International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Emotion Recognition Using Deep Learning Techniques for User Experience

Cheruvu Nagalakshmi Govardhini[1], *Amjuri Venu*[2], *Kotla Lakshminarayana*[3], *Jidagam Shiva Madhav Kumar*[4], *Dr. Ajit Kumar Rout*[5]

[1]*Student, Palakonda, Srikakulam, 532440, India,*
[2]*Student, Rajam, Vizainagaram, 532127, India,*
[3]*Student, Palakonda, Srikakulam, 532440, India,*
[4]*Student, Palakonda, Srikakulam, 53244, India.*
[5]*Professor, GMRIT, Rajam, India.*

DOI: https://doi.org/10.55248/gengpi.4.1023.102819

ABSTRACT

Emotion recognition is critical in comprehending human behavior because it allows us to perceive and respond correctly to people's emotional states. Researchers work hard to study physical traits across several modalities, such as the face, voice, and body motions, in order to deepen our understanding of complicated human behavior. The spike in interest in spontaneous multi-modal emotion identification reflects its ability to collect and understand real-time emotions, emphasizing the importance of combining facial expressions with audio input. In order to address the difficulty of voice emotion recognition, this research uses the RAVDESS dataset and deep learning approaches such as CNNs and LSTM models. The spectrograms were extracted with the librosa library, and LSTM models were used to recognize emotions in speech. This study used point-wise convolution to examine video frames and included techniques from video-based emotion identification. Our research sought to advance real-time emotion recognition, advancing fields including software engineering, internet personalization, education, and gaming. A thorough evaluation of the proposed multimodal emotion detection system can be accomplished by incorporating performance metrics like accuracy, precision, and confusion matrix.

Keywords — Emotion recognition, MFCCs, Convolution Neural Network, Long Short Term Memory.

1. Introduction

Emotion identification is a fluid method focusing on an individual's emotional state, meaning that the emotions associated with each person's acts differ. People regularly communicate their feelings through numerous modes of communication. Recognizing emotions in our daily encounters is critical for social engagement since emotions influence human behaviors and decisions. Individuals employ a variety of strategies to express their emotions, both verbally and nonverbally, including expressive language, facial expressions, bodily movements, and other nonverbal signs. As a result, an individual's emotional state can be anticipated by applying emotional clues from numerous sources. However, a single method is insufficient for effectively analyzing a person's emotions. It is difficult to ascertain someone's emotional state just based on an object or event viewed. This emphasizes the importance of approaching emotional awareness as a multifaceted task.

In recognition of the limitations imposed by unimodal approaches, encompassing factors like temporal dynamics, performance variations, and decreased accuracy, the field has evolved to incorporate multimodal recognition systems. These systems, in acknowledgment of these constraints, leverage a spectrum of emotional cues from diverse sources. The multimodal emotion identification system seamlessly integrates various feature extraction techniques, including Gaussian Models, Natural Language Processing, automata, and Hidden Markov Models. These collective strategies yield a richer understanding of expressive speech, as they synthesize data from multiple channels, ultimately delivering a more comprehensive and precise depiction of an individual's emotional state.

Deep learning has emerged as a successful technique in recent years, accompanied by the rapid growth of modern science, achieving astounding feats across a variety of rules, including signal alter, machine intelligence, and emotion detection. Deep Belief Networks (DBNs), recurrent neural networks (RNNs), and Convolutional Neural Networks (CNNs) have proven to be particularly adept in resolving the complexity of multimodal emotional recognition. These strategies provide a comprehensive perspective of emotional cues by merging input from several modalities, improving the overall accuracy and reliability of emotion identification systems. As technology advances, the incorporation of multimodal recognition inside deep learning frameworks promises a promising approach for improving our ability to effectively perceive and interpret human emotions.

2. Literature Survey

Schoneveld, L., et.al., introduces an Audio-Visual Emotion Recognition (AVER) model where the author takes an audio input from RECOLA dataset and visual input from AffectNet and Google Facial Expression Comparison and AffectNet datasets. This paper proposes a model which involves combining two deep neural networks: (i) a deep CNN model, which has undergone training using knowledge distillation for FER, and (ii) a customized and carefully adjusted and fine-tuned VGGish model for SER. Audio-visual emotion recognition models used in this paper are more robust to noise and occlusion than models that only use audio or visual information. The authors note that the proposed approach is limited by the availability of extensive datasets for pre-processing and training both the audio and visual modules.[1]

Issa, Dias..,et.al., has taken the input which is a combination of five different audio features, and the output is a classification of the emotion present in the speech signal. . In this paper, three datasets are used mainly by the author, those are: SAVEE ,EMO-DB and RAVDESS. The authors proposed a new framework using one-dimensional deep CNN with five different audio features as input data. They used a baseline model, fine-tuned the models, and evaluated the performance using classification accuracy and confusion matrices. The incremental method mentioned in this paper is used to improve classification accuracy may not be applicable for all types of speech emotion recognition tasks.[2]

Gupta.,et.al., In this paper, the authors have taken the speech data as input, and the output is the recognized emotion from the speech data. Here, the RAVDESS dataset is used, which has 24 actors speaking and singing in North-American accent, and comprises 7536 video recordings. The paper proposes a feature extraction technique using image spectrograms of speech-based physiological signals, and employs different classification models including deep neural networks, support vector machines, and decision trees to address the drawbacks of existing approaches, a DSCNN mechanism with somewhat different functions is proposed. The current study only utilized a single corpus for training and evaluation, potentially limiting the model's generalizability to different datasets and populations.[3]

Li, Y., Zhao.,et.al., The paper proposes a model that takes raw audio as input and outputs emotion and gender classification. In this paper, The IEMOCAP dataset was utilized to assess the proposed approach. The proposed method includes a spectrogram-based self-attentional CNN-BLSTM model for emotion classification, multitask learning with gender classification, and and a self-attention mechanism to focus on prominent emotional stages in speech. The key advantage of the suggested method is that it achieves an absolute gain in overall accuracy of 7.7% above the best available findings. The method is only evaluated on the IEMOCAP dataset, so it is not clear how well it would generalize to other datasets.[4]

Sajjad., et.al., The paper proposes a model that takes speech signals as input and outputs emotion labels. In this study, the author proposed a novel SER technique based on RBF to process some interesting segments from an entire audio stream picked using the K-means clustering algorithm. Using a CNN model named Resnet101, the selected speech segments will be transformed to spectrograms and high-level differentiating features retrieved. The main advantage of the proposed approach achieves high accuracy in recognizing emotions from speech signals. Investigating the effect of different normalization techniques on the proposed approach; exploring the use of transfer learning to improve performance.[5]

Nayak, Satyajit, et al. This paper provides a methodology that generates a multivariable time series thermal video series, as well as the ability to discern human emotions and corresponding psychological suggestions. The authors collected their own dataset of thermal video sequences from 20 participants, which they call the "IITKGP Thermal Facial Expression (ITFE) dataset". This paper utilizes deep learning for face registration, employs MIL algorithm for ROI tracking, and applies SVM classifier for emotion recognition. The proposed framework achieves high accuracy in recognizing human emotions from thermal video sequences, even in real-time scenarios. The limited number of participants in the data set limits our ability to draw generalized conclusions about human emotions, and a larger dataset is desirable for future studies.[6]

Zhang, K., et.al., The paper proposes a model that takes thermal video sequence as input and recognition of human emotions and recognized emotion as output. The authors used two datasets for their experiments: the AffectNet Thermal (AT) dataset and the Thermal Facial Expression (TFE) dataset. The paper introduces an innovative Human-Computer Interaction (HCI) framework comprised of four key modules: (1) Preprocessing, (2) Feature Extraction, (3) Emotion Recognition, and (4) post-processing. One notable strength of this study lies in its pioneering HCI framework designed for the recognition of emotions using time-series thermal video sequences. This framework demonstrates cutting-edge performance when evaluated on two widely recognized benchmark datasets. The authors mention that the proposed framework has some limitations, such as the need for a high-quality thermal camera and the limited availability of thermal datasets.[7]

The input to the emotion recognition system is audio-visual data of human faces expressing seven basic emotions, and The study focuses on the identification of various emotions, including anger, disgust, fear, happiness, sadness, surprise, and neutrality, and utilizes datasets sourced from the Cohn-Kanade (CK) database, YouTube videos, and films for analysis of human emotions. The system employs various methods for audio and video preprocessing, including DFT, DCT, feature selection, face detection, feature extraction, clustering, autoencoder, concatenation, normalization, and classification using a feed-forward neural network trained by back-propagation. The main advantage of the system is its real-time recognition of human emotions using both audio and visual data, with high accuracy and adaptability to changes in parameters and emotion classes. The limitation and future work of the system include the limited availability of labeled datasets for emotion recognition, potential bias in dataset selection, and the need for further research on cross-cultural and cross-linguistic emotion recognition.[8]

The input to this paper was videos of facial expressions, and The outcome of the study involved the identification of ten distinct emotions using the Amsterdam Dynamic Facial Expression Set-Bath Intensity Variations dataset, along with contributions from two additional datasets. The methods used in this paper included preprocessing techniques such as face detection and alignment, feature extraction using VGG16 and Inception-v3 deep convolutional neural networks, and classification process employed Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) techniques. The main advantage of this paper was achieving high accuracy in recognizing emotions from videos using deep convolutional neural networks. The main limitation/future work of this paper was the small dataset size, which required the authors to consider two procedures for training and testing.[9]

Do, L. N., Yang, H. J., Nguyen.,et.al., The authors used two public datasets, AFEW 2016 and SAVEE, to evaluate the performance of the proposed model. The proposed model employs some of the techniques like CNN-based architectures, VGG16, InceptionV3, and ResNet50, and additional modules, BLSTM, TCN, and TAM, to obtain high-level features from the facial image sequence and effectively capture sequential patterns within the entire video dataset, this research focuses on the development and training of deep neural networks. A key strength of this investigation lies in the creation of robust neural networks tailored for emotion recognition tasks involving visual data, applicable to both unconstrained and constrained settings. The authors suggest that the proposed model can be further improved by incorporating additional modalities such as audio and text and can be extended to other types of visual data, such as images or live video streams.[10]

Luna-Jiménez.,et.al., This paper presents a solution for recognizing emotions from multimodal information using transfer learning. The system takes as its input both speech signals and facial expressions, and the output is emotion recognition. The system employs a late fusion approach to combine two separate models. The speech model undergoes pre-training using the VoxCeleb dataset and subsequent fine-tuning using the RAVDESS dataset, and the facial expression model pre-trained on the Aff-Wild2 dataset and fine-tuned on RAVDESS. The system incorporates a subject-wise cross-validation approach to ensure a robust evaluation process. Notably, a key benefit of this system is its attainment of top-tier performance levels on the RAVDESS dataset. Future work includes exploring the use of additional modalities and improving the robustness of the system.[11]

Denis Dresvyanskiy.,et.al., The author proposes a system that is designed as a comprehensive multimodal emotion recognition solution, capable of processing both audio and video inputs to generate predictions for emotional labels. The system was tested on the AffWild2 dataset, which is a massive dataset of audiovisual recordings of people expressing emotions in natural settings. The author proposed a system that consisting of several modules, including a feature extraction module, a deep audiovisual fusion module, and a transfer learning module. The feature extraction module uses pre-trained models to extract audio and visual features. The deep audiovisual fusion module combines these features using a deep neural network. The transfer learning module fine-tunes the fusion model on the target dataset. The system also showcases the successful application of transfer learning in the realm of multimodal emotion recognition. The paper mentions that the system could be improved by incorporating additional modalities, such as text or physiological signals.[12]

Arselan Ashraf., et.al., ., The proposed model presented in this paper for multimodal emotion recognition takes both speech and video as inputs and delivers the corresponding recognized emotional states. To validate their approach, the authors employed two distinct datasets: the Toronto Emotional Vids and Audio (TESS) dataset and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Various techniques have been implemented for audio input, including division and overlapping, multiplication of frames, Fourier transforms, and more. Similarly, for video input, the process involves tasks such as face detection, histogram equalization, image cropping, resizing, frame selection, and extraction. The model employs a 2D-CNN for audio data and a 3D-CNN for video data, in addition to Extreme Learning Machine (ELM), Softmax layers, and SVM for the final classification. To enhance feature extraction, the proposed model has been trained using an extensive dataset with the aid of data augmentations, with a notable focus on Convolutional Neural Networks (CNNs). The model's accuracy can be further improved by using more advanced techniques for feature extraction and classification, and its performance can be evaluated on more diverse datasets to test its generalizability.[13]

Singh., et.al., This paper presents a method based on deep neural networks for identifying emotions in audio and video. The proposed multi-modal architecture combines audio and video features to improve accuracy in predicting emotion labels. The audio processing module uses a CNN and LSTM network to extract and capture temporal dependencies in audio spectrograms, while the video processing module uses a similar approach for video frames. The fusion module combines the audio and video features using a weighted sum and a fully connected layer. The paper uses the IEMOCAP corpus as the dataset and achieves promising results. One of the main limitations of this paper is the limited diversity of the dataset used. The IEMOCAP corpus used in the study consists of conversations between actors, which might not accurately reflect situations in the actual world. This might make the proposed approach less applicable in other scenarios.[14]

Hossain.,et.al., In this study, the author proposes a system for recognizing emotions has been developed, which accepts both speech and video signals as input and provides the corresponding emotion recognition as output. Big Data and non-Big Data databases were employed in the two databases.. The Big Data database contains 10,000 audio-visual clips, while the other database contains 1,040 audio-visual clips. The system uses preprocessing, deep networks using CNN, two-stage ELM based fusion, and an advanced methodology is employed for the selection of pivotal frames, encompassing various informative patterns from these key frames for subsequent feature extraction. The integration of an Extreme Learning Machine (ELM) enhances the accuracy of the system. The proposed system has undergone rigorous training using a vast dataset of emotional content, resulting in the effective training of deep neural networks. However, the obtained accuracies are still below expectation, and further research is needed to improve the system's performance. Future work could focus on exploring other fusion strategies and improving the accuracy of the system.[15]

3. Data Collection

Unimodal systems are ill-equipped to cope with the substantial surge in data originating from diverse sources in real-time. Consequently, a multi-modal platform for recognizing human emotions has been devised to effectively manage three distinct types of data. So, Our proposed system takes three types of inputs namely, text, speech and video in real-time as well as in online. As it is multi-modal emotion recognition system, three modules are integrated.

4. Deep Learning

4.1. CNN:

Modern deep learning, which is based on Convolutional Neural Networks (CNNs), has completely changed several industries, most notably image analysis and computer vision. These neural networks are made to analyze and comprehend visual information with astounding precision and efficiency. CNNs use convolutional layers, which apply filters or kernels to input data, to extract hierarchical characteristics from pictures. Intricate features and representations may be learned by CNNs straight from raw pixel data thanks to these filters' ability to recognize patterns like edges, textures, and complex visual structures. Pooling layers are frequently used in CNN designs to minimize spatial dimensions and avoid overfitting, and fully linked layers are then used for classification tasks.

CNN's general workflow begins with an input image and creates a feature map by applying a variety of filters to it.

- Increases non-linearity by using a ReLU function.
- Adds a layer of pooling to each feature map.
- It converts the gathered photos into a single, lengthy vector.
- It feeds the vector into an artificial neural network that is fully connected.

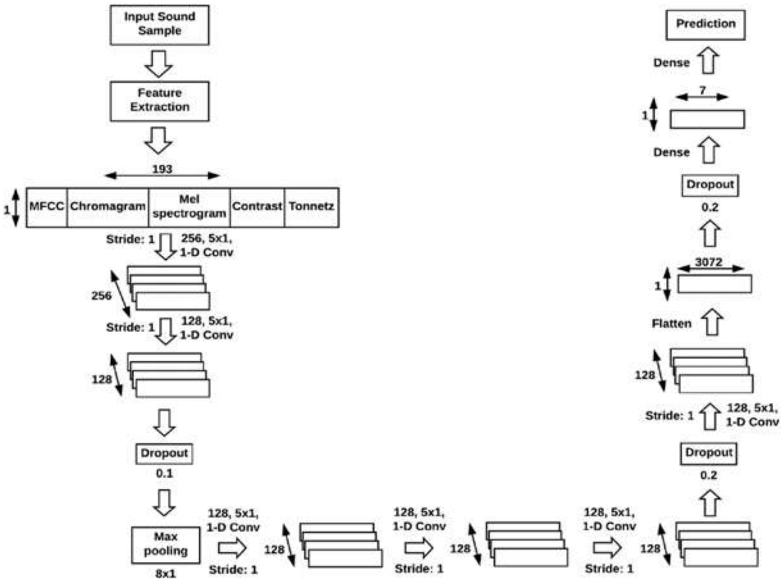


Figure 4.1: Workflow of CNN in Emotion recognition

5. Methodology

Unimodal systems are ill-equipped to cope with the substantial surge in data originating from diverse sources in real-time. Consequently, a multi-modal platform for recognizing human emotions has been devised to effectively manage three distinct types of data. So, Our proposed system takes three types of inputs namely, text, speech and video in real-time as well as in online. As it is multi-modal emotion recognition system, three modules are integrated those are,

A. Textual Emotion Recognition

B. Audio Emotion Recognition

C. Video Emotion Recognition

A. TEXTUAL EMOTION RECOGNITION:

Here, we are training the model with help of Stream-of-consciousness dataset. This module is mainly concentrating on 1D-CNN, LSTM model for extracting features and generalizing emotions. The main procedure involves the following steps:

1. Taking input to the model which consists of a text sequence, such as a sentence or a document, is the primary focus.
2. The text is then pre-processed to eliminate stop words, punctuation, and other extraneous information. The text is then tokenized into individual words.
3. Now, the words are converted into numerical vectors using an embedding layer. This allows the model to represent the meaning of the words in a high-dimensional space. Here, the vectors are given to 1D-CNN model which gives vector of features as output.
4. This vector of features are then given to LSTM model to capture the global dependencies in the input text.
5. The model generates an output in the form of a feature vector, which can be utilized for the classification of text into various emotional categories.

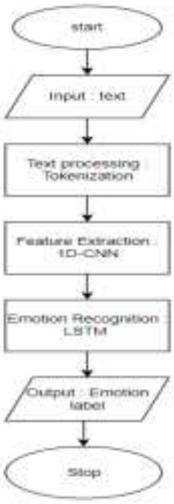

start
Input : text
Text processing : Tokenization
Feature Extraction : 1D-CNN
Emotion Recognition : LSTM
Output : Emotion label
Stop

Fig 5.1 : Process of Textual Emotion Recognition

The models including in this textual emotion recognition are:

a) 1D-CNN:

1D-CNN stands for one-dimensional convolutional neural network. It is a particular class of neural network that is frequently employed for handling sequential input. Here, 1D-CNN model is applied to the embedded text for the purpose of obtaining features important for emotion recognition. The convolution operation is applied to the embedded text to extract features, and the utilization of an activation function serves the purpose of introducing non-linear characteristics into the model. The pooling operation is then applied to The purpose of reducing the dimensionality of the feature maps is to create a 1D-CNN model that yields a feature vector. This vector captures the local dependencies present within the input text, as outlined in a student paper submitted to the University of Warwick.

b) LSTM:

LSTM, an acronym for Long Short-Term Memory, belongs to the category of recurrent neural networks (RNNs). Its design specifically addresses the vanishing gradient issue often encountered in conventional RNNs. Here, The LSTM model is applied to the vector of features produced by the 1D-CNN model which is a way to visualize temporal dependencies between the words. The LSTM model sequentially processes the input sequence, considering each word individually, and continually updates its internal state based on the current input as well as its previous state. The output of the LSTM model is a vector of features that captures the global dependencies in the input text.

B. AUDIO EMOTION RECOGNITION:

We are using Ryerson Audio-Visual Database of Emotional Speech and Song for audio data sets (RAVDESS) dataset in order to train the LSTM model to extract features and detect emotion The main procedure involves the following steps:

1. The system takes speech as input.
2. That means, The system records the audio signal and Apply pre-processing techniques to remove noise and artifacts.
3. Now, Mel-Frequency Cepstral Coefficients (MFCCs) are extracted from the audio signal by dividing frames of a certain size for the audio signal
4. Finally, the features are given to the LSTM model in order to recognize emotional label from it.

a) LSTM:

LSTM, short for Long-Short-Term Memory, is a type of recurrent neural network (RNN) used to model sequential data. LSTM aims to overcome the vanishing gradient problem encountered in conventional RNNs. LSTM is used for feature extraction and classification of emotions in speech emotion recognition. . The initial step involves converting the input audio into discrete frames, followed by the extraction of frame-specific features using the LSTM model. The extracted features are then used to classify the emotion of the input speech. LSTM is particularly useful in speech emotion recognition because it possesses the capability to capture extended-term dependencies within the speech signal, which plays a vital role in achieving precise emotion recognition. For example, the pitch and tone of a person's voice may change gradually over time, and LSTM can capture these changes and use them to classify the emotion of the input speech.

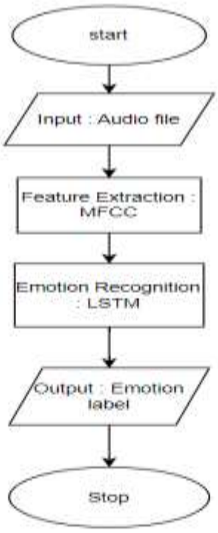


Fig 5.2: Process of Speech Emotion Recognition

C. VIDEO EMOTION RECOGNITION:

Regarding Video emotion recognition, we utilize the FER2013 Kaggle Challenge dataset, which comprises grayscale facial images with a resolution of 48x48 pixels. This aspect of emotion recognition primarily involves employing point-wise convolution and depth-wise convolution techniques for feature extraction. For emotion classification we are using the combination of SVM, LSTM, CNN. The procedure of Video emotion recognition involves the following steps:

1. The first step is to identify the facial landmarks of the individual in the video,in which Dlib library is employed By using this way, the most important aspects such as the eyes, nose and mouth are identified.

2. Once the facial landmarks have been detected, the following step is to take characteristics out of the video frames using point-wise convolution along with depth-wise convolution.

3. After the features have been extracted, they are employed to categorize the video's subject's emotional state. The suggested system uses a combination of Support Vector Machines (SVM), Long-Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) for emotion classification.

a) Point wise convolution:

Point-wise convolution is utilized to draw features from the source image by utilizing and applying a 1x1 convolution filter. This helps to decrease the input's dimensionality and capture significant features. In this system, point-wise convolution is used in conjunction with depth-wise convolution to analyze facial expressions in video.

b) Depth wise convolution:

Depth-wise convolution is used to apply a filter to the input image and extract spatial information from it and provides it to each channel of the input separately. This helps to capture the spatial relationships between different parts of the face and extract more detailed features. In the proposed system, depth-wise convolution is employed to examine the person's facial gestures in videos.

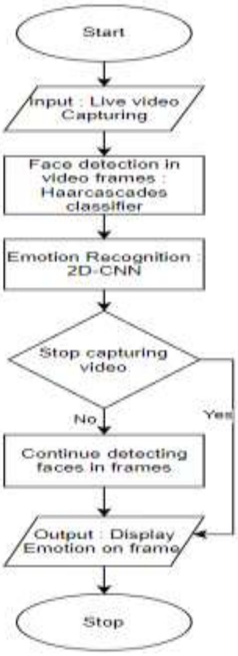

Start
Input : Live video Capturing
Face detection in video frames : Haarcascades classifier
Emotion Recognition : 2D-CNN
Stop capturing video
No
Yes
Continue detecting faces in frames
Output : Display Emotion on frame
Stop

Fig 5.3: Process of Video Emotion Recognition

6. Results and Discussion

A. Text emotion recognition:

Here, we have developed a database which consisting of 10 emotions by giving sentences along with the emotional labels to train our model and it shows the following confusion matrix. As a critical milestone in our research, we have evaluated the performance of our emotion recognition model and generated a confusion matrix to gauge its accuracy. This matrix serves as a vital tool in assessing the model's effectiveness in correctly classifying emotions. It provides a detailed breakdown of the model's predictions, revealing instances where it correctly identified emotions and instances where it may have faltered. By scrutinizing this matrix, we gain valuable insights into areas where our model excels and where it may require further refinement.

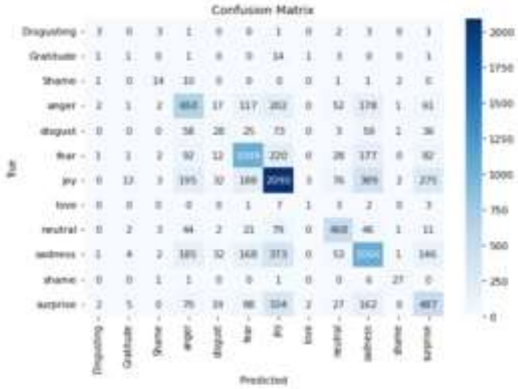


Fig 6.1: Confusion matrix of textual emotion recognition

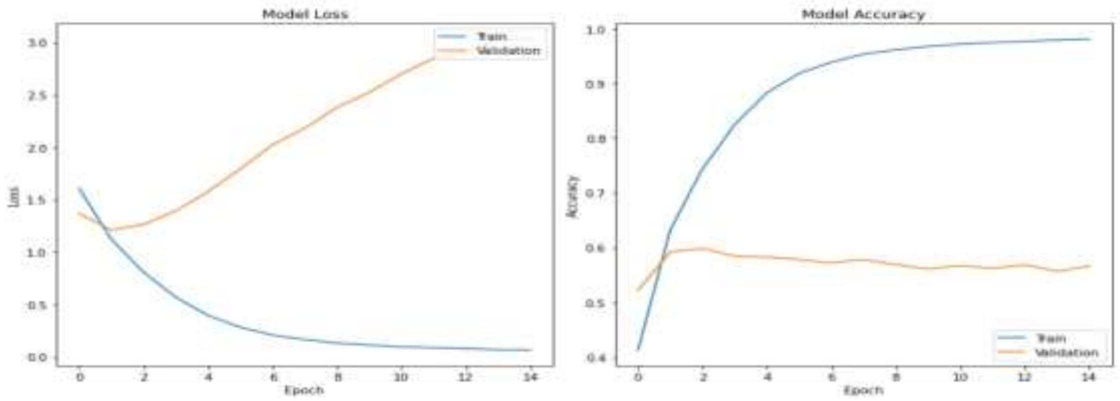
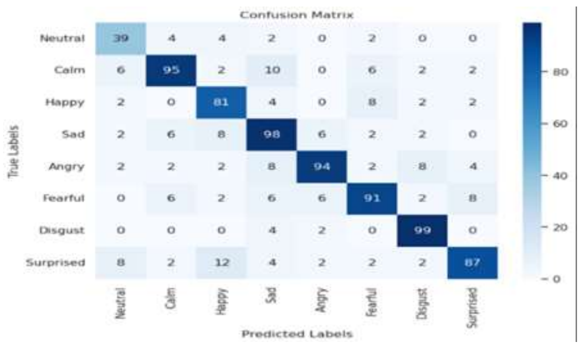


Fig 6.2: Accuracy and Loss curves for text emotion recognition

B. Audio Emotion Recognition:

This model which is used to recognize the emotion from audio is built using a CNN architecture in Keras, demonstrated impressive results with good accuracy. The provided parameters were fine-tuned to achieve optimal performance. Classification of emotions by the model with high accuracy from audio signals holds great promise for applications in fields like speech analysis, virtual assistants, and emotion-aware technology. The specific metrics and detailed results can be found in the provided summary, showcasing the model's effectiveness in capturing emotional cues from audio data. This achievement opens up avenues for enhancing human-machine interaction and understanding emotional nuances in various audio-based contexts.



	Neutral	Calm	Happy	Sad	Angry	Fearful	Disgust	Surprised
Neutral	39	4	4	2	0	2	0	0
Calm	6	95	2	10	0	6	2	2
Happy	2	0	81	4	0	8	2	2
Sad	2	6	8	98	6	2	2	0
Angry	2	2	2	8	94	2	8	4
Fearful	0	6	2	6	6	91	2	8
Disgust	0	0	0	4	2	0	99	0
Surprised	8	2	12	4	2	2	2	87

Fig 6.3: Confusion matrix

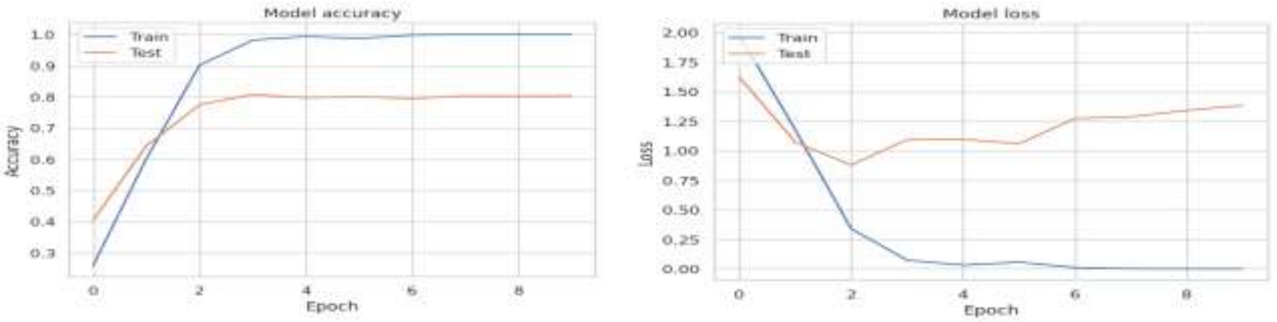


Fig 6.4: Accuracy and Loss vs Epochs

C. Video Emotion Recognition:

The evaluation of the trained model on a dataset of 28059 images for real-time emotion classification yielded insightful results. An extensive breakdown of the model's performance is given by the confusion matrix across different emotion classes, highlighting its ability to accurately classify emotions. Additionally, the validation curve likely demonstrates the convergence and generalization capabilities of the model, offering insights into its learning behavior over training epochs. These results collectively underscore the model's efficacy in real-time emotion classification tasks, preparing the way for its possible use in many fields, including human-machine interaction, sentiment analysis, and personalized user experiences.

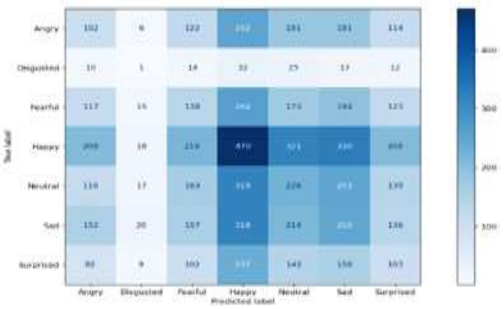


Fig 6.5: Confusion Matrix

8. Conclusion

In conclusion, this work sheds light on the expansive landscape of multi-modal emotion recognition, skillfully employing sophisticated deep neural network approaches such as 1D CNN, 2D CNN, RNN, and LSTM to transcend unimodal limitations and adeptly handle the dynamic influx of real-time data. The resultant platform emerges as a standout performer, particularly evident in its superior performance during virtual interviews when compared to conventional unimodal counterparts. Beyond its immediate applications, the potential ramifications span diverse domains. From employing point-wise and depth-wise convolutions in video analysis to utilizing varied neural networks in processing textual, speech and video data, this study beckons further research and ethical contemplation. Ultimately, multi-modal emotion recognition stands poised as a transformational tool, seamlessly fusing technology and nuanced human emotions across multifarious sectors.

REFERENCES

1. Schoneveld, L., Othmani, A., & Abdelkawy, H. (2021). "Leveraging recent advances in deep learning for audio-visual emotion recognition"(ELSEVIER) Pattern Recognition Letters, 146, 1-7.

2. Issa, Dias, M. Fatih Demirci, and Adnan Yazici. "Speech emotion recognition with deep convolutional neural networks." Biomedical Signal Processing and Control(ELSEVIER) 59 (2020): 101894.

3. Gupta, V., Juyal, S., Singh, G. P., Killa, C., & Gupta, N. (2020). Emotion recognition of audio/speech data using deep learning approaches. Journal of Information and Optimization Sciences, 41(6), 1309-1317..

4. Li, Y., Zhao, T., & Kawahara, T. (2019, September). Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning. In Interspeech (pp. 2803-2807)

5. Sajjad, Muhammad, and Soonil Kwon. "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM." IEEE access 8 (2020): 79861-79875

6. Nayak, Satyajit, et al. "A Human–Computer Interaction framework for emotion recognition through time-series thermal video sequences." Computers & Electrical Engineering (ELSEVIER) 93 (2021): 107280.

7. Zhang, K., Li, Y., Wang, J., Cambria, E., & Li, X. (2021). Real-time video emotion recognition based on reinforcement learning and domain knowledge. IEEE Transactions on Circuits and Systems for Video Technology, 32(3), 1034-1047.

8. Moskvin, A. A., and A. G. Shishkin. "Deep Learning Based Human Emotional State Recognition in a Video." Journal of Modeling and Optimization 12.1 (2020): 51-59.

9. Abdulsalam, Wisal Hashim, Rafah Shihab Alhamdani, and Mohammed Najm Abdullah. "Facial emotion recognition from videos using deep convolutional neural networks." Int. J. Mach. Learn. Comput 9.1 (2019): 14-19.

10. Do, L. N., Yang, H. J., Nguyen, H. D., Kim, S. H., Lee, G. S., & Na, I. S. (2021). Deep neural network-based fusion model for emotion recognition using visual data. The Journal of Supercomputing, 1-18

11. Luna-Jiménez, C., Griol, D., Callejas, Z., Kleinlein, R., Montero, J. M., & Fernández-Martínez, F. (2021).(MDPI) Multimodal emotion recognition on RAVDESS dataset using transfer learning. Sensors, 21(22), 7665.

12. Denis Dresvyanskiy, Elena Ryumina, Heysem Kaya, Maxim Markitantov, Alexey Karpov, Wolfgang Minker. "End-to-End Modeling and Transfer Learning for Audiovisual Emotion Recognition in-the-Wild." Multimodal Technologies and Interaction 2022, 6, 11. https://doi.org/10.33 (MDPI) 90/mti6020011 (2022).

13. Arselan Ashraf, Teddy Surya Gunawan, Fatchul Arifin, Mira Kartiwi, Ali Sophian, Mohamed Hadi Habaebi. "On the Audio-Visual Emotion Recognition using Convolutional Neural Networks and Extreme Learning Machine." Indonesian Journal of Electrical Engineering and Informatics (IJEEI) 10.3 (2022).

14. Singh, Mandeep, and Yuan Fang. "Emotion recognition in audio and video using deep neural networks." arXiv preprint arXiv:2006.08129 (2020).

15. Hossain, M. S., & Muhammad, G. (2019). Emotion recognition using deep learning approach from audio–visual emotional big data. Information Fusion, 49, 69-78.