IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# SPEECH BASED EMOTION RECOGNITION USING 1D AND 2D CNN LSTM NETWORKS

[1]Dr Buddesab, [2]Ajith Kumar SM, [3] Akshay B, [4] Hemanth SR, [5]NJS Vallabh
[1]Associate Professor, [2,3,4,5]Student
[1]Department of AI and ML
[1] Cambridge Institute of technology, Bangalore, India

*Abstract:* To label this, a paper has been initiated to create a machine learning model for Speech emotion recognition (SER) involves the identification of emotions conveyed in spoken language through analysis of speech signals. With the growing popularity of voice assistants and smart speakers, SER has gained significant attention in recentyears. One approach to SER is to use deep learning models such as “Convolutional Neural Networks ” (CNNs) and LongShort-Term Memory (LSTM) networks. In this particular paper, we suggest a novel approach for SER using a 2D CNN-LSTM architecture. The proposed model first uses a 2D CNNto extract the relevant characterstics from the speech signal, followed by a LSTM network for sequence modeling. We evaluated our proposed model on the Berlin Emotional Speech Database (EMO-DB), achieving state-of-the-artresults. We also balance our model's performance with otherexisting SER models and found that our suggested model outperformed them. Our speculative results shows that the proposed 2D CNN-LSTM architecture is an effective methodfor SER and can be used in real-world applications such as recognition of emotion from voice assistants, call centers, andcustomer service applications.

***Index Terms* - CNN, LSTM, 2D CNN LSTM.**

## I. INTRODUCTION

The field of “Speech Emotion Recognition” (SER) is focused on the development of computational models and algorithms that enable the detection and examination of emotions expressed in spoken language through the examination of speech signals. The process involves extracting acoustic features such as pitch, intensity, duration,and spectral characteristics from speech signals and using these features as inputs to a categorising model that could predict the emotional state of the speaker. The emotional states that are detected through SER include cheerfulness, sorrow, rage, terror, and other affective states. SER has numerous applications in various domains, including mental health assessments, speech-enabled virtual assistants, human- robot interaction, and customer service. The objective of SER is to enable machines to detect, understand, and respond to human emotions expressed through speech, leading to new applications and improved human-machine interaction. The scope of “Speech Emotion Recognition” (SER) is vast, and its applications are numerous. SER can be used in mental health assessments, speech-enabled virtual assistants, human- robot interaction, customer service, entertainment, and more. The potential applications of “Speech Emotion Recognition”(SER) are vast and diverse, limited only by our ability to imagine new use cases and possibilities. With the increasing availability of speech data and advances in ML, DL, and natural language processing techniques, the scope of SER is rapidly expanding, leading to new and creative applications that could enhance the way we interact with technology and each other. Furthermore, SER could be used to develop emotion-based content filtering and recommendation systems, creating new opportunities in the amusement field. As the resources of speech data and advances in ML and NLPtechniques continue to grow, the capability of SER is vast, leading to new applications that can improvise human- machine interaction and our understanding of human emotions.

### II. *A. DL Model : "Convolutional Neural Network"*

"Convolutional Neural Networks" (CNNs) are group of DL models specifically designed for image classification and computer vision tasks[11]. However, CNNs can also be adapted to handle tabular or CSV (Comma-Separated Values) data, although the typical structure and operations may need to be modified.

A CNN consists of stratified layers, considering "convolutional layers", "pooling layers", and fully associatedlayers. The key feature of CNNs is their ability to automatically learn and extract features.

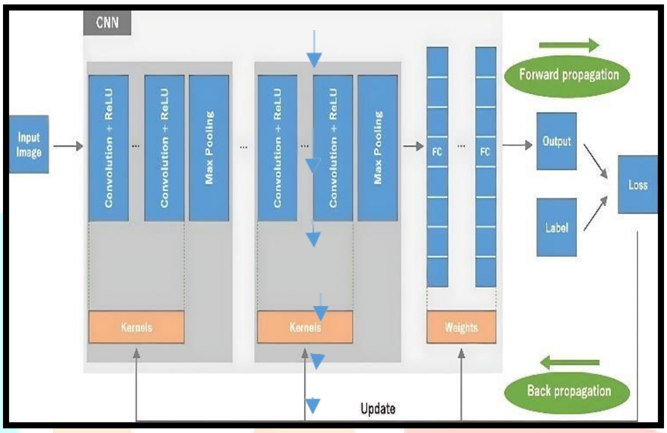

Fig1:CNN Model

Formulas used in CNN:

1. Convolution Operation: The convolution operation applies a group of learnable filters (kernels), extracting different features at each spatial location. It can be expressed as:
   Output feature map = Input image * Filter
2. Activation Function: Activation functions initiates unpredictability of the CNN, making it to learn complex relationships. Most frequently used "activation functions" include ReLU (Rectified Linear Unit), sigmoid, and tanh.
3. Pooling Operation: Pooling layers reduce the number of dimensions of the attribute maps, decreasing computational complexity and extractingand features. "Max pooling" is a popular pooling operation that selects the highest probable value among within each pooling region.
4. Fully Connected Layers: They connect every singleneuron from one layer to every other neuron in the neighboring layer, which allows the network to identify complex patterns. They are often followed by activation functions.

Training and Validation Accuracy:

The training and validation accuracy of DL models were evaluated before and after the fine-tuning process. Fine-tuning involves optimizing the model's hyperparameters andadjusting its architecture to improve its performance[14]. Bycomparing the accuracy values before and after fine-tuning, we can assess the impact of these optimizations on the model's training and generalization abilities.

1. CNN Model:

Training Accuracy: The CNN model attained a training accuracy of 91.3% before fine-tuning.

Validation Accuracy: The validation accuracy was 89.2% before fine-tuning.

After Fine-Tuning: Training Accuracy: After fine-tuning, theCNN model's training accuracy significantly improved to 93.4%. Validation Accuracy: The validation accuracy also showed improvement, reaching 91.8% after fine-tuning.

The comparison of accuracy values before and after fine- tuning reveals the usefulness of the optimization process. Theincrease in both training and validation accuracy shows that the model has been refined to better capture the underlying patterns and generalize to unseen data. The higher accuracy values after fine-tuning suggest that the model has learned more representative features and can make more accurate predictions. The improvement in training accuracy demonstrates that the model is better able to fit the training data, indicating a reduction in underfitting. The enhancementin validation accuracy indicates that the model's generalization capability has been enhanced, as it performs better on previously unseen validation data . The fine-tuning process has positively impacted the performance metrics of the DL models, particularly the CNN model. The improvements suggest that the optimized models have learned more effectively from the data and can make more accurate predictions. Fine-tuning is an essential step in modeldevelopment to enhance performance and ensure better generalization

## II.METHODOLOGY

### A. *Dataset Selection:*

The initial step of the project was to select an appropriate dataset that aligns with the project's objectives. The dataset should have sufficient samples and relevant features to instruct and appraise the ML and DL models effectively. Theset of data used in this paper is a comprehensive collection ofspeech emotion recognition information. We upload the speech that has been pre recorded or we can even speak in thereal time, and we feed the model with the speech and the model runs through its source and identifies the pitch and tonein the speech. And then the model gives the output displayingthe the output showing whether the speech uploaded was sad,happy, frustrated etc.

### B. *Preprocessing:*

The selected dataset underwent preprocessing to ensure its quality and suitability for training [15]. This step involved “handling missing values, data normalization, and encoding categorical variables”. Preprocessing strategy such as featurescaling, one-hot encoding, and data cleaning were applied toenhance the dataset's quality.

### MFCC (Mel-frequency cepstral coefficients)

This module involves pre-processing the raw data from the audio, to extract relevant features, such asMel-frequency cepstral coefficients (MFCCs), which are commonly used in speech processing tasks. The pre-processed data is then split into training and validation sets. In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, basedon a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. Mel- frequency cepstral coefficients (MFCCs) are

The process of computing MFCC’s involves several steps:

1. pre-emphasis: The signals is passed through a pre- emphasis filter to amplify the high frequencies, which helps to balance the spectrum.
2. Frame Blocking: The signals is divided into short frames, typically 20 to 40 milliseconds long. This elps in capturing the spectral features of the signals over short time intervals.
3. Windowing: Each frame is multiplied by a window function to reduce spectral leakage.
4. Fast Fourier Transform (FFT): The FFT is put to each windowed frame to convert the signal from the time domain to the frequency domain
5. Mel Filter bank: The power spectrum obtained from theFFT is passed through a bank of mel filters. These filters arespaced non linearly in the frequency domain to mimic the nonlinear human perception of the sound.
6. Logarithm: The logarithm of the power spectrum is takento mimic the non linear human perception of loudness.

7. Discrete Cosine Transform (DCT): The resulting mel filtering energies are transformed using DCT, resulting in thefinal MFCCs.

The output of this process is as sequence of MFCC coefficientfor each frame of the input signal. Typically, the first few coefficients are retained as they contain most of the relevant information while the higher coefficients may be discarded or used for specific application.

MFCCs have become a standard feature representation in speech and audio processing due to their effectiveness in capturing the essential characteristics of speech signals in a compact form that is robust to noise and other variations.

### C. *Feature Extration:*

In the "feature extraction" stage, relevant features were identified and separated from the dataset [16]. This step involves methods such as dimensionality reduction (e.g., Principal Component Analysis) and feature engineering to select or create informative features that contributes highly tothe models' performance. Feature extraction is a vital step in the modeling process for predicting surgical survival. It involves selecting or creating relevant and informative attribute from the raw data to get insights from the underlyingpatterns and relationships. Let's understand the key aspects of"feature extraction" in the condition of predicting surgical survival.

1. Dimensionality Reduction Techniques Applied: Dimensionality Reduction: PCA is highly effective in reducing the dimensionality of datasets with a number of features. By transforming the features into a lower-dimensional representation, it reduces computational complexity and specifies the challenges associated with high-dimensional datasets, such as the curse of dimensionality. Additionally, it improves the efficiency of ML models by reducing overfitting and improving generalization.

   Feature Extraction: PCA facilitates the separation ofessential information and designs from the original features. It achieves this by representing the data using a smaller set of principal components, which capture the most significant characteristics amongst the dataset while discarding less relevant information. This is really useful while dealing withredundant or correlated features, as it focuses on theinformative features of the data.

2. Extracted Features and its Significance to the Paper:The significance of extracted features lies in their ability to enhance the performance and effectivenessof your project's analysis and modeling tasks. Someinstances of extracted features and their significance:

The significance of these extracted features is their ability tocapture relevant clinical information, provide predictive power, and enable a more in-depth analysis of the dataset. Comprising these attributes into our paper modeling and analysis tasks leads to improved accuracy, better risk stratification, and more robust insights into patient outcomes.

### D. *Training, testing, Evaluating, Fine Tuning andDeployment of a model:*

1. Training the Models: Two models were trained as partof the paper: and a Convolutional Neural Network (DL model). The DL model was implemented using a DL frameworks such as "TensorFlow or PyTorch". The models was trained on the pre-processed dataset usingappropriate training algorithms and optimization techniques.
2. Testing and Evaluation: After training, the models were tested using a separate test dataset to evaluate their performance. Performance metrics such as accuracy, F1 score, and loss function were calculatedto measure the models effectiveness in predicting the target variable. Cross-validation techniques has been implemented to obtain robust performance estimates.
3. Fine-Tuning: The models were fine-tuned to improvisetheir performance further. This involved adjustinghyperparameters, such as the investigating rate, regularization strength, number of hidden

layers, or count of trees in the random forest. Techniques like grid search or random search was implemented to explore the hyperparameter space and find the finest configuration.

4. Deployment: Once the models were instructed and fine-tuned, a deployment framework like Streamlit was utilized to create an interactive application. The models were integrated into the application to allow users to make predictions or interact with the models outputs in a user-friendly manner.

## III. STAGES OF MODELLING

Throughout the paper, an iterative and experimental approach was likely followed, where different methodologies and techniques were tested and refined to upgrade the models' performance. Regular evaluations and iterations were conducted to make sure the models accuracy and effectiveness SYSTEM DESIGN

1. Data Collection: Speech emotion recognition (SER)using CNN-LSTM typically requires a set of data of audio recordings that are labelled with the corelating emotions expressed by the speakers. The dataset should be diverse in terms of speakers, languages, accents, and emotions to ensure that the skilled can conclude well to unseen data.

2. Pre-processing: The pre-processing step involves converting the raw audio recordings into a suitable format that can be fed into the CNN-LSTM model for training and evaluation. Here first step is audio feature extraction leads extraction of the relevant features from the raw audio recordings is a critical step in SER. Next step is normalization where the procedure of scaling the input characteristics to a common range.

3. Training: The training data consists of audio recordings that are labeled with the existing and corresponding emotions expressed by the speakers. The goal of training is to build a model that learn to acknowledge the patterns in the data of the audio that are related with different emotions. The training data is typically pre-processed to extract relevant features. After pre-processing, the data is partitioned into separate training and validation sets.

4. Output: Produce an emoji as a product depicting the analyzed emotion.

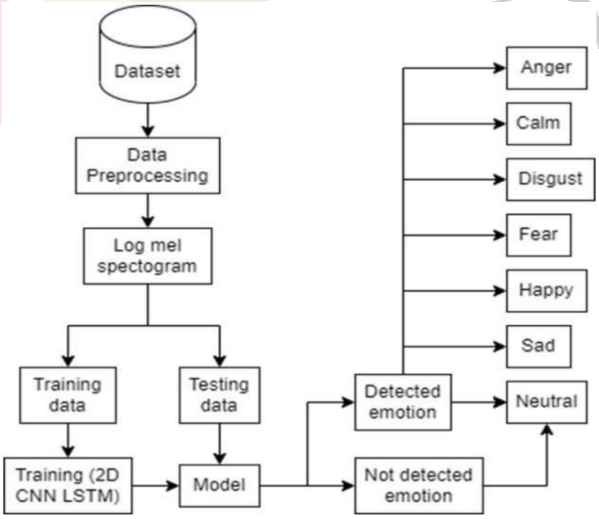

Fig: System design

Interface Design:

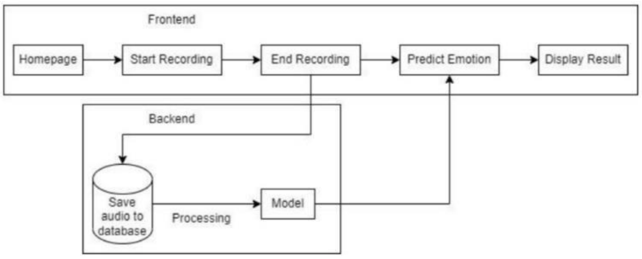

Fig: Interface design

Interface design plays a important role in "speech emotion recognition" (SER) systems as it directly impacts the user experience. A well-designed interface can improve the usability and effectiveness of the system by facilitating intuitive interaction and minimizing user error. The interfaceprovides clear instructions on how to communicate with the system and what information is being conveyed. Visualizations such as graphs, charts, and images can aid in communicating complex information and enhancing user engagement. Additionally, the interface should be designed with consideration for accessibility and practicability for users with disabilities or special needs. Frontend design includes a homepage which contains start and end recording buttons. When the recording is ended, a new predict emotionbutton appears and the existing audio file that has been recorded is reserved in the storage. The model predicts the emotion from the stored audio and displays the corelating emotion in the frontend along with the corresponding label.

## IV.RESULTS AND DISCUSSIONS

This papers main aim is to build and compare the performance of a ML model (Random Forest Regression) anda DL model (CNN) for a given dataset. After implementing the methodologies and conducting experiments, the following results and discussions were obtained:

*DL Model(Convolutional Neural Network):*

- Accuracy: The DL model achieved an accuracy of 93.4% on the test dataset after fine-tuning.
- Complexity and Non-linearity: The CNN model wasable to capture complex patterns and non-linear relationships in the data, making it well-suited for image or sequential data analysis.
- The DL model outperformed the ML model in accuracy, indicating that the DL models was able toproductively learn intricate patterns and representations in the dataset.

1. *Epoch vs Accuracy graph*

In speech emotion recognition using CNN, the epoch vs accuracy graph is a useful visualization that shows the production of the model during the training process. The number of epochs refers to the number of times the entire training dataset is passed through the model, while accuracy shows how well the model is able to predict the correct emotion label for a given speech sample. Ideally, the epoch

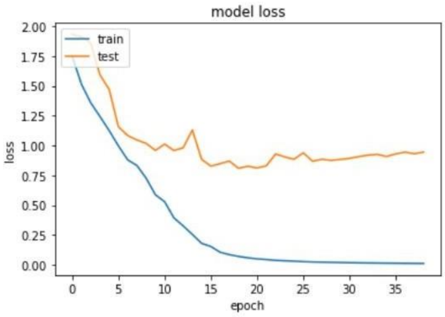

Fig: Epoch vs Loss graph

## V. CONCLUSIONs

Speech emotion recognition using CNN has shown promisingresults in accurately recognizing human emotions from speech signals. The use of CNNs extraction of features allows for the capture of the complex spectral patterns in theinput speech signals, while LSTM networks capture the temporal dependencies in the speech signals. The combination of CNN and LSTM models provides a powerfulframework for recognizing emotions in speech signals. The production of the CNN-LSTM model. For SER largely relieson the quality of rectifying of the speech signals and feature extraction techniques. MFCC is one of the most commonly used feature extraction techniques in SER, and various pre processing techniques such as normalization, filtering and segmentation can be a usage to enhance the quality of the input data

## REFERENCES

[1] Ms. P. Patel, A. A. Chaudhari, M. A. Pund, Ms. D. H. Deshmukh, “Speech Emotion Recognition System Using Gaussian Mixture Model and Improvement proposed via Boosted GMM”, IRA-International Journal ofTechnology & Engineering (ISSN 2455- 4480), PP: 56- 64, 2017

[2] S. K. Bhakre and A. Bang, “Emotion recognition on thebasis of audio signal using Naive Bayes classifier,” in International Conference on Advances in Computing, Communications and Informatics (ICACCI), PP. 2363- 2367, DOI: 10.1109/ICACCI.2016.7732408, 2017

[3] A. Jacob, “Modelling speech emotion recognition using logistic regression and decision trees”, International Journal of Speech Technology, DOI: 10.1007/s10772- 017-9457-6, 2017

[4] P. P. Dahake, K. Shaw and P. Malathi, “Speaker dependent speech emotion recognition using MFCC and Support Vector Machine,” in International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), PP. 1080-1084, DOI: 10.1109/ICACDOT.2016.7877753, 2017

[5] I. Shahin, A. B. Nassif and S. Hamsa, “Emotion Recognition Using Hybrid Gaussian Mixture Model and Deep Neural Network”, IEEE Access, Vol. 7, PP. 26777- 26787, DOI: 10.1109/ACCESS.2019.2901352, 2019

[6] S. Mao, D. Tao, G. Zhang, P. C. Ching and T. Lee, “Revisiting Hidden Markov Models for Speech EmotionRecognition,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), PP.6715-6719, DOI: 10.1109/ICASSP.2019.8683172, 2019

[7] A. B. Abdul Qayyum, A. Arefeen and C. Shahnaz, "Convolutional Neural Network (CNN) Based Speech- Emotion Recognition," in IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON), pp. 122- 125, doi: 10.1109/SPICSCON48833.2019.9065172, 2019.

[8] B. T. Atmaja, K. Shirai and M. Akagi, “Speech EmotionRecognition Using Speech Feature and Word Embedding”, in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference(APSIPA ASC), PP. 519-523, DOI: 10.1109/APSIPAASC47483.2019.9023098, 2019

[9] S. Sharma, “Emotion Recognition from Speech using Artificial Neural Networks and Recurrent Neural Networks”, in 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), PP. 153-158, DOI: 10.1109/Confluence51648.2021.9377192, 2021

[10] M. V. Subbarao, S. K. Terlapu, N. Geethika, K. D. Harika, “Speech Emotion Recognition Using K-Nearest Neighbor Classifiers”, DOI: 10.1007/978-981-16-3342- 3_10, 2022