human emotions from given audio-visual content is a very challenging problem. However, with the improvements in deep learning techniques,analyzing audio-visual content towards emotion recognition is possible. The existing deep learning methods focused on audio content or video frames for emotion recognition. An integrated approach consisting of audio and video frames in a single framework is needed to leverage efficiency. This paper proposes a deep learning framework with specific optimizations for facial expression and emotion recognition from videos. We proposed an algorithm, Learning Human Emotion Recognition (LbHER), which exploits hybrid deep learning models that could process audio and video frames toward emotion recognition. Our empirical study with a benchmark dataset, IEMOCAP, has revealed that the proposed framework and the underlying algorithm could leverage state-of-the-art human emotion recognition. Our experimental results showed that the proposed algorithm outperformed many existing models with the highest average accuracy of 94.66%. Our framework can be integrated into existing computer vision applications to recognize emotions from videos automatically.*

Keywords: *Emotion Recognition, Spatial Expression Analysis, Deep Learning, Artificial Intelligence, Hyperparameter Tuning*

Received: June 11, 2024; Received in revised form: August 23, 2024; Accepted: August 30, 2024

1. INTRODUCTION

Human emotion recognition is the skill of interpersonal relationships. Therefore, it has a vital role in day-to-day communications. Humanscan intuitively understand communication through text, audio, and facial expressions. The ability to recognize emotions depends on the level of perception. With the advancements in Artificial Intelligence (AI), there has been an increased number of computer vision applications used to solve problems in the real world. With the help of deep learning techniques, solutions are being provided for various issues. Automatic recognition of human emotions is one of the challenging problems researchers have considered. However, recognizing emotions accurately is a problematic phenomenon. Many researchers contributed to building deep learning models meant for emotion recognition.

Profiled sentiment analysis is made possible by profound learning advancements. The state of the art in AI facial expression recognition is reviewed in this study

[1]. Provided a wealth of user-generated material for studying emotions, it is made more accessiblefor recognizing emotions. A 72% accurate approach based on facial expressions is suggested for automated video subtitle annotation [2]. LLEC is a model used for emotion cognition using entropy and similarity models on unlabeled data. Experiments show that enhanced LLEC significantly increases emotion recognition accuracy [3]. Deep reinforcement learning and algorithm optimization are among the tasks that lie ahead. Emotion identification is enhanced by this end-to-end method without the need for human feature engineering. Upcoming research will refine deep learning models for EEG-based emotion identification and enhance cross-subject categorization [4]. It was observed from the literature that most of the existing works considered textual content or audio or video. There is a need for an integrated approach that considers audio and video information processing to efficiently recognize human emotions. The contributions in this paper are as follows.

1. We proposed a deep learning framework with specific optimizations for facial expression and emotion recognition from videos.
2. We proposed a Learning-based Human Emotion Recognition (LbHER) algorithm that exploits hybrid deep learning models that could process audio and video frames towards emotion recognition.
3. Our empirical study, which used a prototype and the IEMOCAP benchmark dataset, has revealed the significance of our hybrid deep learning methodology.

The remainder of the paper is structured as follows—section 2 reviews prior worksabout human emotion recognition using deep learning models. Section 3 presents our methodology for efficiently detecting human emotions using hybrid deep learning models. Section 4 presents the results of our empirical study. Section 5 discusses the research findings in this paper and provides the study's limitations. Section 6 concludes our research work, besides giving directions for future research.

2. RELATED WORK

Human emotion recognition is an important research area that has attracted many researchers across the globe. Zhang *et al.* [1] profiled sentiment analysis is made possible by profound learning advancements. The state of the art in AI facial expression recognition is reviewed in this study. Villegas-Ch *et al.* [2] provided a wealth of user-generated material for the study of emotions. A 72% accurate approach based on facial expressions is suggested for automated video subtitle annotation. Casado *et al.* [3] presented LLEC for emotion cognition using entropy and similarity models on unlabeled data. Experiments show that enhanced LLEC dramatically increases the accuracy of emotion recognition. Deep reinforcement learning and algorithm optimization are among the tasks that lie ahead. Hassouneh *et al.* [4] used deep CNN for EEG emotional feature learning; the proposed technique outperforms conventional classifiers on the DEAP dataset by 3.58%. Emotion identification is enhanced by this end-to-end method without the need for human feature engineering. Upcoming research will refine deep learning models for EEG-based emotion identification and enhance cross-subject categorization. Pise *et al.* [5] recognized emotions thanks to recent developments in information fusion and machine learning, especially when using EEG signals for accurate emotion detection. Building higher-dimensional emotion models and enhancing the techniques for classifying emotion-related datasets are examples of future studies. Patel *et al.* [6] evaluated student gestures to identify emotions and provide instantaneous feedback to enhance instruction. However, identifying nuanced emotions and dealing with skewed data present hurdles for AI. Bazgir *et al.* [7] harmed by depression. It's critical to discover early. A new technique that shows promise for depression screening uses face recordings to extract physiological information. Anbarjafari *et al.* [8] used facial landmarks and EEG data; research focuses on real-time emotion identification for physically disabled people and children with autism. Diamantini *et al.* [9] addressed the absence of online learning. The accuracy of the suggested deep learning model for e-learning's face emotion identification is enormous.

Revina *et al.* [10] used EEG data, an emotion detection system was created, broken down into frequency bands, and then categorized using SVM with 91.3% accuracy— Cimtay *et al.* [11] with compound emotions like happily-disgusted, affective computing improvements in emotion recognition. The 50-category iCV-MEFED dataset aids research. Kumar *et al.* [12] examined the phases, functionality, databases, and applications of FER approaches. Social communication relies heavily on Face Expression Recognition (FER).

Zulfiqar *et al.* [13], with 81.2% accuracy on LUMED-2 and 91.5% accuracy on DEAP datasets, multimodal emotion recognition incorporates facial expressions, EEG, and GSR. Topic and Russo [14], automated face identification has become increasingly important with the growth of picture databases. The methods, difficulties, and applications are covered in this overview. Chen *et al.* [15], with uses like biometric authentication and video monitoring, say that facial recognition is becoming increasingly important. The recognition accuracy of a CNN-based system is 98.76%. Song *et al.* [16] Because of noise, interpreting EEG signals for emotions is complex. Deep learning, HOLO-FM, and TOPO-FM improve the recognition of datasets. The approach could help the realms of medicine and authentication. Future objectives are to enhance cross-validation and add more features.

Khan *et al.* [17] presented a method for recognizing emotions called the Multi-Modal Physiological Emotion Database (MPED). A new A-LSTM technique en-

hances emotion identification feature extraction. The database is open to the public for use in research. Chen *et al.* [18] presented face recognition smart glasses that help with security by providing a 98% detection rate to identify offenders using Haar-like characteristics. With immense accuracy, it uses Convolutional Neural Networks (CNN) for facial recognition. Parui *et al.* [19] proposed an approach that combines linear reconstruction with FRI theory to represent pictures as piecewise smooth functions for single-image super-resolution. It is superior to current techniques. Zheng *et al.* [20] suggested that the Emotion Meter achieves 85.11% accuracy in multimodal emotion identification by integrating EEG and ocular movements. EEG is best at cheerful, fearful eye movements.

Qing *et al.* [21] presented a machine-learning approach to interpretable emotion identification from EEG data. Emotional activation curves utilizing entropy and correlation coefficients are suggested to improve emotion identification accuracy. Gandhi *et al.* [22] depend heavily on sentiment analysis (SA) with both AI and NLP. Sentiment identification in text and videos is improved by Multimodal Sentiment Analysis (MSA), which is investigated in eleven fusion categories employing machine learning and deep learning. Sarkar and Etemad [23], by acquiring representations through pretext tasks, self-supervised deep multi-task learning improves ECG-based emotion identification and achieves state-of-the-art performance across datasets. Ayata *et al.* [24] presented a framework for music selection that enhances the functionality of current systems by utilizing wearable sensors to identify user moods. He *et al.* [25] presented a novel NIR-VIS facial image generation method that streamlines the procedure and raises the accuracy of HFR.

Feng *et al.* [26] explained an automated technique that uses EDA signals to categorize children's emotions. The dataset includes one hundred children's recordings with annotations for acceptance, boredom, and joy. Comparatively speaking, time-frequency analysis with CMorlet wavelets enhances SVM classifier performance. Upcoming projects will focus on real-time processing and growing datasets to capture more diverse emotional patterns. Tolosana *et al.* [27] discussed the rise of lifelike fake content and looked at DeepFakes, datasets, facial modification methods, and benchmarks. Real-world detection poses obstacles, which has led to research on generalization and fusion methods. Hazarika *et al.* [28] suggested employing pre-trained dialogue models to facilitate transfer learning for emotion identification in discussions. Experiments demonstrate enhanced robustness and performance—Baltrusaitis*et al.* [29] integrated data from several senses, known as multimodal machine learning. In classifying the field's obstacles, this poll highlights the promise of co-learning. Davison *et al.* [30] presented a 3D HOG technique for micro-expression detection verified on the SAMM and CASME II datasets. It focuses on 26 FACS-based areas and performs better than current techniques. Future efforts will focus on enhancing sensitivity and quickness. It was observed from the literature that most of the existing works considered textual content or audio or video. There is a need for an integrated approach that considers audio and video information processing to recognize human emotions efficiently.

3. PROPOSED FRAMEWORK

The section presents the proposed methodology, including a deep learning-based framework, preprocessing approaches, proposed algorithm, and evaluation methodology.

3.1. PROBLEM DEFINITION

If any given test video is provided, developing a deep learning-based framework that exploits audio and video frames with a hybrid deep learning approach toward automatic recognition of human emotions is a challenging problem.

3.2. OUR FRAMEWORK

As shown in Fig. 1, we proposed a deep learning-based framework for human emotion recognition. The framework is based on supervised learning, exploiting hybrid learning models for emotion recognition from given video content. The given data set is subjected to pre-processing, which includes a specific methodology, as illustrated in Section 3.3 and Section 3.4. After completion of pre-processing, the dataset is divided into training and test sets of 80% and 20%, respectively. The hybrid deep learning model proposed in this paper is trained with the training data. The model is persisted for future reuse and incremental learning or transfer learning. The saved model is loaded, and test samples will be subjected to emotion recognition. The experimental results are then compared with the ground truth to evaluate the proposed framework.

The proposed framework is based on a hybrid deep learning approach considering multiple modalities while processing the video content. As illustrated in Fig. 2, the proposed framework exploits audio content and video frames from the given input video. From the audio content, an audio spectrogram is generated. The video frames and the audio spectrogram are used to train the hybrid deep learning model. The pre-trained hybrid deep learning model performs emotion recognition from a given test video. Since it is a supervised learning process, it includes training and testing phases. In the training phase, the hybrid deep learning model is trained with 80% of training data to gain the required knowledge. Any given test video is subjected to an emotion recognition process in the testing phase by considering both audio content and video frames. Eventually, the proposed framework can classify emotions into happy, sad, angry, and neutral.

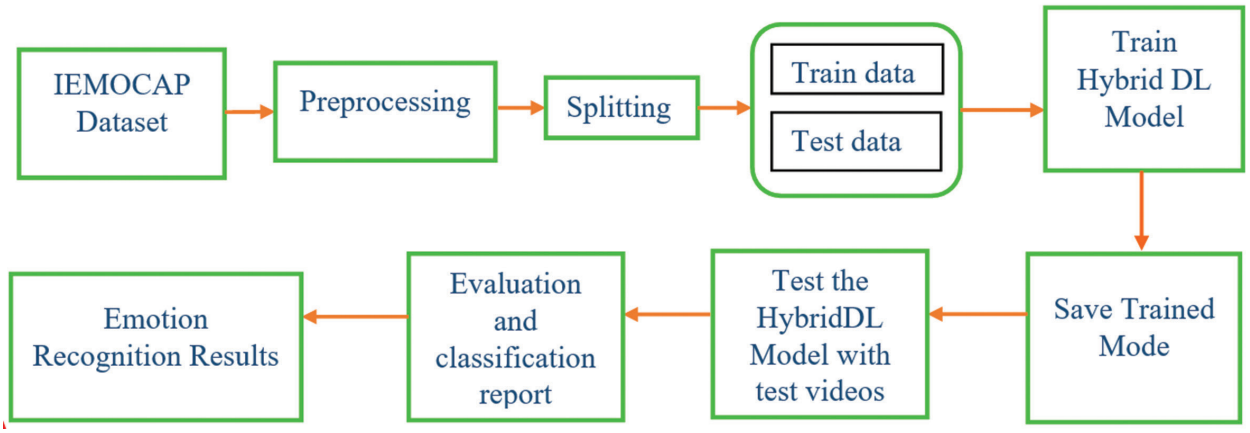

Fig. 1. The proposed deep learning-based framework

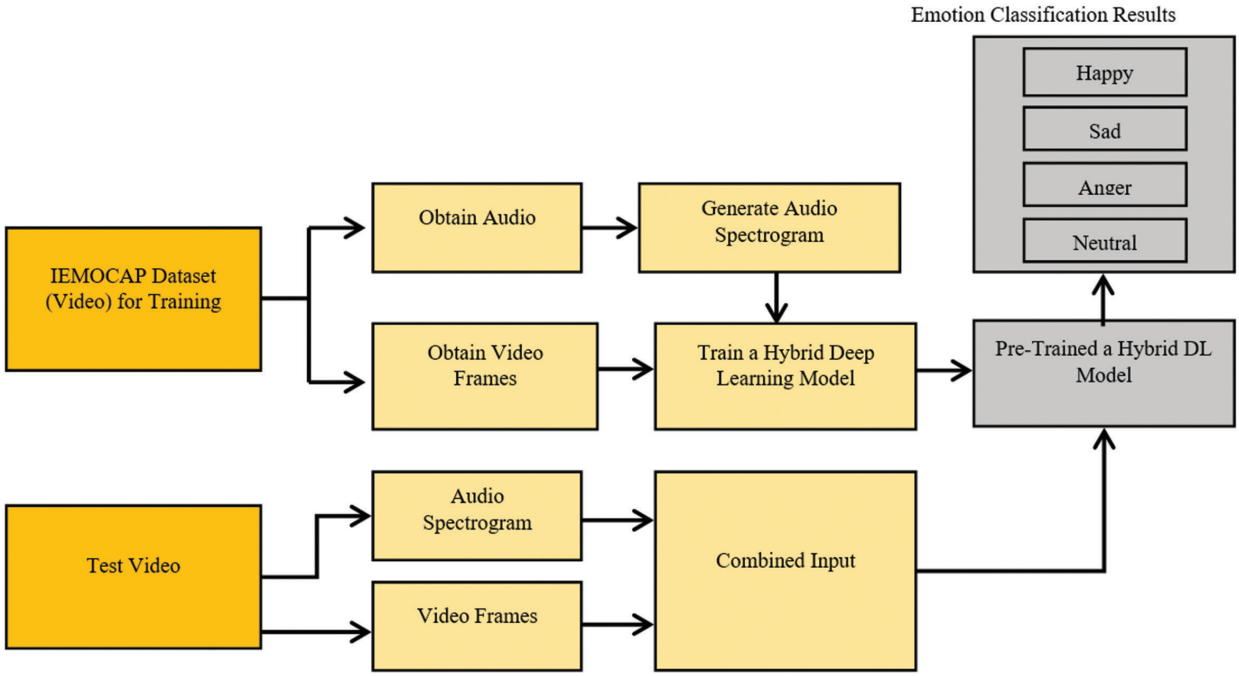

Fig. 2. Illustrates the training and testing phases involved in the proposed framework

A benchmark dataset, IEMOCAP, contains labeled data meant for supervised learning. The dataset is divided into a labeled training set and an unlabeled test set. The labeled data is used to train the hybrid deep learning model, while the unlabeled data is used to test the proposed hybrid deep learning model's performance.

3.3. PREPROCESSING AUDIO DATA

The IEMOCAP data corpus comprises audio wave files of varying durations, each labeled with the honest emotion for the relevant time segment. IEMOCAP generates its audio waves at a 22 KHz sample rate. Using the librosa1 Python library and a 44KHz sample rate, the audio spectrogram is recovered from the WAV file. 44 KHz sample rate was chosen because, according to the Nyquist-Shannon sampling theorem, the sampling frequency must be at least twice the signal frequency to recover the signal correctly. 20Hz to 20KHz is the frequency range of the audio signal. Therefore, the most widely utilized sampling rate is 44 KHz. The spectrograms were created in two parts: the initial duration of the speech or emotion and divided into three-second halves. Data segmentation was also carried out both with and without noise removal.

We used a bandpass filter ranging from 1Hz to 30KHz to eliminate the ambient noise. Work in [31] is also followed by denoising, or noise cleaning, of the input audio stream for data augmentation. To ensure consistency in noise level and frequency relative to other signal sections, noise is injected into sentence utterances lasting less than three seconds. The original audio signal was distorted when noise was introduced with a signal-to-noise ratio (SNR) of 1 throughout the signal duration. Zero padding was also experimented with to have a 3-second time scale. After that, the signal is denoised. We then denoise the resultant signal. To improve prediction accuracy for each emotion, denoising

enhances the visibility of the input audio signal's frequency, time scale, and amplitude components. Audio spectrograms are constructed using the same color bar intensity scale (+/- 60dB) to preserve the spectrum analysis's consistency that spans various emotional states. Normalization of data is comparable to this. The signal with the accurate information remains with a high power intensity or signal amplitude after denoising. The power level of some places in the spectrogram is lower than that of the actual signal of interest. In contrast, some signal strength is seen across the time scale, which, throughout the time scale, a signal intensity, that is, noise, is seen. The resulting spectrogram pictures have a pixel size of 200x300.

The cheerful emotion count is noticeably low, as can be seen. So, to get the final figure of 1600, we repeated the joyful data. The emotion count for fury was likewise replicated. The number of data points for the sad and neutral emotions was lowered to 1600 for each. The model is trained using a total of 6400 photos. Equilibrium data is essential for practical model training. To validate the model, 400 pictures representing each emotion are employed. The training set never contains the photos used for validation. Before realizing that including axis and scale may harm prediction accuracy, we began using audio spectrograms with xy axis and colorbar scale. Rotation and cropping of the input audio spectrograms were done to see an improvement in class accuracy. Every picture was reduced to 200x300 pixels and cropped by 10 pixels at the top. This cropping is done to mimic a little shift in emotion frequency. Likewise, a +/-10-degree rotation was applied to every picture. This rotation modifies the time scale in addition to simulating frequency changes. The rotation was done to a minor degree of 10 degrees because augmenting data that affects the temporal scale is not preferable. After cropping and rotation, 19200 total data points are included in the training set. Using both the original and data-augmented pictures for comparison, the model was trained independently. Images were not horizontally flipped since doing so would cause the timeframe to be flipped and simulate someone speaking backward, reducing the accuracy of the model's predictions.

Executing separate model training on an audio spectrogram with a complete duration of not only three seconds was necessary. The entire time length spectrogram was substituted for the provided 3-second audio spectrogram, preserving the data needed for balance. Approximately one hundred audio spectrograms were visually analyzed. The highest frequency found in all of these spectrograms was found to be around 8 KHz. This indicates that about 60% of the spectrogram picture is blue and contains no emotional information. Every supplied audio spectrogram was scaled to 200 by 300 pixels after being 60% chopped from the top. If the frequency range is known beforehand, creating spectrograms with a defined frequency scale would be the best action.

3.4. PREPROCESSING VIDEO DATA

We also performed video data pre-processing because part of our study involves creating a video model to evaluate where forecast accuracy of emotion identification might be improved. We first divided each video file into sentences using the same method as the audio files to handle the video data. This ensured that the video file we searched matched the specified audio spectrogram. Next, from each video avi file, we retrieved 20 pictures every 3 seconds, which matched the 3-second audio spectrum. Since the film has two performers, the frames were cropped to the appropriate left or right to only include the actor whose emotion was being captured. After that, we further cropped the video frames to hide the actor's head and face. The video frames have a final resolution of 60 x 100. One drawback of the dataset is that because the performers are not speaking directly to the camera in the film, it is impossible to see their entire facial expressions when it comes to a particular emotion. It was discovered that the computer was using more than 12GB of RAM to extract audio spectrograms and video frames. Computer crashes resulted from this. Each audio and video file was analyzed separately in batches to retrieve the data.

3.5. PROPOSED HYBRID DEEP LEARNING METHOD

We proposed a hybrid deep learning model comprising CNN+RNNto process audio content and enhanced 3D CNN to process video frames. Our cross-entropy loss, expressed in Eq. 1,trains the model.

$$L_{\text{cross entropy}} = \frac{1}{N}\sum_{n=1}^{N} -\log\left(\frac{exp(x_c^n)}{\sum_j exp(x_j)}\right) \quad (1)$$

N denotes the total amount of data in the dataset, x_c^n the accurate class score of the n-th data point, and, x_j the class score of the jth input data points out of the n-th data. Because the cross entropy loss will only be minimal when the true class's score for a given data point is noticeably higher than the scores of all other classes, minimizing the loss will compel our model to learn the emotion-related features from the audio spectrogram.

As seen in Fig. 3, our hybrid model is a two-stream network made up of two sub-networks, which was inspired by the work of [32]. We have selected to employ the best-performing audio model we have ever built, CNN + RNN since the first sub-network is the audio model. As the CNN+RNN model illustrates, it dumps the original output layer to get high-level properties of audio spectrograms. The second sub-network is the video model, which consists of two fully connected layers, three 3D max-pooling layers, and four 3D convolutional layers (3DCNN). Ultimately, the last layer of both sub-networks is joined together, succeeded by one displaying the output layer. Semi-supervised and supervised training are the two approaches we

use to train this model. We first pre-train our model for the semi-supervised training strategy using video frames and audio spectrograms from the same and distinct videos. As a result, the model is compelled to discover how a video's visual and aural components correlate. Three different kinds of inputs are used in the pre-training process: positive, where the audio spectrogram and video frames are from the same video; complicated negative, where the audio spectrogram and video frames are from different videos with different emotions; and super hard harmful, where the audio spectrogram and video frames are from other videos with the same emotion. Contrastive loss, as expressed in Eq. 2, is the loss function we employ in pre-training.

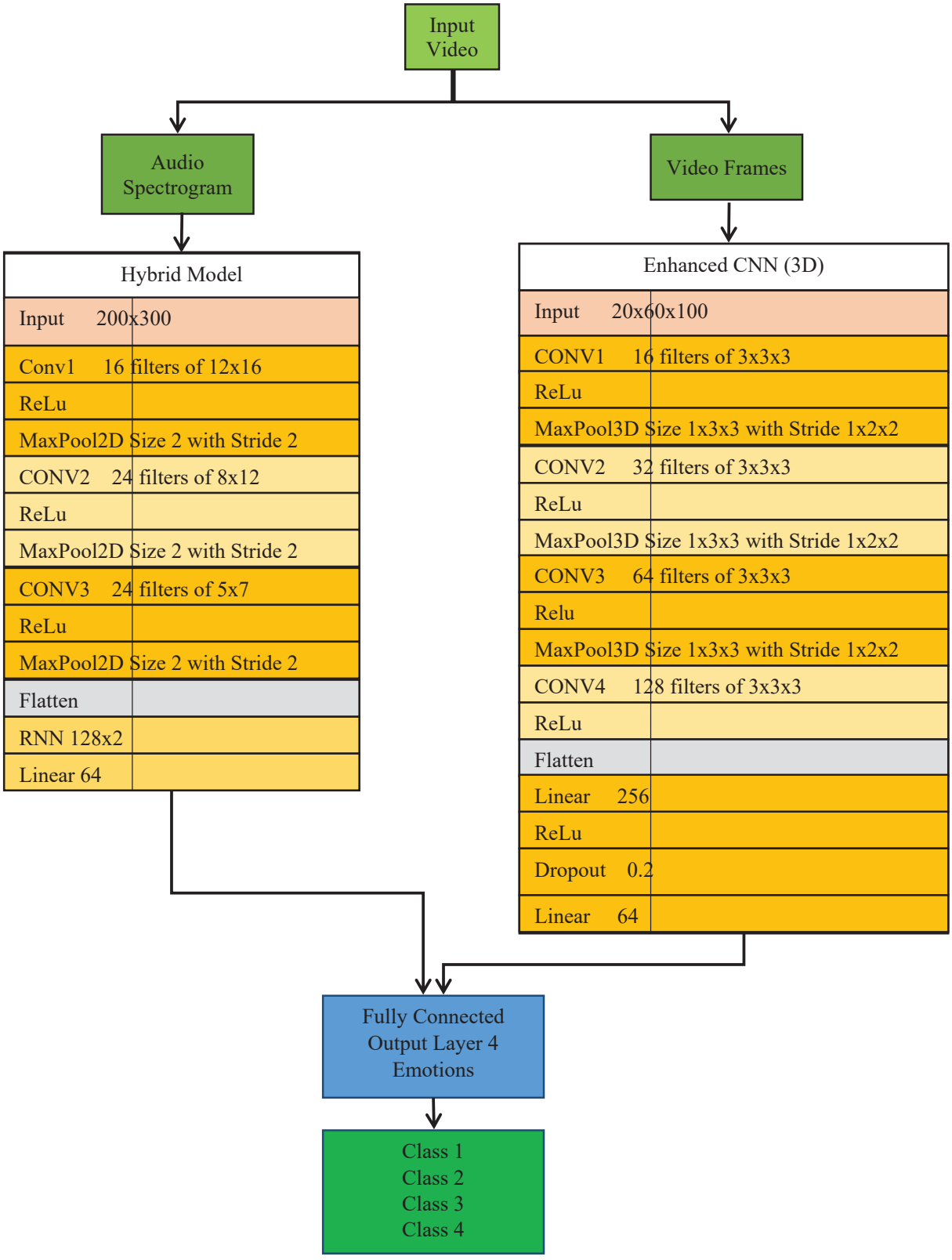

Fig. 3. Proposed hybrid deep learning model for emotion recognition

$$L_{\text{contrastive loss}} = \frac{1}{N}\sum_{n=1}^{N} L_1^n + L_2^n \quad (2)$$

Where

$$L_1^n = (y^n)\|f_v(v^n) - f_a(a^n)\|_2^2$$

$$L_2^n = (1 - y^n)max(\eta - \|f_v(v^n) - f_a(a^n)\|_2, 0)^2$$

The number N denotes the number of data points in the dataset; the variables represent the video and audio sub-networks f_v, f_a; if the video frames and audio spectrogram are from the same movie, y^n is one; if not, it is zero. The margin hyperparameter is denoted by η. When the audio spectrogram and video frames are from the same video, $\|f_v(v^n) - f_a(a^n)\|_2$ should be tiny; otherwise, it should be significant. Thus, the audio and video models are driven to output identical values when their inputs are from the same video and highly different values not reducing the contrastive loss. This enables the model to understand how the same video's audio and visual components relate. Supervised learning is performed on the pre-trained model following pre-training. The output is the anticipated emotion, with the input being an audio spectrogram and video frames from a video.

3.6. PROPOSED ALGORITHM

We proposed an algorithm, Learning Human Emotion Recognition (LbHER), which exploits hybrid deep learning models that could process audio and video frames toward emotion recognition.

Algorithm: Learning based Human Emotion Recognition (LbHER)

Input: IEMOCAP dataset D

Output: Emotion classification results R, performance statistics P

1. Begin
2. D'←Preprocess()
3. ($T1$, $T2$)←SplitData(D')

Training Phase

4. audios←getAudio($T1$)
5. spectrograms←GenerateSpectrogram(audios)
6. videoFrames←getVideoFrames($T1$)
7. Configure hybrid DL model m (as in Fig. 3)
8. Compile m
9. m'←TrainModel(spectrograms, videoFrames)
10. Save m'

Testing Phase

11. audios←getAudio($T2$)
12. spectrograms←GenerateSpectrogram(audios)
13. videoFrames←getVideoFrames($T2$)
14. Load m'
15. R←RecognizeEmotions (spectrograms, videoFrames, m')
16. P←Evaluate(R, ground truth)
17. Display R
18. Display P
19. End

Algorithm 1. Learning-based Human Emotion Recognition (LbHER)

As presented in Algorithm 1, it takes IEMOCAP dataset as input and performs human emotion recognition. The algorithm is designed to process audio-visual data to classify human emotions and provide performance statistics. The algorithm starts with preprocessing the IEMOCAP dataset (D), which is then split into training (T1) and testing (T2) sets. During the training phase, audio data from T1 is extracted and converted into spectrograms, while video frames are also obtained. A hybrid Deep Learning (DL) model (m) is configured, compiled, and trained using spectrograms and video frames. The trained model is then saved for later use. In the testing phase, audio from T2 is converted into spectrograms, and video frames are extracted. The saved model (m') is loaded, and emotions are recognized using the test data and the model. The emotion recognition performance is evaluated against ground truth data, and the results (R) and the performance statistics (P) are displayed. The algorithm follows a structured approach to emotion recognition, leveraging audio and video data to train a hybrid DL model, which is then used to classify emotions in new data. The output includes both the emotion classification results and performance statistics, indicating the effectiveness of the LbHER algorithm in recognizing human emotions from audio-visual cues.

3.7. PERFORMANCE EVALUATION

Since we used a learning-based approach, metrics derived from the confusion matrix, shown in Fig. 4, evaluate our methodology.

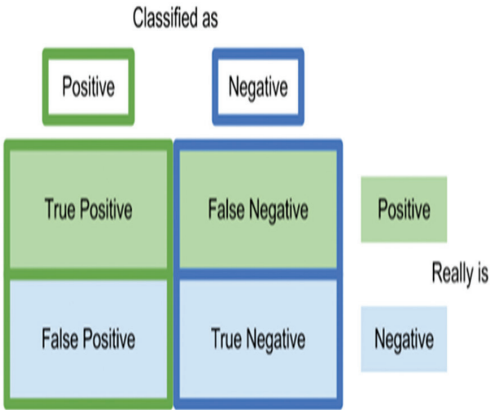

Fig. 4. Confusion matrix

Our method's predicted labels are compared with the ground truth based on the confusion matrix to arrive at performance statistics. Eq. 3 to Eq. 6 express metrics used in the performance evaluation.

$$\text{Precision (p)} = \text{TP}/(\text{TP+FP}) \quad (3)$$

$$\text{Recall (r)} = \text{TP}/(\text{TP+FN}) \quad (4)$$

$$\text{F1-score} = 2*((\text{p}* \text{r}))/((\text{p+r})) \quad (5)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

The measures used for performance evaluation result in a value that lies between 0 and 1. These metrics are widely used in machine learning research.

4. EXPERIMENTAL RESULTS

This section presents the experimental results of the proposed hybrid deep learning model, which automatically recognizes human emotions from a given video. The proposed approach's novelty is that it considers both audio content and video frames from the given input video.

Table 1. Class labels on the corresponding description

Class Label	Description
0	Happy
1	Anger
2	Sad
3	Neutral

The hybrid deep learning model exploits a benchmark dataset named IEMOCAP for getting trained towards automatic detection and classification of emotion. The proposed model performs multi-class classification. The class labels and the description are given in Table 1.

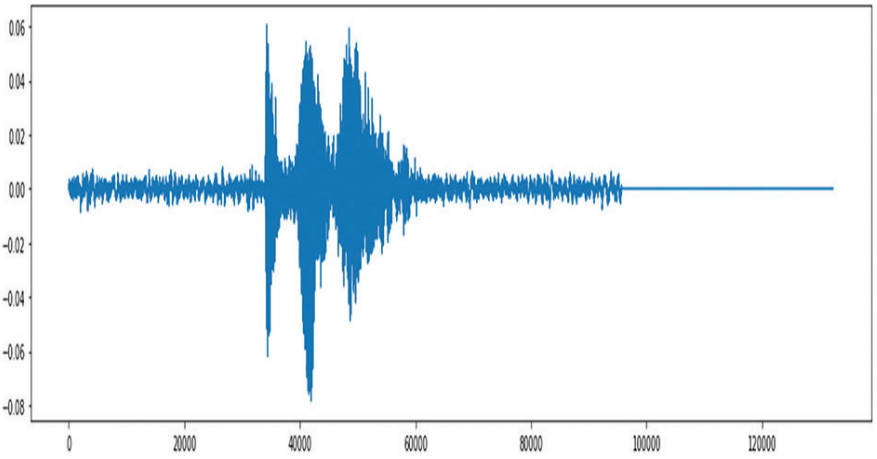

Fig. 5. Original audio content

Fig. 5. presents the content of the original audio file associated with the given test video.

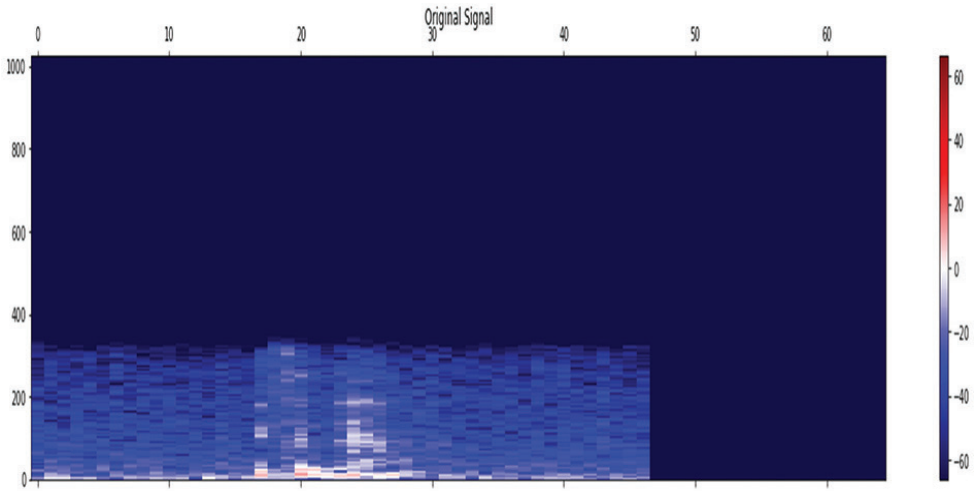

Fig. 6. Shows the spectrogram of the original audio file

As presented in Fig. 6, the given original audio file is converted to a spectrogram because the proposed hybrid deep learning model needs spectrogram input.

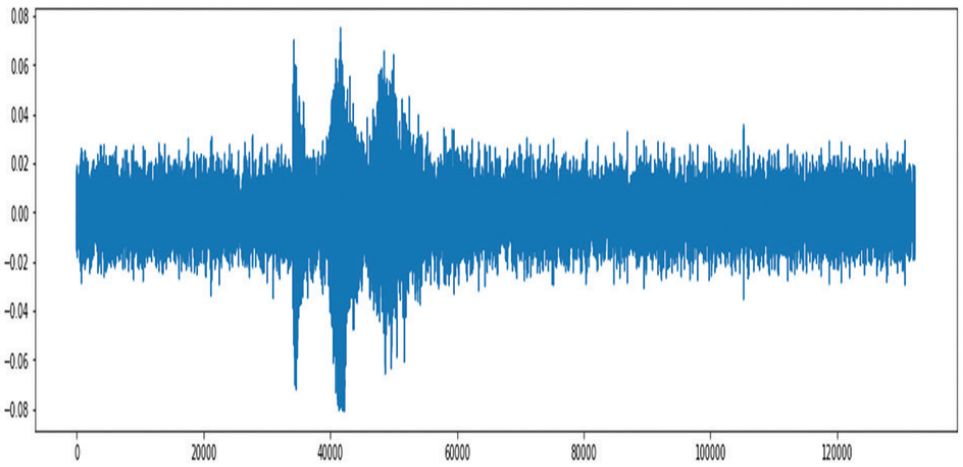

Fig. 7. Shows noise signal

As presented in Fig. 7, the noise signal associated with the given audio content is provided with visualization.

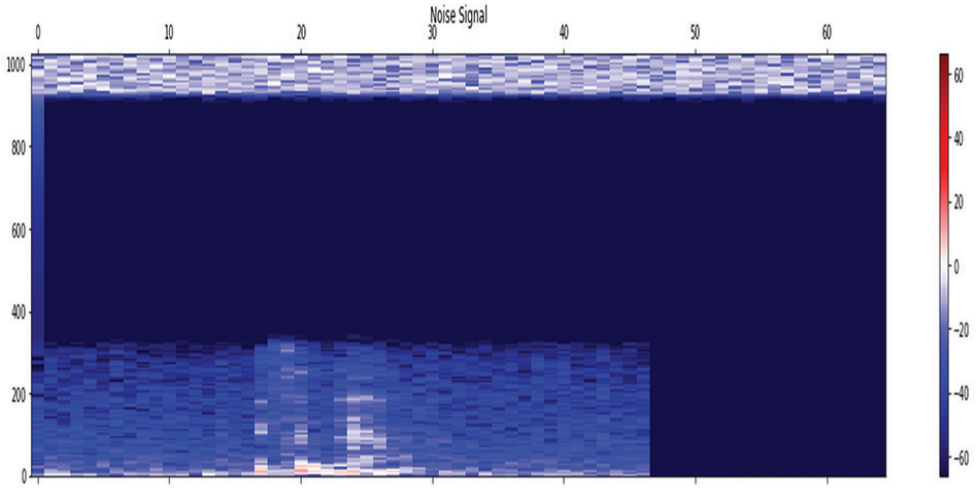

Fig. 8. Spectrogram of the audio with noise

The spectrogram of audio with noise is provided, as visualized in Fig. 8, to understand the data distribution in the form of the spectrogram.

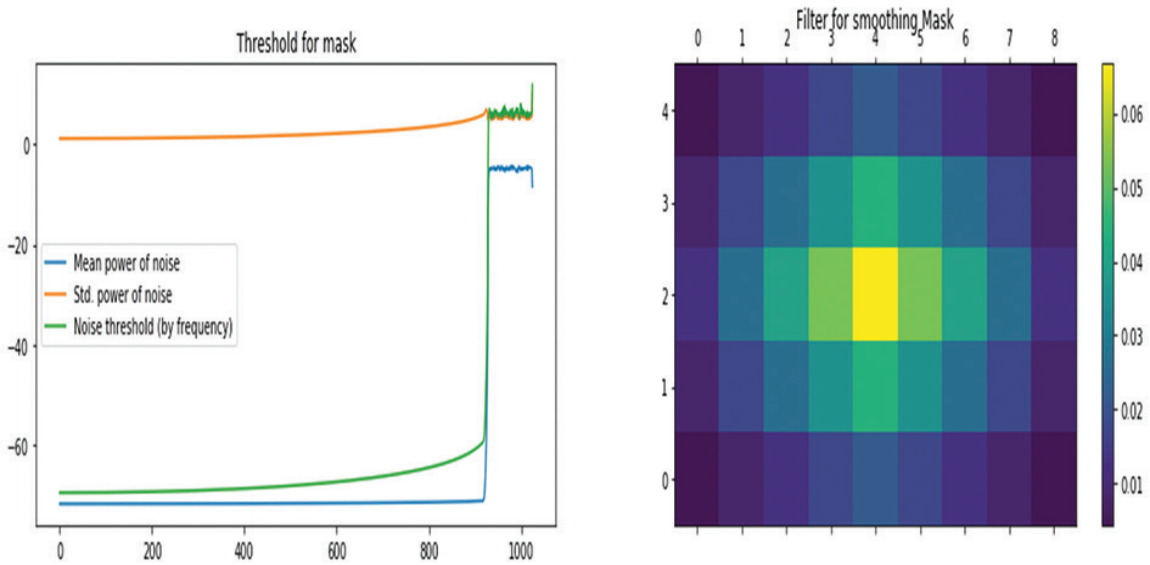

Fig. 9. Illustrates threshold for mask (left) and filter for smoothing mask (right)

As presented in Fig. 9, the threshold for the mask is provided in terms of mean power of noise, standard power of noise, and a noise threshold by frequency besides the filter for smoothing the mask.

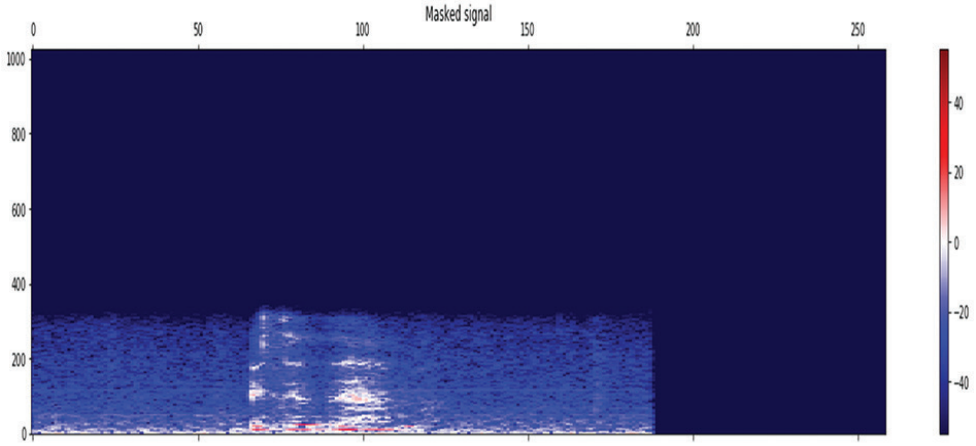

Fig. 10. Shows masked signal in terms of spectrogram

As presented in Fig. 10, the spectrogram of the masked signal is provided with visualization. It is the resultant spectrogram of the masked signal.

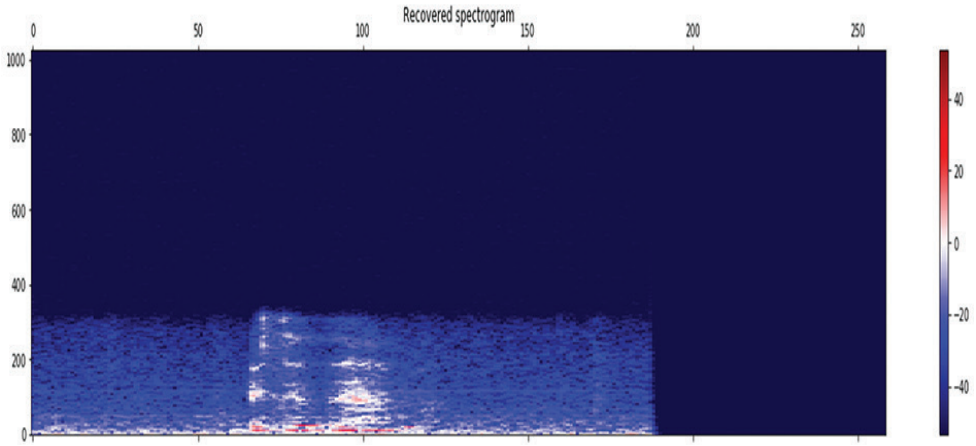

Fig. 11. Shows recovered spectrogram

As presented in Fig. 11, the recovered spectrogram of the masked signal is provided with visualization.

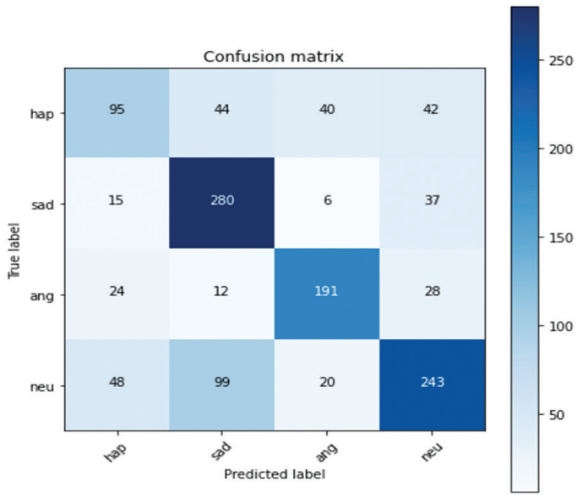

Fig. 12. shows the resultant confusion matrix reflecting the performance of the proposed model.

As presented in Fig. 12, the output of the proposed hybrid model for human emotion recognition is presented as a confusion matrix for different classes. It is the result of multiclass classification from which the model's accuracy is computed.

Table 2. Performance comparison

Model	Accuracy (%)		
	90-10 (Train-Test)	80-20 (Train-Test)	70-30 (Train-Test)
CNN	0.8942	0.8895	0.8833
CNN+LSTM	0.9174	0.9124	0.906
CNN+RNN	0.9341	0.9293	0.9228
CNN+RNN+3DCNN (Proposed)	0.952	0.9472	0.9406

Table 2 shows that the proposed model's performance is compared to many state-of-the-art deep learning models.

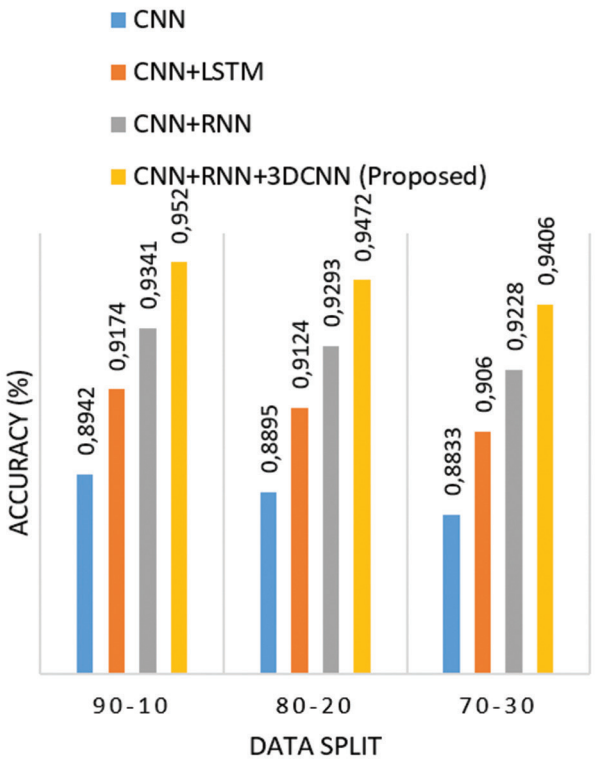

Fig. 13. Shows performance comparison among different models

As presented in Fig. 13, it is understood that different deep learning models showed varied levels of performance in human emotion recognition from a given video. The baseline CNN model, CNN and LSTM hybrid model, and CNN and RNN hybrid model are the existing models compared with the proposed hybrid deep learning model, which comprises CNN, RNN, and enhanced CNN model. CNN model exhibited 88.90% average accuracy, CNN + LSTM 91.19% accuracy, and CNN + RNN hybrid 92.87%, while the proposed hybrid deep learning model exhibited the highest average accuracy with 94.66%. The results show that the proposed hybrid model could perform better than existing and baseline CNN models.

5. DISCUSSION

Human emotion recognition has its utility in many real-world applications. Particularly in healthcare and psychology domains, it is essential to understand emotions to make well-informed decisions. Deep learning models are widely used in computer vision applications to perform various tasks. Since the proposed framework aims to recognize emotions from a given video, deep-learning models are preferred in this paper. However, from the empirical study, it was understood that deep learning models like CNN could help extract features from the input video but lack ability in multi-class classification. To overcome this problem, we proposed a hybrid deep learning model that exploits CNN, RNN, and enhanced CNN in this paper. The framework is designed in such a way that it makes use of both audio content and video frames from the given input. The CNN + RNN combination is used to extract features from audio content, while enhanced CNN is used to extract features from video frames. A fully connected layer is used to perform multi-class classification. The empirical study shows that the proposed hybrid deep learning model could outperform many existing deep learning models in human emotion recognition. However, the proposed methodology has certain limitations, as discussed in Section 5.1.

5.1. LIMITATIONS

In this paper, the proposal framework has certain limitations. The framework comprises a hybrid deep learning model trained with a particular dataset. It is essential to use diversity in data to have generalized findings. This limitation must be overcome with diversified datasets to train the proposed hybrid deep learning model. Another significant limitation is that the proposed hybrid model needs further optimization with improved hyperparameter tuning strategies. The proposed framework can also be enhanced by introducing Generative Adversarial Network (GAN) architecture suitable for human emotion recognition.

6. CONCLUSION AND FUTURE WORK

We proposed a deep learning framework with specific optimizations for facial expression and emotion recognition from videos. We proposed an algorithm, Learning Human Emotion Recognition (LbHER), which exploits hybrid deep learning models that could process audio and video frames toward emotion recognition. Our empirical study with a benchmark dataset, IEMOCAP, has revealed that the proposed framework and the underlying algorithm could leverage state-of-the-art human emotion recognition. Our experimental results showed that the proposed algorithm outperformed many existing models with the highest average accuracy of 94.66%. In the future, we intend to improve our deep learning framework with a Generative Adversarial Network (GAN) architecture. Another direction for future work is to investigate and evaluate our deep learning framework with multiple diversified datasets to generalize the findings.

7. REFERENCES

[1] J. Zhang, Z. Yin, P. Chen, S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review", Information Fusion, Vol. 59, 2020, pp. 103-126.

[2] W. E. Villegas-Ch, J. García-Ortiz, S. Sánchez-Viteri, "Identification of emotions from facial gestures in

a teaching environment with the use of machine learning techniques", IEEE Access, Vol. 11, 2023, pp. 38010-38022.

[3] C. Á. Casado, M. L. Cañellas, M. B. López, "Depression recognition using remote photoplethysmography from facial videos", IEEE Transactions on Affective Computing, Vol. 14, No. 4, 2023, pp. 3305-3316.

[4] A. Hassouneh, A. M. Mutawa, M. Murugappan, "Development of a real-time emotion recognition system using facial expressions and EEG based on machine learning and deep neural network methods", Informatics in Medicine Unlocked, Vol. 20, 2020, pp. 1-9.

[5] A. Pise, H. Vadapalli, I. Sanders, "Facial emotion recognition using temporal relational network: an application to E-learning", Multimedia Tools and Applications, Vol. 81, No. 19, 2022, pp. 26633-26653.

[6] K. Patel, D. Mehta, C. Mistry, R. Gupta, S. Tanwar, N. Kumar, M. Alazab, "Facial sentiment analysis using AI techniques: state-of-the-art, taxonomies, and challenges", IEEE Access, Vol. 8, 2020, pp. 90495-90519.

[7] O. Bazgir, Z. Mohammadi, S. A. H. Habibi, "Emotion recognition with machine learning using EEG signals", Proceedings of the 25th national and 3rd International Iranian Conference on Biomedical Engineering, Qom, Iran, 29-30 November 2018, pp. 1-5.

[8] G. Anbarjafari, J. Guo, Z. Lei, J. Wan, E. Avots, N. Hajarolasvadi, B. Knyazev, A.Kuharenko, "Dominant and Complementary Emotion Recognition From Still Images of Faces", IEEE Access, Vol. 6, pp. 26391–26403.

[9] C. Diamantini, A. Mircoli, D. Potena, E. Storti, "Automatic annotation of corpora for emotion recognition through facial expressions analysis", Proceedings of the 25th International Conference on Pattern Recognition, Milan, Italy, 10-15 January 2021, pp. 5650-5657.

[10] I. M. Revina, W. R. S. Emmanuel, "A survey on human face expression recognition techniques", Journal of King Saud University-Computer and Information Sciences, Vol. 33, No. 6, 2021, pp. 619-628.

[11] Y. Cimtay, E. Ekmekcioglu, S. Caglar-Ozhan, "Cross-subject multimodal emotion recognition based on hybrid fusion", IEEE Access, Vol. 8, 2020, pp. 168865-168878.

[12] A. Kumar, A. Kaur, M. Kumar, "Face detection techniques: a review", Artificial Intelligence Review, Vol. 52, pp. 927-948.

[13] M. Zulfiqar, F. Syed, M. J. Khan, K. Khurshid, "Deep face recognition for biometric authentication", Proceedings of the International Conference on Electrical, Communication, and Computer Engineering, Swat, Pakistan, 24-25 July 2019, pp. 1-6.

[14] A. Topic, M. Russo, "Emotion recognition based on EEG feature maps through deep learning network", Engineering Science and Technology, an International Journal, Vol. 24, No. 6, 2021, pp. 1442-1454.

[15] M. Chen, Y. Hao, " Label-less learning for emotion cognition", IEEE Transactions on Neural Networks and Learning Systems, Vol. 31, No. 7, 2019, pp. 2430-2440.

[16] T. Song, W. Zheng, C. Lu, Y Zong, X. Zhang, Z. Cui, "MPED: A multi-modal physiological emotion database for discrete emotion recognition", IEEE Access, Vol. 7, 2019, pp. 12177-12191.

[17] S. Khan, M. H. Javed, E. Ahmed, S. A. A. Shah, S. U. Ali, "Facial recognition using convolutional neural networks and implementation on smart glasses", Proceedings of the International Conference on Information Science and Communication Technology, Karachi, Pakistan, 9-10 March 2019, pp. 1-6.

[18] J. X. Chen, P. W. Zhang, Z. J. Mao, Y. F. Huang, D. M. Jiang, Y. N. Zhang, "Accurate EEG-based emotion recognition on combined features using deep convolutional neural networks", IEEE Access, Vol. 7, 2019, pp. 44317-44328.

[19] S. Parui, A. K. R. Bajiya, D. Samanta, N. Chakravorty, "Emotion recognition from EEG signal using XGBoost algorithm", Proceedings of the IEEE 16th India Council International Conference, Rajkot, India, 13-15 December 2019, pp. 1-4.

[20] W. L. Zheng, W. Liu, Y. Lu, B. L. Lu, A.Cichocki, "Emotionmeter: A multimodal framework for recogniz-

ing human emotions", IEEE Transactions on Cybernetics, Vol. 49, No. 3, 2018, pp. 1110-1122.

[21] C. Qing, R. Qiao, X. Xu, Y. Cheng, "Interpretable emotion recognition using EEG signals", IEEE Access, Vol. 7, 2019, pp. 94160-94170.

[22] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions", Information Fusion, Vol. 91, 2023, pp. 424-444.

[23] P. Sarkar, A. Etemad, "Self-supervised ECG representation learning for emotion recognition", IEEE Transactions on Affective Computing, Vol. 13, No. 3, 2020, pp. 1541-1554.

[24] D. Ayata, Y. Yaslan, M. E. Kamasak, "Emotion based music recommendation system using wearable physiological sensors", IEEE Transactions on Consumer Electronics, Vol. 64, No. 2, 2018, pp. 196-203.

[25] R. He, J. Cao, L. Song, Z. Sun, T. Tan, "Adversarial cross-spectral face completion for NIR-VIS face recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 42, No. 5, 2019, pp. 1025-1037.

[26] H. Feng, H. M. Golshan, M. H. Mahoor, "A wavelet-based approach to emotion classification using EDA signals", Expert Systems with Applications, Vol. 112, 2018, pp. 77-86.

[27] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection", Information Fusion, Vol. 64, 2020, pp.131-148.

[28] D. Hazarika, S. Poria, R. Zimmermann, R. Mihalcea, "Conversational transfer learning for emotion recognition", Information Fusion, Vol. 65, 2021, pp. 1-12.

[29] T. Baltrušaitis, C. Ahuja, L. P. Morency, "Multimodal machine learning: A survey and taxonomy", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 41, No. 2, 2018, pp. 423-443.

[30] A. Davison, W. Merghani, C. Lansley, C. C. Ng, M. H. Yap, "Objective micro-facial movement detection using facs-based regions and baseline evaluation", Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition, Xi'an, China, 15-19 May 2018, pp. 642-649.

[31] D. Amodei et al. "Deep speech 2: End-to-end speech recognition in english and mandarin", Proceedings of the 33rd International Conference on International Conference on Machine Learning, New York, NY, USA, 19-24 June 2016, pp. 173-182.

[32] A. Torfi, S. M. Iranmanesh, N. Nasrabadi, J. Dawson, "3D convolutional neural networks for cross audio-visual matching recognition", IEEE Access, Vol. 5, 2017, pp. 22081-22091.