Speech Emotion Recognition Using Deep Learning

Dr.G.Prathibha, M.Tech, Ph.D

Y. Kavya, L. Poojita, P.Vinay Jacob
Dept. of ECE, Dr YSR ANUCET,
Acharya Nagarjuna University, Guntur, A.P

ABSTRACT:

Speech is one of the primary forms of expression and is important for Emotion Recognition. Emotion Recognition is helpful to derive various useful insights about the thoughts of a person. Automatic speech emotion recognition is an active field of study in Artificial intelligence and Machine learning, which aims to generate machines that communicate with people via speech. In this work, deep learning algorithms such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) are explored to extract features and classify emotions such as calm, happy, fearful, disgust, angry, neutral, surprised and sad using the Toronto emotional speech set (TESS) dataset which consists of 2800 files. The features like Mel-frequency cepstral coefficients(MFCC), chroma and mel spectrogram are extracted from speech using the pre-trained networks such as Xception, VGG16, Resnet50, MobileNetV2, DenseNet121, NASNetLarge, EfficientNetB5, EfficientNetV2M, InceptionV3, ConvNeXtTiny, EfficientNetV2B2, EfficientNetB6, ResNet152V2. Features of the two different networks are fused using the fusion techniques such as Early, Mid, Late to get better optimum results. Features are then classified initially with the Long Short Term Memory (LSTM) finally resulted in the accuracy of 99%. In this paper the work is extended to RAVDESS dataset also which consists of seven emotions such as calm, joyful, sad, surprised, afraid, disgust and angry in total of 1440 files.

Keywords: Convolution Neural Network, Recurrent Neural Network, speech emotion recognition, MFCC, Chroma, Mel, LSTM.

I. INRODUCTION

Speech is a very good way to get emotional information and it is quick, efficient and necessary mode of human communication. Emotion plays a significant role in daily interpersonal human interactions. It helps us to match and understand the feelings of others by conveying our feelings and giving feedback to others.

In this rapidly advancing AI world, human computer interactions are more important. In the present world, Siri and Alexa are physically closer than other humans .So, Speech Emotion Recognition (SER) is introduced for human–computer interaction. The goal of Speech Emotion recognition is to predict emotional content of speech and to classify speech according to one of several labels (i.e. calm, happy, fearful, disgust, angry, neutral, surprised and sad). It is used in many applications such as Marketing, Healthcare, Stress monitoring, E-learning etc. For successful SER, three key issues need to be taken into considerations which are choice of a good emotional speech database, extracting effective features and designing reliable classifiers using deep learning algorithms.

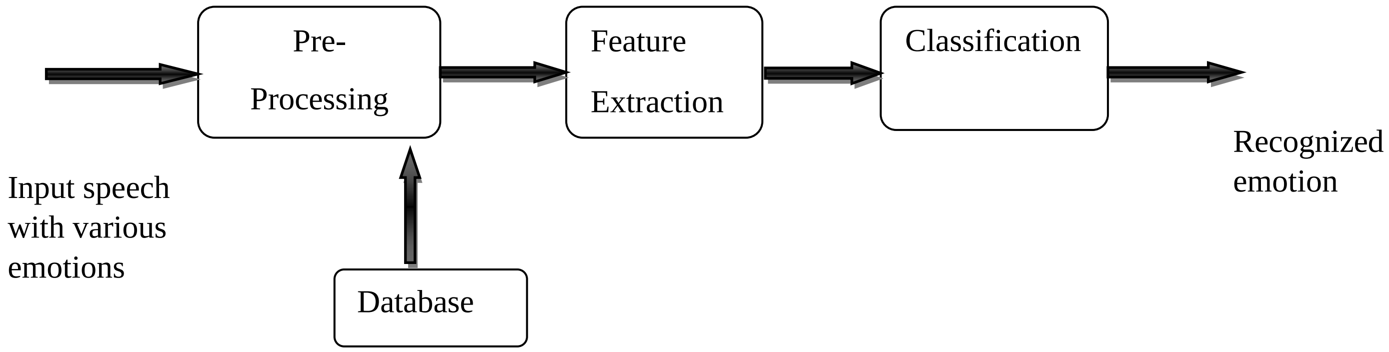


Fig.1 A typical speech recognition algorithm

As shown in Fig.1, Feature extraction is the most crucial step in speech signal processing. For predicting emotions from real time audio, text need to be extracted from audio. For classifying emotional state, number of classifier schemes have been used for the SER, such as Multilayer Perceptron (MLP), Long Short Term Memory (LSTM) classification algorithm etc. SER's deep learning techniques like Convolutional Neural Network and Recurrent Neural Network provide advantages over previous methods like machine learning, including the capacity to recognize complicated structures and features without human feature extraction and tuning. Advantages include the capacity to extract low-level characteristics from raw data and manage unlabeled data. For emotion recognition task, TESS dataset[1] is used.

In this present work, features are extracted from the audio and classified using two different classifiers. Section II provides the information of literatures on existing techniques for speech emotion recognition. Section III deals with methodology used for extraction, classification and recognition. Section IV discusses about the proposed techniques. Section V focuses on experimental results and finally Section V includes conclusion and references used in this work.

II. LITERATURE REVIEW

Seunghyun yoon.et.al.,[2] have proposed a novel deep dual recurrent model for understanding of speech data. It is used to predict the emotion class. The proposed model is performed on four emotions happy, sad, neutral, angry) when applied on Interactive Emotional Dyadic Motion Capture(IEMOCAP)dataset .Finally, they obtained accuracy of 71.8% by applying the above dataset. However, in the present work only four emotions are considered.

Kun-Yi Huang.et.al.,[3]have suggested a convolutional neural network(CNN) model with audio-based embedding for emotion modeling. They utilized NC-KU database containing seven emotion categories (happiness, sad, neutral, surprise, disgust, boredom, anger) are considered and five-fold cross validation was used evaluate the performance of the proposed CNN-based method for speech emotion recognition. They concluded that emotion recognition accuracy of 82.34%, which is improved by 8.7% compared to the Long short-term memory (LSTM)-based method.

S. Lalitha.et.al., [4] have focussed on investigation of the effective performance of perceptual based speech features on emotion detection. They used preceptron features like MFCC'S,PLPC, MFPLPC, BFCC, RPLP and IMFCC for emotion detection. The algorithm using these auditory cues are evaluated with Deep Neural Networks (DNN). They analyzed the algorithm on publicly available Berlin database and obtained 80% of accuracy. Their future aim is to work on multi-corpus acted and natural database with perceptual speech features.

Leila Kerkeni.et.al., [5] have suggested Machine learning technique. They extracted Mel-frequency cepstrum coefficients (MFCC) and Modulation spectral (MS) features from the speech signals and used to train different classifiers. They utilized Recurrent neural network (RNN) classifier to classify seven emotions. In this work, Berlin and Spanish databases are used. Finally, they achieved 83% accuracy on Berlin database and 94% accuracy on Spanish database by using RNN classifier. However, the accuracy can be improved using Pre-Trained Networks of CNN.

A. Christy.et.al., [6] proposed a convolution neural network model. Here the acoustic speech signal are split into short frames, fast Fourier transform is applied and relevant features are extracted using MFCC and MS. In this paper they used algorithms like linear regression, decision tree, Support Vector Machine (SVM) and convolutional neural networks(CNN) for classification. Here human emotions like neutral, calm, happy, sad, fearful, disgust and surprise are classified using above classifiers by testing their model with RAVDESS dataset using CNN resulting in an accuracy of 78.20%.

Mehmet Bilal Er [7] has suggested a novel approach for classification of Speech Emotion based on deep acoustic features. Here acoustic features such as RMS, MFCC and Zero-crossing Rate are obtained from voice records. They classified emotions using RAVDESS, Berlin(EMO-DB) and IEMOCAP datasets. Finally, they achieved accuracy of 79.41%, 90.21% and 85.37% for RAVDESS, EMODB, and IEMOCAP datasets respectively. In future their aim is to develop new techniques for the determination of optimum feature sets in order to improve the accuracy rates of classification.

Krishna Chauhan.et.al.,[8] have proposed CNN architecture, based on log-mel spectrograms of segmented speech utterances. They extracted emotion related features using Interactive Emotional Dyadic Motion Capture (IEMOCAP) and Berlin database of emotional speech (EMODB) datasets. For Speaker-independent analysis they obtained the accuracy of 65.47% for IEMOCAP and 72.02% for EMODB.

Asiya UA.et.al.,[9] have worked on acoustic data such as mel frequency cepstral coefficients, chromagram, mel spectrogram. Developed system can recognize emotions like calm, happy, disgust, angry, surprise, neutral and sad which are obtained from both RAVDESS and TESS datasets . For this combined dataset they achieved in an accuracy of 89%.

Ammar, Amjad.et.al., [10] have proposed a technique of Deep Convolutional neural network to extract the features from speech emotion databases. So, they adopted feature selection approach to find discriminative and important features of SER. They used classifiers such as Random Forest, Decision tree, Support Vector Machine (SVM),Multilayer perceptron (MLP), K-Nearest Neighbors(KNN) to classify seven emotions from publicly available datasets such as EMODB, SAVEE, RAVDESS, IEMOCAP and obtained accuracies as 92.02%,88.77%,93.61% and 77.23% respectively.

Loan Trinh Van.et.al., [11] have suggested a model of deep neural networks such as CNN, CRNN, GRU using Interactive Emotional Dyadic Motion Capture(IEMOCAP) with four emotions (anger, happiness, sadness, neutrality).For this they obtained highest accuracy for GRU of 97.47%. Their future work is to work on more emotions

III. METHODOLOGY

TESS DATASET:

The performance and robustness of recognition systems will suffer if system is not trained with sufficient dataset. To train and assess the performance of an emotion recognition system, it's crucial to have a sufficient dataset. So, in this work, Toronto Emotional Speech Set (TESS) dataset [1] is used. This dataset consists of female voice recordings and is of very high-quality audio. Most of the other datasets are focused on 2800 male voices which leads to imbalance representation. This dataset holds a set of 200 target words were spoken in the carrier phase “Say the word_” by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, fear, sad, disgust, pleasant surprise, neutral, happiness).

RAVDESS DATASET:

RAVDESS [27] includes 1440 files: 60 trials per actor times 24 actors equals 1440. The RAVDESS has 24 professional actors (12 female, 12 male) who deliver two lexically-matched lines in a neutral North American dialect. Speech emotions include calm, joyful, sad, angry, afraid, surprised, and disgusted expressions. Each expression is created in two levels of emotional intensity (normal and strong), as well as a neutral expression.

CONVOLUTION NEURAL NETWORK:

A CNN, or convolution neural network, is a form of artificial neural network used in deep learning for object and image recognition and classification. In CNNs(fig.1), there could be multiple hidden layers, which perform feature extraction from the image by doing calculations. Convolution Neural Network consists of multiple layers like the input layer, Convolution layer, Pooling layer, and fully connected layers. The input image is processed by the Convolution layer to extract features, the Pooling layer reduces computation by down-sampling the image, and the fully connected layer generates the final prediction. The network uses gradient descent and back propagation to determine the best filters.

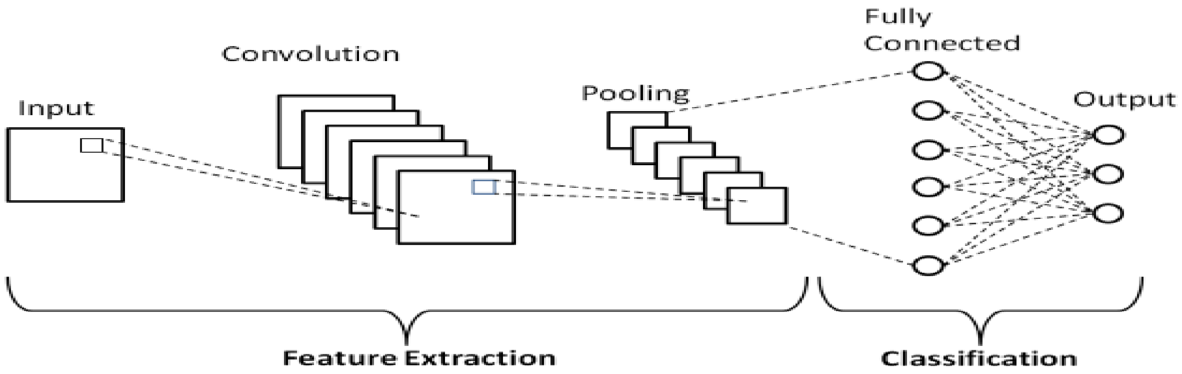


Fig.3 Block diagram of CNN

LONG SHORT-TERM MEMORY:

Long short-term memory (LSTM) network is a recurrent neural network (RNN) that avoids vanishing gradient problem. It is used for classification of emotions.

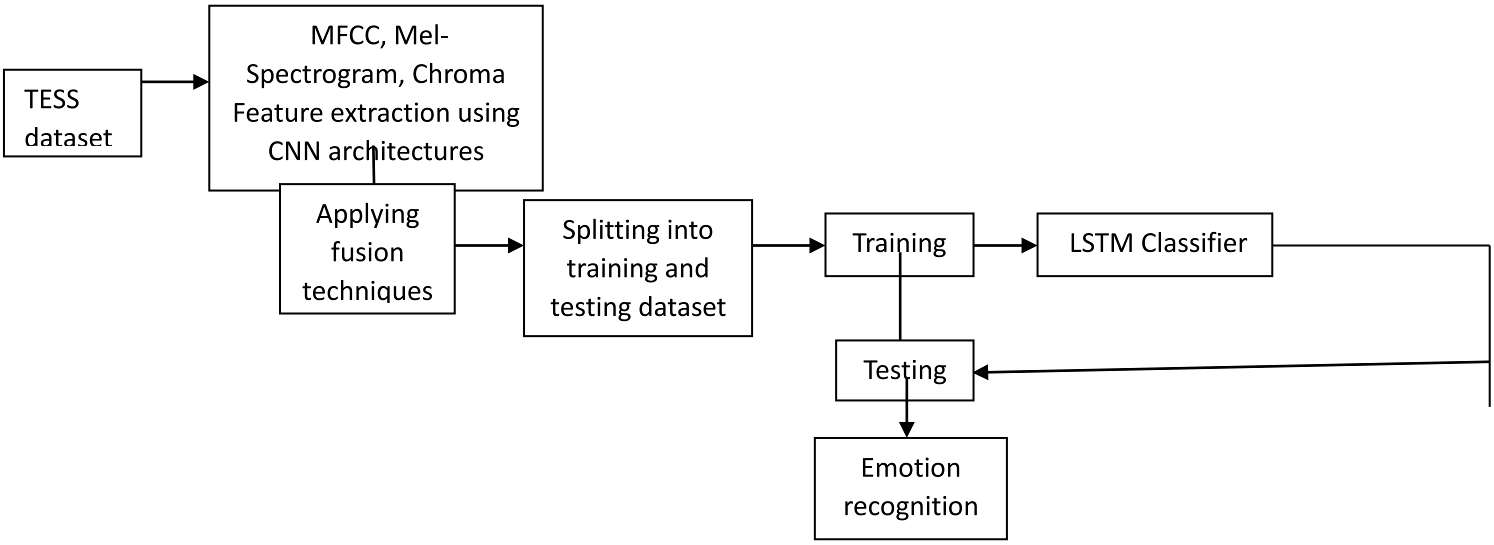

TESS dataset
MFCC, Mel-Spectrogram, Chroma Feature extraction using CNN architectures
Applying fusion techniques
Splitting into training and testing dataset
Training
LSTM Classifier
Testing
Emotion recognition

Fig.3 Block diagram of speech emotion recognition

IV. PROPOSED TECHNIQUES

PRE TRAINED NETWORKS:

Xception [16] is a deep convolutional neural network design that employs depthwise separable convolutions which reduces the computational complexity and the number of parameters in convolutional neural networks (CNNs). VGG16 [15] is crucial for its deep yet simple architecture that demonstrated the power of small, uniform convolutional filters to achieve significant improvements in image classification tasks, influencing subsequent deep learning model designs. ResNet152v2 [18], an improved version of the Residual Network (ResNet) architecture with 152 layers, that uses residual connections to solve the vanishing gradient problem, allowing for the training of incredibly deep networks. InceptionV3 [19] uses auxiliary classifiers during training to reduce the vanishing gradient issue and improve gradient flow across the network. MobileNetV2 [20] additionally uses shortcut connections and linear bottlenecks to improve information flow and gradient propagation, resulting in more efficient training and greater performance.DenseNet121 [21] is a convolutional neural network design with dense connection patterns that enable feature reuse and gradient flow across the network.

NASNetLarge [22] uses a cell-based architecture that is automatically found using reinforcement learning to optimize both the network's topology and hyper parameters. EfficientNetB5 [23] achieves cutting-edge performance on a wide range of computer vision tasks while requiring fewer parameters and computations than other models of comparable accuracy, making it ideal for resource-constrained environments like mobile and edge devices. EfficientNetB6 [17] has a hierarchical design made up of many blocks, each of which includes efficient convolutional procedures including depthwise separable convolutions and squeeze-and-excitation modules for effectively capturing rich hierarchical information.EfficientNetV2B2 [24] hierarchical structure with numerous blocks, each integrating sophisticated convolutional techniques such as depthwise separable convolutions and squeeze-and-excitation modules to effectively capture complicated features at various scales.EfficientNetV2M [25] is a variation of the EfficientNetV2 convolutional neural network architecture that aims to reconcile computational efficiency and good performance in computer vision applications. ConvNeXtTiny [26] is a convolutional neural network architecture that is optimized for efficient image categorization on resource-constrained devices

Fusion techniques:

Three fusion techniques are employed to exploit the network strength. Fusion techniques enable better representation of complex phenomena, improve classification or regression performance. Early fusion [12], also known as feature-level fusion, is a technique in data analysis and machine learning that combines information from different sources or modalities at an early stage of processing, usually before any learning or inference occurs. Mid fusion [13], also known as intermediate fusion, is a multi-modal learning approach in which features taken from distinct modalities are fused at an intermediate level before being passed into later layers for further processing. Late fusion [14], also known as decision-level fusion, is a multi-modal learning technique in which predictions or judgments from distinct models taught in different modalities are integrated at a later stage.

Emotion	precision	Recall	F1-Score
angry	1.00	1.00	1.00
disgust	0.96	1.00	0.98
fear	1.00	1.00	1.00
happy	1.00	0.97	0.99
neutral	1.00	1.00	1.00
ps	0.98	0.99	0.99
sad	0.99	0.97	0.98

V. EXPERIMENTAL RESULTS

Since deep learning has many advantages, Convolutional Neural network is considered for speech emotion recognition. Features are extracted from the audio's available in TESS dataset which is taken from the kaggle. In this work, some of libraries from pytorch are considered for emotion recognition. Adam optimizer is used, which has fast convergence and gives effective optimization. During training 50 epochs are considered and each of batch size 64 resulting in 80% accuracy.

The above graph represents the ROC curve which is plotted between False Positive Rate and True Positive Rate with different AUC values.

Pre-trained networks are considered to provide learned representations, enabling efficient transfer learning and feature extraction for new tasks.

By observing the results of all the networks EfficientB5 has good precision, recall, f1-score values for seven emotion with the highest accuracy of 99%.

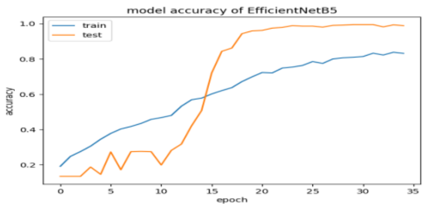
	MobilenetV2			ConvNextTiny			NasNetLarge		
Emotion	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
angry	0.54	0.99	0.70	0.97	1.00	0.99	0.56	0.97	0.71
disgust	1.00	1.00	1.00	1.00	0.94	0.97	0.54	1.00	0.70
fear	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	0.99
happy	1.00	0.08	0.16	0.99	0.94	0.96	0.97	1.00	0.98
neutral	0.99	1.00	0.99	1.00	1.00	1.00	0.24	0.04	0.07
ps	0.99	1.00	0.99	0.92	1.00	0.96	1.00	0.97	0.98
sad	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.03	0.06

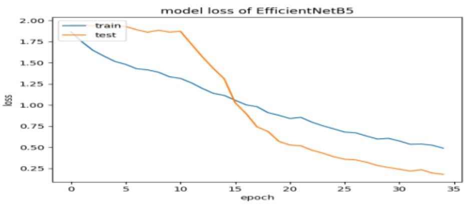
	Resnet152V2			EfficientnetV2B2			EfficientnetB6		
Emotion	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
angry	0.56	0.97	0.71	0.99	0.98	0.99	0.43	1.00	0.60
disgust	0.54	1.00	0.70	1.00	0.27	0.42	0.00	0.00	0.00
fear	0.98	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00
happy	0.97	1.00	0.98	0.55	0.99	0.71	1.00	0.91	0.96
neutral	0.24	0.04	0.07	0.00	0.00	0.00	1.00	1.00	1.00
ps	1.00	0.97	0.98	0.96	0.98	0.97	0.00	0.00	0.00
sad	1.00	0.03	0.06	0.48	1.00	0.65	0.99	1.00	1.00

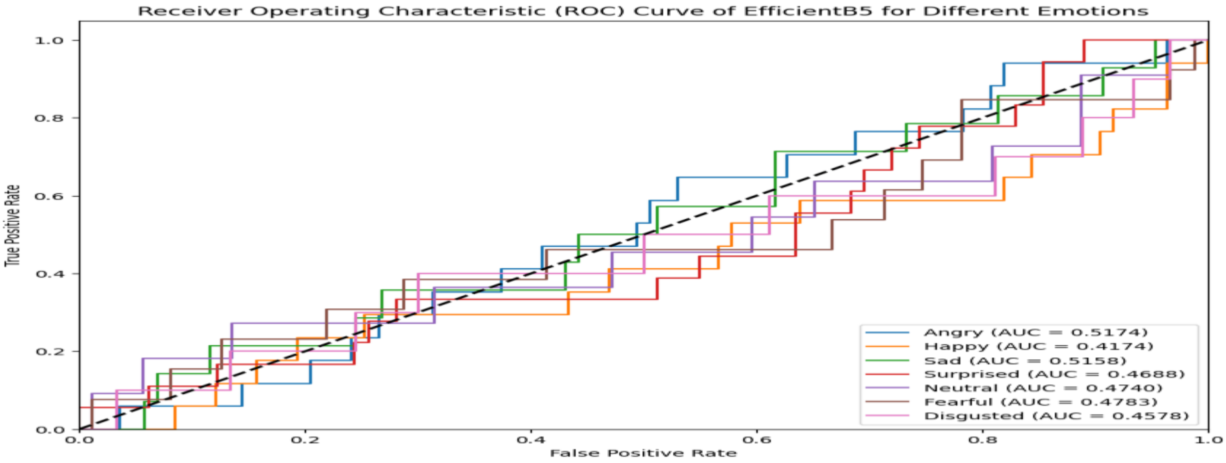
	Xception			VGG16			InceptionV3		
Emotion	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
angry	0.00	0.00	0.00	0.95	1.00	0.97	0.93	1.00	0.96
disgust	0.96	0.98	0.97	0.98	0.99	0.98	0.99	0.99	0.99
fear	1.00	1.00	1.00	1.00	0.91	0.95	0.98	0.94	0.96
happy	0.53	1.00	0.69	0.97	0.98	0.98	0.97	0.99	0.98
neutral	0.96	0.98	0.97	0.47	1.00	0.64	0.99	1.00	0.99
ps	0.98	0.97	0.98	0.97	0.98	0.98	0.99	0.90	0.95
sad	0.96	1.00	0.98	0.00	0.00	0.00	0.98	1.00	0.99

	Xception			VGG16			InceptionV3		
	EfficientB5			EfficientV2M			Densenet121		
Emotion	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
angry	0.97	0.99	0.98	1.00	0.78	0.88	1.00	0.99	1.00
disgust	1.00	0.99	0.99	0.99	0.98	0.98	0.49	0.98	0.65
fear	0.99	1.00	0.99	0.91	1.00	0.95	0.99	1.00	0.99
happy	1.00	0.96	0.98	0.87	1.00	0.93	1.00	0.12	0.21
neutral	1.00	1.00	1.00	0.99	0.98	0.98	1.00	0.99	0.99
ps	0.97	0.98	0.98	0.97	0.96	0.97	0.98	0.99	0.99
sad	0.99	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00

	Early fusion of Mobilenetv2 &	Mid fusion of ConvNextTiny &	Late fusion of Resnet152V2 &



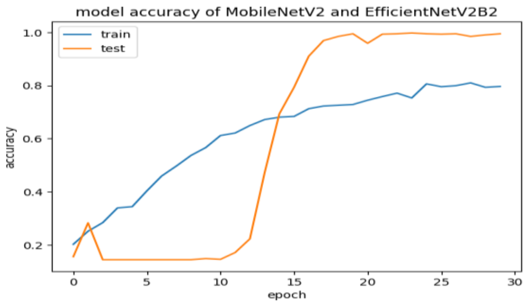


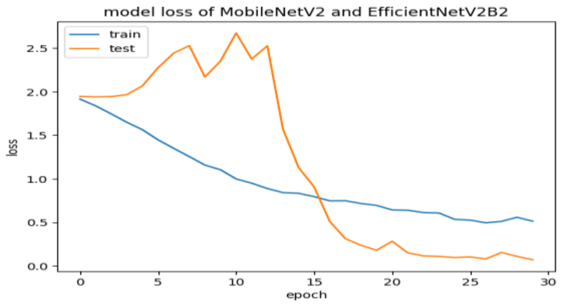


Features of two different networks are fused to get optimum results:

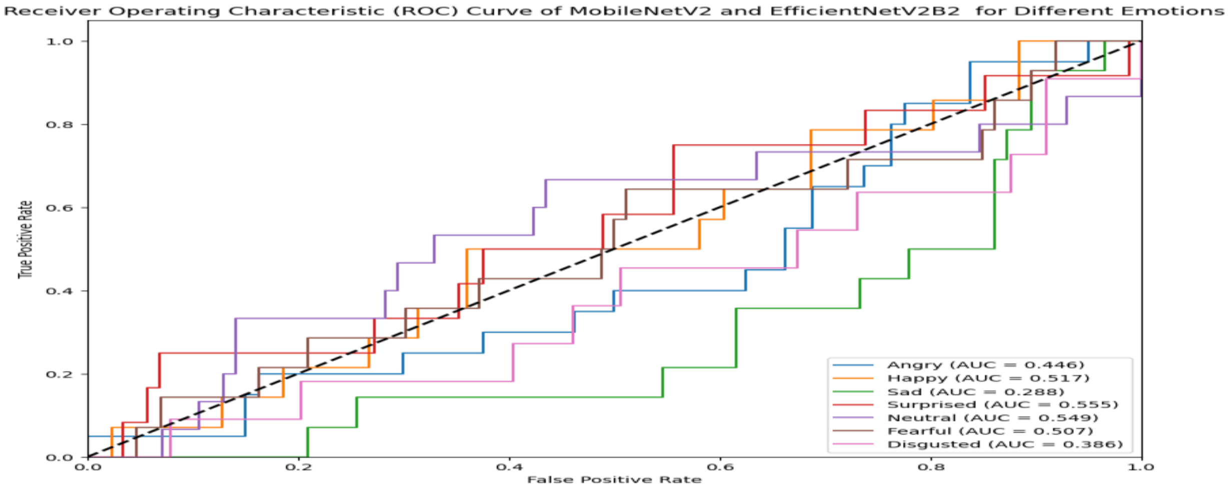
	EfficientNetV2B2			NasNetLarge			EfficientnetB6		
Emotion	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
angry	0.99	0.99	0.99	0.97	1.00	0.99	0.49	0.97	0.65
disgust	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.01	0.02
fear	0.99	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00
happy	1.00	0.98	0.99	1.00	0.98	0.99	1.00	0.97	0.98
neutral	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00
ps	0.99	0.99	0.99	0.99	0.95	0.97	0.96	1.00	0.98
sad	1.00	1.00	1.00	0.96	0.99	0.98	1.00	1.00	1.00

By observing the results of above information metrics, early fusion of MobilenetV2 & EfficientNetV2B2 achieved best precision, recall, f1-score values with an accuracy of 99%.



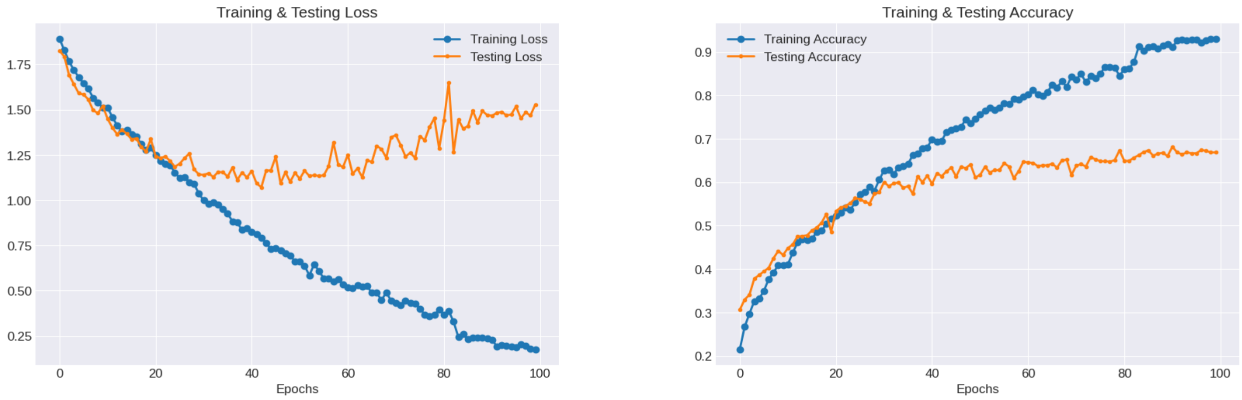


	Early fusion of EfficientB5 & Densenet121			Mid fusion of Xception & EfficientV2M			Late fusion of VGG16 & InceptionV3		
Emotion	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
angry	1.00	0.90	0.95	1.00	0.98	0.99	0.98	1.00	0.99
disgust	1.00	0.93	0.96	1.00	0.94	0.97	0.97	0.99	0.98
fear	0.99	0.96	0.97	1.00	1.00	1.00	1.00	1.00	1.00
happy	0.98	0.97	0.98	1.00	0.98	0.99	1.00	1.00	1.00
neutral	1.00	1.00	1.00	0.99	1.00	0.99	0.00	0.00	0.00
ps	0.83	0.99	0.90	0.92	1.00	0.96	0.99	0.95	0.97
sad	1.00	1.00	1.00	1.00	1.00	1.00	0.53	1.00	0.70



Using RAVDESS DATASET:

During training of dataset 100 epochs are considered which takes long time for training and resulted in the less accuracy.



Emotion	Precision	Recall	F1-Score
angry	0.69	0.68	0.69
disgust	0.63	0.57	0.60
fear	0.70	0.55	0.62
happy	0.51	0.55	0.53
neutral	0.81	0.83	0.82
ps	0.52	0.62	0.57
sad	0.76	0.77	0.76

CONCLUSION:

In this work features such as Mel-frequency cepstral coefficients (MFCC), chroma, mel- spectrogram are extracted from the audio's available in TESS dataset using the CNN network architectures like Xception, Vgg16, Mobilenetv2 etc. Extracted features from two different networks are combined using fusion techniques such as early, late, mid fusions. Finally features are classified using the algorithm of LSTM resulted in the accuracy of 99%. Furthermore, the work is extended to the RAVDESS dataset, which took a lengthy time to train using 100 epochs but resulted in a lower accuracy of 6%.Considering the accuracies of TESS and RAVDESS, the TESS dataset is further employed using pre-trained networks and fusion methods, resulting in an increased accuracy of 99%.

REFERENCES:

1. Kate Dupuis and M. Kathleen Pichora-Fuller (2010), Toronto emotional speech set (TESS), University of Toronto, Psychology Department
2. Yoon, Seunghyun, Seokhyun Byun, and Kyomin Jung. "Multimodal speech emotion recognition using audio and text." *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018.
3. Huang, Kun-Yi, et al. "Speech emotion recognition using convolutional neural network with audio word-based embedding." *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2018.
4. Lalitha, S., Shikha Tripathi, and Deepa Gupta. "Enhanced speech emotion detection using deep neural networks." *International Journal of Speech Technology* 22 (2019): 497-510.
5. Kerkeni, Leila, et al. "Automatic speech emotion recognition using machine learning." (2019).
6. Christy, A., et al. "Multimodal speech emotion recognition and classification using convolutional neural network techniques." *International Journal of Speech Technology* 23 (2020): 381-388.
7. Er, Mehmet Bilal. "A novel approach for classification of speech emotions based on deep and acoustic features." *IEEE Access* 8 (2020): 221640-221653.
8. Chauhan, Krishna, Kamalesh Kumar Sharma, and Tarun Varma. "Speech emotion recognition using convolution neural networks." *2021 international conference on artificial intelligence and smart systems (ICAIS)*. IEEE, 2021.
9. Asiya, U. A., and V. K. Kiran. "Speech Emotion Recognition-A Deep Learning Approach." *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*. IEEE, 2021.
10. Amjad, Ammar, Lal Khan, and Hsien-Tsung Chang. "Effect on speech emotion classification of a feature selection approach using a convolutional neural network." *PeerJ Computer Science* 7 (2021): e7.
11. Trinh Van, Loan, et al. "Emotional speech recognition using deep neural networks." *Sensors* 22.4 (2022): 1414.
12. Schuller, Björn, et al. "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles." (2005).
13. Boulahia, Said Yacine, et al. "Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition." Machine Vision and Applications 32.6 (2021): 121.
14. Ibrahim, Hemin, Chu Kiong Loo, and Fady Alnajjar. "Speech emotion recognition by late fusion for bidirectional reservoir computing with random projection." *IEEE Access* 9 (2021): 122855-122871.

15. Aggarwal, Apeksha, et al. "Two-way feature extraction for speech emotion recognition using deep learning." Sensors 22.6 (2022): 2378.

16. Reimao, Ricardo Amaral Martins. "Synthetic speech detection using deep neural networks." (2019).

17. Tatar, Bahadır. *Convolutional neural networks (CNN) based binary classifiers for construction machinery detection*. MS thesis. 2022..

18. Kamble, Ashwin, Pradnya H. Ghare, and Vinay Kumar. "Deep-learning-based BCI for automatic imagined speech recognition using SPWVD." IEEE Transactions on Instrumentation and Measurement 72 (2022): 1-10..

19. Liu, Dong, et al. "Multi-modal fusion emotion recognition method of speech expression based on deep learning." *Frontiers in Neurorobotics* 15 (2021): 697634.

20. Zhong, Ying, et al. "A Lightweight Model Based on Separable Convolution for Speech Emotion Recognition." *Interspeech*. Vol. 11. 2020.

21. Pusarla, Nalini, Anurag Singh, and Shrivishal Tripathi. "Learning DenseNet features from EEG based spectrograms for subject independent emotion recognition." *Biomedical Signal Processing and Control* 74 (2022): 103485.

22. Zaman, Khalid, et al. "Driver emotions recognition based on improved faster R-CNN and neural architectural search network." *Symmetry* 14.4 (2022): 687.

23. Lu, Quy Thanh. "An Approach for Classification of Diseases on Leaves." *International Journal of Advanced Computer Science and Applications* 14.10 (2023).

24. Hasan, Yumnah, et al. "A Convolutional Neural Network Based Patch Classifier Using Mammograms." *ICAART (3)*. 2023.

25. Saxena, Aditya, et al. "Efficient Net V2 Algorithm-Based NSFW Content Detection." *International Conference on Information Technology*. Singapore: Springer Nature Singapore, 2023.

26. Jose, Jiby Mariya, and Shajulin Benedict. "DeepASD Framework: A Deep Learning-Assisted Automatic Sarcasm Detection in Facial Emotions." *2023 8th* International Conference on Communication and Electronics Systems (ICCES). IEEE, 2023.

27. Livingstone, S.R. RAVDESS Emotional Speech Audio Emotional Speech Dataset. 2018.