Emotion Detection using Telugu speech

Aditya Rajesh Sakri
CSE-AI
Amrita Vishwa Vidyapeetam
Bangalore, India
adityasakri@gmail.com

vaka satwik reddy
CSE-AI
Amrita Vishwa Vidyapeetam
Bangalore, India
satwik91@gmail.com

Vamsi krishna Vunnam
CSE-AI
Amrita Vishwa Vidyapeetam
Bangalore, India
vamsivunnam129@gmail.com

Dr. Suja.P
CSE-AI
Amrita Vishwa Vidyapeetam
Bangalore, India
p_suja@blr.amrita.edu

I. ABSTRACT

Emotion detection from speech is a challenging task that has gained significant attention in recent years. The ability to recognize emotions from speech signals has important applications in fields such as human-computer interaction, affective computing, and healthcare. Deep learning, a subset of machine learning, has shown promising results in emotion recognition tasks. Telugu, one of the major languages spoken in India, has not received much attention in emotion recognition from speech signals using deep learning techniques. In this report, we aim to investigate the performance of deep learning algorithms for emotion detection in Telugu speech signals. The quickest and most natural form of human communication is speech. We assess the emotional state of the person we are interacting with and respond to them appropriately in our daily interactions. Spoken emotion recognition is tough and complex for a number of reasons. how well different speech parameters may identify emotions in Telugu speech. We here make use of 550 audio signals of Telugu emotional speech that spans five emotional categories: surprise, anger, happiness, neutrality, and sadness. We first train the model with the dataset consisting of five different emotions using the neural networks and then test the model based on the remaining untrained dataset. For training we will explore various deep learning techniques such as Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks. The Main goal of this report is to develop a robust and accurate emotion recognition system for Telugu speech signals using deep learning techniques. By doing so, we hope to contribute to the advancement of emotion recognition research in Telugu and facilitate the development of emotion-aware applications in this language.

Keywords— Deep Learning, CNN, LSTM, Emotion Detection

II. INTRODUCTION

The process of determining a speaker's underlying emotional state based on their speech patterns is known as emotion detection from speech. Applications for this technology can be found in a number of industries, including healthcare, customer service, and education. Due to its capability to automatically learn and extract features from unprocessed speech signals, it has demonstrated great promise in the field of emotion detection from speech.

Deep learning techniques for emotion detection from speech typically use spectrograms or mel-frequency cepstral coefficients (MFCC) extracted from the audio signal as input data. These characteristics are fed into an architecture of a neural network that is trained on labelled data to discover the patterns connected to various emotions. The model's output is a forecast of the speaker's emotional state.

Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and their combinations, such as the convolutional recurrent neural network, are among the neural network architectures that have been used for emotion detection from speech (CRNN). RNNs can detect temporal dependencies in the audio signal, while CNNs are proficient at learning local patterns in spectrograms. The strengths of both CNNs and RNNs are combined in CRNNs, which makes them ideal for speech emotion recognition.

Deep learning-based emotion detection from speech is an exciting new field of study with numerous real-world applications. Future emotion detection systems should become more precise and reliable as labelled datasets become more readily available and deep learning models advance.

III. LITERATURE SURVEY

Theses are the literature review of related works and existing systems in the Emotion Detection from speech of Different Languages using Deep Learning.

The paper[1] by Yi-Lin Lin and Gang Wei proposes a speech emotion recognition system that combines Hidden Markov Model (HMM) and Support Vector Machine (SVM) algorithms. The proposed system first extracts Mel Frequency Cepstral Coefficients (MFCCs) and delta coefficients from speech signals, which are then used as features for the HMM-based modeling of emotional states. The SVM classifier is then used to classify the emotional states predicted by the HMM. The proposed system was evaluated on a dataset of Mandarin Chinese speech signals and achieved a recognition rate of 80.7%.

In their paper[2] titled "Emotion Detection From Speech Using Mfcc and Gmm," Patil, Zope, and Suralkar suggest a method for identifying emotions in speech by combining MFCCs and Gaussian Mixture Models (GMMs) The Berlin Emotional Speech Database (EmoDB), which contains speech samples of actors expressing various emotions, was used by the authors to evaluate the proposed system. According to the findings, the suggested system had an overall accuracy of 84.4%. For each class of emotion, the system also attained high precision and recall values.

The paper[3] provides a thorough analysis of recent developments in speech emotion recognition using deep learning techniques. It covers a variety of topics, such as feature extraction, emotion recognition databases, and deep learning models. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention-based models are just a few examples of the various deep learning architectures and techniques that are covered by the authors.

The paper [4] "Automatic Speech Emotion Recognition Using Recurrent Neural Networks with Local Attention" by Mirsamadi, Barsoum, and Zhang proposes an emotion recognition system that utilises RNNs with a local attention mechanism. The authors tested their system using speech samples of actors expressing different emotions from the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset. The results showed that the proposed system had an accuracy rate of 62.2%.

The paper[6] "Speech Emotion Recognition Using Attention-Based LSTM" by Xie, Yue, et al. proposes a system for speech emotion classification using a combination of Long Short-Term Memory (LSTM) and attention mechanisms.

The paper[7] discusses speech based emotion recognition. The authors did many number of methods, such as prosody, spectral, and cepstral features, for extracting emotional features from speech signals. They also talk about how different machine learning algorithms can be used to identify speech signals that contain emotions.

The paper[8] presents a study on speech Er using machine learning techniques. The authors extract Mel-frequency cepstral coefficients (MFCC) as features from the Berlin Emotional Speech Database. For emotion classification, four machine learning algorithms—K-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM)—are used. The authors show that using machine learning for speech emotion recognition is feasible by achieving an accuracy of up to 84% using the SVM algorithm.

The paper[9] suggests a deep learning-based method for speech emotion recognition using long short-term memory (LSTM) networks and 1D and 2D convolutional neural networks (CNNs). The proposed method was tested on a dataset of recordings of emotional speech, and it produced high recognition accuracy, proving its efficacy. Potential applications for this work include the healthcare industry and human-computer interaction.

Using Mel-Frequency Cepstral Coefficients (MFCCs) and Support Vector Machines, the paper[10] focuses on the task of emotion recognition from Telugu speech (SVMs). The Telugu Emotional Speech Corpus is used by the authors to extract MFCCs as features. The happy, sad, angry, and neutral categories are used to train the SVM algorithm to categorise speech utterances. According to the authors, the emotion classification task had an overall accuracy rate of 81.2%.

The paper[11] talks about the task of speech emotion recognition (SER), which entails inferring a speaker's emotional state from their speech signal. In a number of applications, including human-computer interaction, speech therapy, and emotion-based marketing, the authors emphasize the value of SER.

A simple speech emotion recognition (SER) system based on deep frequency characteristics and convolutional neural networks is discussed in the paper[12] (CNNs). The importance of SER is emphasised by the authors in a number of contexts, such as customised human-computer interaction and mental health monitoring. The authors first use the speech signal's Mel frequency cepstral coefficients (MFCCs) to extract deep frequency information.

The paper[13] talks about a speech emotion recognition (SER) system based on discriminant temporal pyramid matching and deep convolutional neural networks (CNNs) (DTPM). The authors emphasise the value of SER in a variety of fields, including entertainment, healthcare, and human-computer interaction.

The paper[14] "Speech recognition using deep neural networks: A systematic review" . The paper discusses various DNN architectures used in ASR, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid models. The authors also review different techniques used for feature extraction, such as Mel-frequency cepstral coefficients (MFCCs) and deep feature learning

The paper[15] talks about the various methods and strategies that have been employed for speech emotion recognition, including more recent deep learning methods like convolutional neural networks and long short-term memory networks as well as more established machine learning algorithms like decision trees and support vector machines.

The paper[16] talks about examining the drawbacks of conventional methods for continuous emotion identification, which frequently rely on hand-crafted characteristics and presume that a speaker's emotional state remains constant during brief time periods.. The RECOLA dataset, which comprises continuous emotion annotations for speech signals captured throughout diverse settings, is used in this study to describe experimental findings.

Paper[17] talks about Convolutional neural networks (CNNs) and long short-term memory (LSTM) networks which are only a couple of the deep learning methods that have been employed for voice emotion identification.The study's dataset, which consists of voice recordings of 10 different emotions spoken by six male and six female Telugu speakers, is described in the paper.
The paper[19] proposes a Telugu speech emotion recognition system using Support Vector Machines (SVMs). The authors extract Mel-frequency cepstral coefficients (MFCCs) as features and train SVM classifiers for four emotions

The paper[20] talks about Mel-Frequency Cepstral Coefficients (MFCCs) and Support Vector Machines to identify emotions in Telugu speech (SVMs). The authors

point out that SVMs are a well-liked machine learning method that have been utilised for speech emotion identification in different languages, and that MFCCs are often used features for speech processing.

IV. DATA DESCRIPTION

The Emotion detection using Telugu speech dataset is a collection of audio recordings of individuals speaking Telugu, a Dravidian language spoken in the Indian states of Andhra Pradesh and Telangana, expressing a range of emotions. The dataset was created by us from YouTube and different multimedia applications.

The dataset includes audio recordings of individuals speaking Telugu. The length and quality of each recording may vary.it includes emotions expressed by the speakers such as happiness, sadness, anger, fear, surprise, and neutral. And also, the information about the speakers like male or female. The data is stored in the format of wav. Each audio file is labeled with the corresponding emotion expressed by the speaker. The emotions included in the dataset are happy, sad, angry, fearful, surprised, and neutral. The dataset is divided into train, validation, and test sets. Seventy percent of the data is in the training set, 15% is in the validation set, and 15% is in the test set.

Extracted features for emotion detection from Telugu speech such as mel-frequency cepstral coefficients (MFCCs)

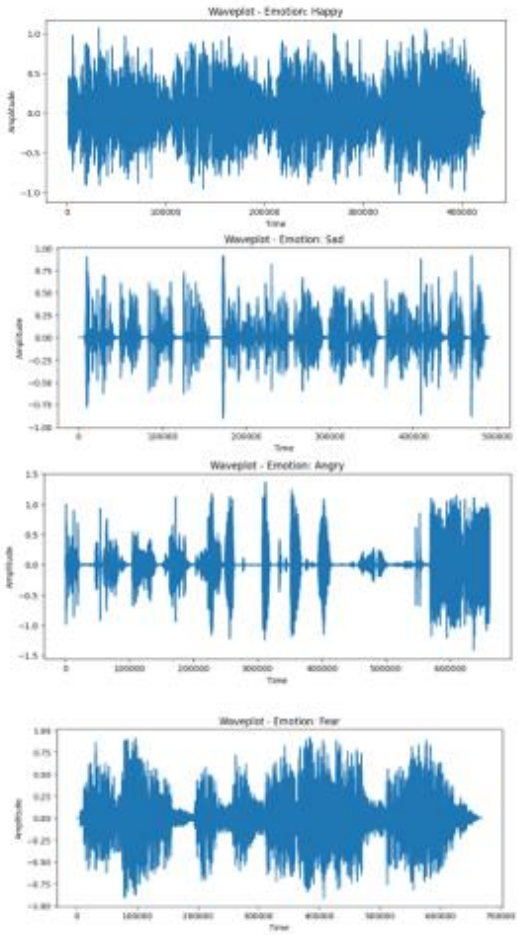

Fig 1: Audio Signals of Four Emotion in dataset

V. METHODOLOGY

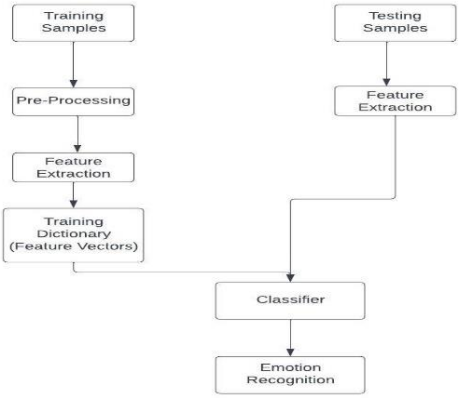

Fig2: Methodology and Flow Chart

To enable machine learning algorithms to process the target variable, the 'label' column was transformed into numeric values. This was accomplished by substituting numerical representations for each emotional state. Specifically, "Angry" was changed to "0," "Sad" to "1," "Happy" to "2," and "Fear" to "3." This encoding makes it easier for the model to comprehend the categorical nature of the target variable.

Using the OneHotEncoder class from the sklearn.preprocessing module, the target labels were further prepared for training. Each label is now represented as a one-hot encoded vector after this transformation. The one-hot encoding, which assigns a binary value to each category in independent dimensions, makes sure that the model comprehends the categorical character of the target variable.

VI. NETWORK ARCHITECTURE

A) MLP classifier

In an MLP classifier for emotion recognition using speech, a network architecture with two hidden layers is used.

The input layer represents the input speech signal, and it has a number of neurons equal to the number of features extracted from the speech signal.

Two Hidden layer consists of a number of neurons that can be tuned during the training phase. The activation function used in this layer is a non-linear function like Relu or sigmoid function.

With learnable parameters, the Alex net has eight layers. Relu activation is used in each of the model's five levels, with the exception of the output layer, where it is used in combination with max pooling and three fully connected layers.

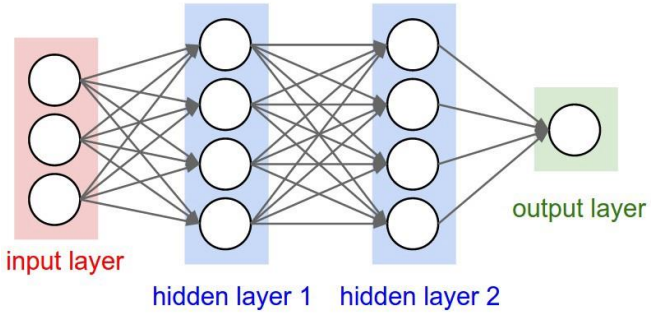

Fig2: Basic Network architecture of two hidden layers

B) CNN architecture

Convolutional Neural Networks (CNNs) are a class of deep learning models that are frequently employed for processing image and sequence data. They are ideal for applications like image classification, object detection, and natural language processing because they are particularly good at capturing spatial and temporal patterns.

For the analysis of sequence data, we chose a CNN design as our architecture. A Convolutional layer (Conv1D) with 64 filters, each of size 3, and a rectified linear unit (ReLU) activation function make up the model's foundation. Convolutions are applied to the input data at this layer to extract local characteristics and improve their representations. To avoid overfitting, regularization is conducted using L2 regularization.

A Max Pooling layer (MaxPooling1D) is then added, with a pool size and stride of 2 and respectively. By lowering the spatial dimensionality and extracting the most important data, this layer samples the feature maps. Two extra Convolutional layers with higher filter sizes (128 and 256) and ReLU activation are added to the architecture.

To identify complex patterns in the data, a dense layer with 512 units and ReLU activation is added. By randomly deactivating a portion of the units during training, dropout regularization is used to avoid overfitting. The output layer, which is represented by a Dense layer with 4 units and a softmax activation function, completes the model. The model can categorize the input sequence into one of the emotions (Angry, Sad, Happy, or Fear) thanks to the probability distribution produced by the softmax activation over the 4 emotion classes.

The categorical cross-entropy loss function, which is suitable for multi-class classification issues, is built into the model. The model's weights are optimized using the Adam optimizer, and its effectiveness in training and testing is assessed using the accuracy metric.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d (Conv1D)             (None, 38, 64)            256

 max_pooling1d (MaxPooling1D  (None, 19, 64)           0
 )

 conv1d_1 (Conv1D)           (None, 17, 128)           24704

 max_pooling1d_1 (MaxPooling  (None, 8, 128)           0
 1D)

 conv1d_2 (Conv1D)           (None, 6, 256)            98560

 max_pooling1d_2 (MaxPooling  (None, 3, 256)           0
 1D)

 flatten (Flatten)           (None, 768)               0

 dense (Dense)               (None, 512)               393728

 dropout (Dropout)           (None, 512)               0

 dense_1 (Dense)             (None, 4)                 2052

=================================================================
Total params: 519,300
Trainable params: 519,300
Non-trainable params: 0
_________________________________________________________________
```

Fig 3: CNN Model Architecture from Colab

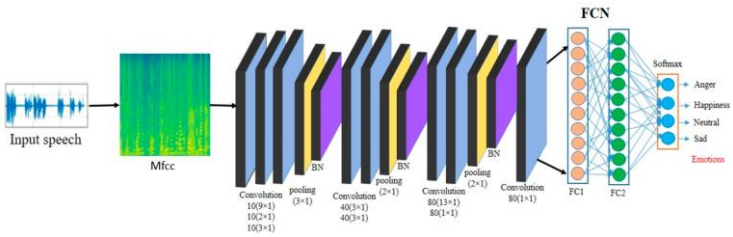

Fig4: Network Architecture Diagram of CNN

C) Alexnet Architecture

This model has several convolutional layers with different filter sizes and strides, followed by max pooling and activation functions. The input tensor shape is (batch_size, 40, 1), and the output tensor shape is (batch_size, 3). The model also includes several dense layers with ReLU activation functions and dropout layers to prevent overfitting. The first dense layer has 4096 units, and the second dense layer has the same number of units. The output layer uses the sigmoid activation function, making this model suitable for multi-label classification tasks.

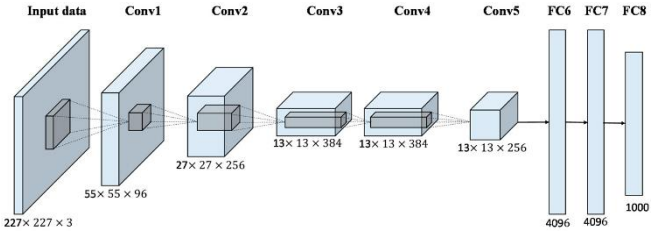

Fig4: Network Architecture Diagram using Alexnet.

D) Applying Ml Classifier at Dense Layer

We trained deep learning models to extract features from the output of the last dense layer. It then uses these features as input to a K-Nearest Neighbors (k-NN), decision tree and random forest classifier to predict the labels of the test data. The output of the last dense layer is reshaped into a 2D array and used to train the classifier. The classifier is trained with a value of k=5 (i.e., the 5 nearest neighbors are used to make a prediction).

The ml classifiers we used is KNN, Decision Tree, Random Forest

E) Regularization Techniques

Regularization in CNN (Convolutional Neural Network) refers to techniques used to prevent overfitting of the model during the training process. Overfitting occurs when a model learns the specific features of the training data. Regularization methods aim to prevent this by adding constraints to the model parameters or by adding noise to the input or hidden layers of the network.

Some Important Regularization Techniques are L1 and L2 Regularization, Early Stopping, Dropout

i. L1 and L2 Regularization

L1 and L2 regularization are techniques used to prevent overfitting in Deep learning models. L1 regularization encourages sparsity and feature selection by imposing a penalty equal to the absolute value of the weights. With L2 regularizations, smaller weights are encouraged for all parameters by adding a penalty proportionate to the squared value of the weights. L2 regularizations makes models less sensitive to minute input changes and more resilient to outliers. The regularizes are used in the given code to do L2 regularization. Large weights in the convolutional and dense layers are penalized by the l2 function, which regulates the model's complexity.

F) Optimization Techniques

Optimization in deep learning is a technique used to improve the performance of artificial neural networks. It involves finding the best set of parameters or weights for the neural network so that it can make accurate predictions or classifications. They have multiple layers of interconnected nodes or neurons, each with their own set of weights that determine how much influence they have on the final output. During training, the model adjusts these weights to minimize the error between its predictions and the actual outputs from the data.

i) Adam

Adam optimizer is used in conjunction with the categorical cross-entropy loss function. To reduce loss and increase the model's ability to correctly predict emotion labels, the optimizer modifies the weights of the model during training. It combines the advantages of RMSprop, which uses moving averages of gradient magnitudes, and AdaGrad, which adjusts the learning rate for each parameter.

G) LSTM (Long Short-Term memory)

LSTM (Long Short-Term Memory) is a type of recurrent neural network (RNN) architecture that addresses the vanishing gradient problem and enables the model to capture long-term dependencies in sequential data. It is frequently used in numerous tasks involving time series or sequential data, including sentiment analysis, speech recognition, and natural language processing. Two Dense layers with 64 units each and 'relu' activation function follow the LSTM layer. These layers introduce non-linearity and facilitate feature extraction and transformation.

The final Dense layer has 4 units, corresponding to the 4 emotional states ('Angry', 'Sad', 'Happy', 'Fear'). The softmax activation function is applied, enabling the model to output probabilities for each class, indicating the predicted emotional state.

The model is compiled with the categorical cross-entropy loss function, which is suitable for multi-class classification tasks. The 'adam' optimizer is used to optimize the model parameters, and the accuracy metric is employed to evaluate the model's performance.

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 256)               264192

 dropout_1 (Dropout)         (None, 256)               0

 dense_2 (Dense)             (None, 64)                16448

 dropout_2 (Dropout)         (None, 64)                0

 dense_3 (Dense)             (None, 64)                4160

 dropout_3 (Dropout)         (None, 64)                0

 dense_4 (Dense)             (None, 4)                 260

=================================================================
Total params: 285,060
Trainable params: 285,060
Non-trainable params: 0
```

Fig 5: LSTM Model Architecture from Colab

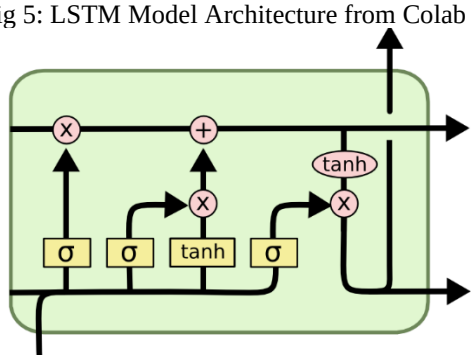

Fig 6: LSTM cell

VII. EXPERMENTAL WORK

A) MFCC Calculation

Mel-frequency cepstral coefficients (MFCCs) are widely used features for audio analysis. They capture the spectral characteristics of an audio signal, particularly useful for speech and music processing. The provided code snippet extracts MFCCs from a Telugu audio file using the Librosa library. It computes the mean MFCC values across each feature dimension and returns a data frame. These features can be used for Telugu emotion classification using deep learning models.

The librosa.feature.mfcc function computes the MFCCs from the audio data, specifying the number of MFCC coefficients to be extracted (n_mfcc). The resulting MFCCs are then averaged along each feature dimension to obtain the mean MFCC values using mfccs.mean(axis=1).

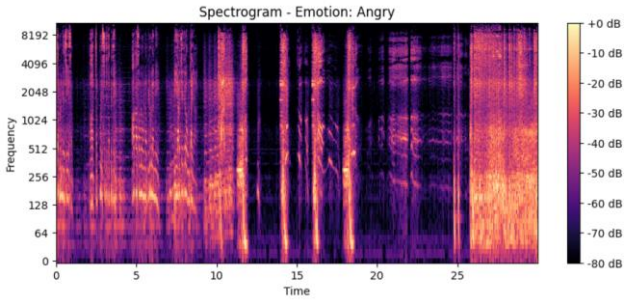

Fig: 7 Spectrogram representation of MFCC features

B) Training

First, split the dataset into training and testing at 80% and 20 % and form that 80% training data, the data split into 75% training and 25% validation. So, training, validation, and testing sets are 60%, 20%, 20%.

C) Models Used

MLP, CNN, ALEXNET, ML-CNN, LSTM

VIII. RESULTS

A) MLP Classifier

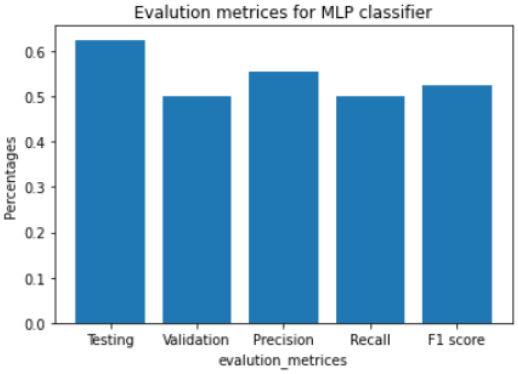

Fig8: Evaluation metrices of mlp classifier

We got accuracy of 64% and validation accuracy 51%. Because The amount of data available for training the MLP model may not be enough to learn the underlying patterns in the data and result in low accuracy.

B) CNN (Convolutional Neural Network)

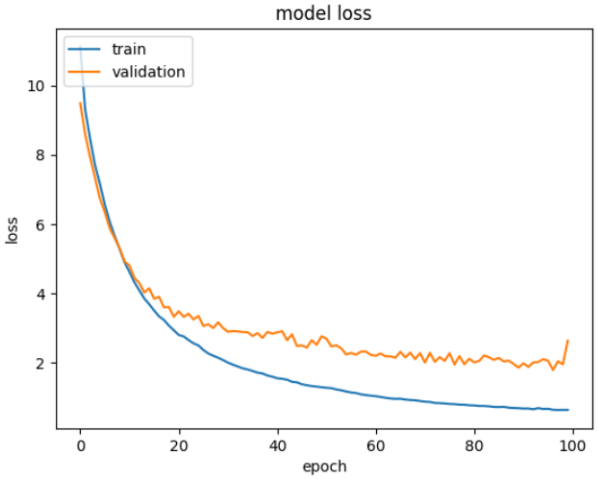

Fig 9: Training and validation loss (CNN)

The train loss is computed by evaluating the model's performance on the training data during each training iteration. A decreasing train loss indicates that the model is effectively learning from the data and minimizing the error. The declining trend in the train loss signifies that the model's ability to fit the training data is improving over time.

The validation loss is computed by evaluating the model's performance on a separate validation dataset. It serves as an estimate of how well the model generalizes to unseen data. Similar to the train loss, a decreasing validation loss indicates that the model is becoming more accurate and generalizing better to new data. This trend suggests that the model is not overfitting, as it performs well not only on the training data but also on previously unseen data.

C) ALEXNET

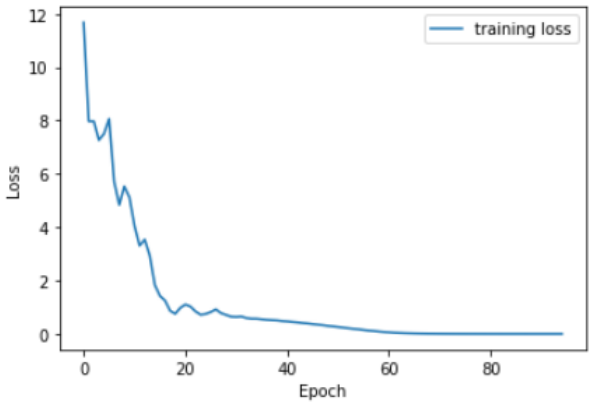

Fig10: Training loss of Trained Network

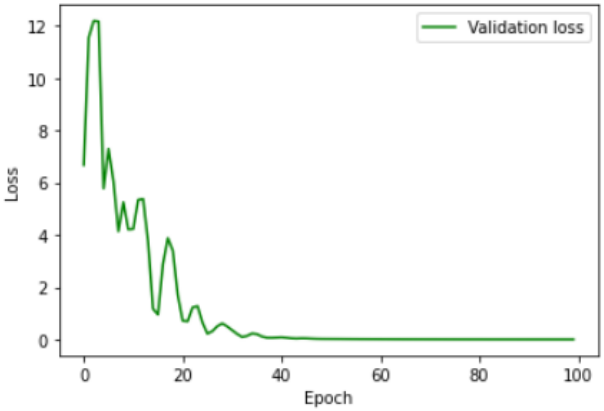

Fig11: Validation Loss of Trained Network

The training loss is decreasing, and the validation loss is also decreasing, then the model is not overfitting and is fitting data well.

D) ML classifier at Dense Layer

The output of the last dense layer is reshaped into a 2D array and used to train the classifier. The classifier is trained with a value of k=5 (i.e., the 5 nearest neighbors are used to make a prediction).

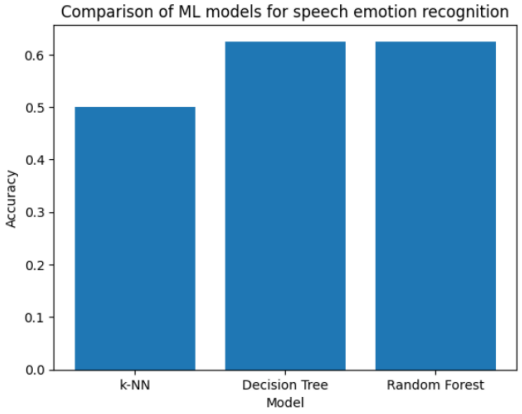

Fig12: ml models by feature extraction

Model	Accuracy	Precision
KNN	0.50	0.19
Decision tree	0.62	0.52
random forest	0.62	0.52

Fig12: At Dense layer ml model accuracies

E) RNN

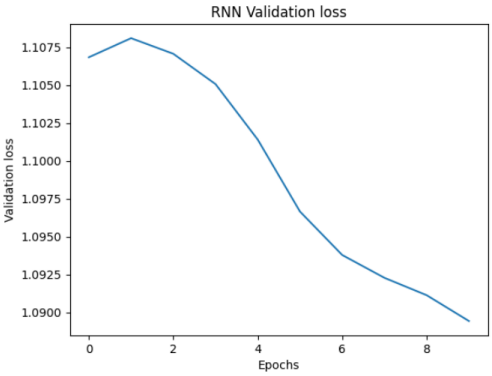

Fig16: RNN Validation Loss

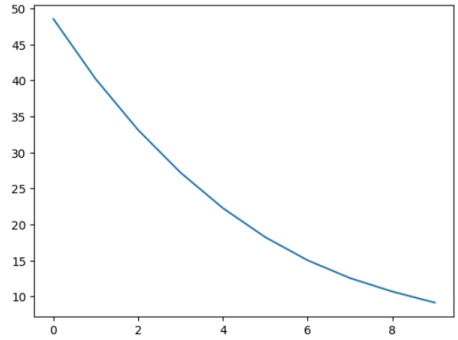

Fig17: RNN Validation Loss

On the test data, the RNN model for audio emotion detection had a 50% accuracy rate. Despite the fact that this outcome is not the best for many applications, it might still be helpful for some tasks where a simple binary classification of positive and negative emotions is sufficient.

because it was not trained on a large enough dataset of audio samples. Increasing the number of layers and units or using a more advanced architecture such as a convolutional LSTM or attention-based model could improve accuracy.

REFERENCES

[1] Yi-Lin Lin and Gang Wei, "Speech emotion recognition based on HMM and SVM," 2005 International Conference on Machine Learning and Cybernetics, Guangzhou, China, 2005, pp. 4898-4901 Vol. 8, doi: 10.1109/ICMLC.2005.1527805.

[2] Patil, K.J.; Zope, P.H.; Suralkar, SSBT's college of Engineering, India Emotion Detection From Speech Using Mfcc and Gmm. Int. J. Eng. Res. Technol. (IJERT) 2012, ISSN: 2278-0181

[3] Abbaschian, B.J.; Sierra-Sosa, D.; Elmaghraby, A. Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models.2021, 21, 1249.

[4] S. Mirsamadi, E. Barsoum and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 2017, pp. 2227-2231, doi: 10.1109/ICASSP.2017.7952552.

[5] Li, Yuanchao, Tianyu Zhao, and Tatsuya Kawahara. "Improved End-to-End Speech Emotion Recognition Using Attention Mechanism and Multitask Learning." Interspeech. 2019.

[6] Xie, Yue, et al. "Speech emotion classification using attention-based LSTM." IEEE/ACM Transactions on Audio, Speech, and Language Processing 27.11 (2019): 1675-1685.

[7] Basu, Saikat, et al. "A review on emotion recognition using speech." 2017 International conference on communication and computational technologies (ICICCT). IEEE, 2017.

[8] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in IEEE Access, vol. 7, pp. 117327-117345, 2019, doi: 10.1109/ACCESS.2019.2936124.

[9] Jianfeng Zhao, Xia Mao, Lijiang Chen, Speech emotion recognition using deep 1D & 2D CNN LSTM networks, Biomedical Signal Processing and Control, Volume 47, 2019, 312-323,ISSN 1746-8094

[10] Trigeorgis, George, et al. "Adieu features? speech emotion recognition using a deep convolutional recurrent network." 2016 IEEE international conference on speech and signal processing (ICASSP). IEEE, 2016.

[11] Harár, P., R. Burget, and M. Kishore Dutta. "Speech Emotion Recognition with studies." Proceedings of the 4th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India. 2017.

[12] Anvarjon, Tursunov, and Soonil Kwon. "Deep-net: A lightweight CNN-based speech emotion recognition system using deep frequency features." Sensors 20.18 (2020): 5212.

[13] Zhang, Shiqing, et al. "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching." IEEE Transactions on Multimedia 20.6 (2017): 1576-1590.

[14] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh and K. Shaalan, "Speech recognition using neural networks: A systematic review", IEEE Access, vol. 7, pp. 19143-19165, 2019.

[15] S. Lalitha, A. Madhavan, B. Bhushan and S. Saketh, "Speech emotion recognition", Proc. Int. Conf. Adv. Electron. Comput. Commun. (ICAECC), pp. 1-4, Oct. 2014.

[16] J. Han, Z. Zhang, F. Ringeval and B. Schuller, "Prediction-based learning for continuous emotion recognition in speech", Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), pp. 5005-5009, Mar. 2017.

[17] P. Pandey and D. Chakraborty, "A Study on Emotion Recognition in Telugu Speech using Deep Learning Techniques," International Journal of Computer Applications, vol. 181, no. 27, pp. 22-27, 2018.

[18] M. Venkata Ramana and M. H. R. Prasad, "Telugu Speech Emotion Recognition using Support Vector Machines," International Journal of Computer Applications, vol. 119, no. 9, pp. 37-42, 2015.

[19] K. Venkata Krishnaiah and R. Sreenivasa Rao, "Emotion Recognition in Telugu Speech using Artificial Neural Network," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 7, no. 8, pp. 615-620, 2017.

[20] V. Gunturi and P. R. Gudem, "Emotion Recognition from Telugu Speech using Mel-Frequency Cepstral Coefficients and Support Vector Machines," International Journal of Innovative Research in Science, Engineering and Technology, vol. 7, no. 1, pp. 239-246, 2018.

[21] R. Anusha, P. Subhashini, D. Jyothi, P. Harshitha, J. Sushma and N. Mukesh, "Speech Emotion Recognition using Machine Learning," 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2021, pp. 1608-1612, doi: 10.1109/ICOEI51242.2021.9453028.