Speech Emotion Recognition with Combination of CNN-LSTM and DNN

Ponna Dinesh, Suman Deb
Department of Electronics Engineering
SVNIT, Surat, India
[u20ec009,sumandeb]@eced.svnit.ac.in

***Abstract*—Emotion is essential to all living things. Understanding emotion is challenging for everyone, but it can solve thousands of problems and save many lives if done correctly. Emotion is represented not only in gestures but also in work and the ability to produce effective results. As a result, throughout the past three decades, numerous researchers have been interested in emotion recognition through speech. In this study, I address the crucial task of speech emotion recognition (SER) by leveraging deep learning architectures, namely Convolutional Neural Networks (CNN) combined with Long Short-Term Memory (LSTM) layers and Deep Neural Networks (DNN). My approach utilises a diverse feature set, including Mel spectrogram, MFCCs, spectrogram, delta of MFCCs, chroma feature, and tonal centroid features extracted from the EmoDB dataset. Through extensive experimentation employing a 5-fold cross-validation methodology, I achieved notable classification accuracies of 77.38% with the CNN+LSTM model and 87.09% with the DNN model. Additionally, I investigated the impact of different hyperparameters and training strategies on the performance of the CNN+LSTM and DNN models, striving to optimise their accuracy and robustness in real-world applications. The experimental results revealed insights into the optimal configuration of these models, shedding light on potential avenues for further improvement. Moreover, I explored the generalisation capability of the trained models by evaluating their performance on unseen data from diverse emotional speech datasets, thus ensuring the reliability and applicability of the proposed approach across different contexts. This comprehensive analysis enhances our understanding of SER methodologies and provides valuable guidance for future research endeavours to push the boundaries of emotion recognition technology.**

***Index Terms*—Speech Emotion Recognition, Emotion Recognition, Feature Extraction, Convolutional Neural Network, Long Short-Term Memory, Deep Neural Networks, and 5-Fold cross-validation.**

I. Introduction

Speech is one of the most fundamental and intuitive forms of human communication, conveying information and expressing various emotions. Emotions are pivotal in interpersonal interactions, influencing behaviour, decision-making, and social relationships. Therefore, recognising these emotions accurately is crucial for enhancing human-computer interaction (HCI). Speech Emotion Recognition (SER) seeks to automate the identification of emotional states from spoken language, using various acoustic and linguistic cues inherent in the human voice [1]. With the advent of sophisticated algorithms and computational models, SER has gained significant attention, promising transformative impacts across various sectors. In customer service, for instance, SER can detect dissatisfaction or stress in voices, allowing for more responsive and tailored service. In healthcare, monitoring emotional cues in speech can assist in diagnosing and treating conditions such as depression and anxiety. Moreover, understanding learners' emotional state in educational technologies can help adapt content delivery to improve engagement and learning outcomes.

Improving human-computer interaction (HCI) through technology, skill development, deep learning, and artificial intelligence techniques is necessary to improve HCI, including emotion recognition. SER is a valuable area of HCI with various real-time applications, including virtual reality, assessing user happiness in contact centres, identifying human emotion in human reboot interactions, and assessing a user's emotional state to determine the proper response in emergency call centres. The SER for the automobile board system is designed to identify drivers' emotional or mental states so that necessary safety measures can be taken for the passengers. Moreover, SER is utilised in automatic translation systems and detecting violent and destructive behaviour in crowds that cannot be completed by hand [2]. SER for intelligent, efficient services to offer methods for showcasing the efficiency of the SER for intelligent healthcare facilities using CNN architectures with rectangular shape filters [3]. The effectiveness of SER for real-time applications that use discriminative and salient features analysis to extract features to increase the significance of HCI.

Dario Bertero, Farhad Bin Siddique, Chien-Sheng Wu, Yan Wan, Ricky Ho Yin Chan, and Pascale Fung described their approach of enabling an interactive dialogue system to recognise user emotion and sentiment in real time. These modules enable ordinarily standard dialogue systems to show "empathy" and respond to the user while being aware of their emotion and intent. Emotion identification from speech was formerly achieved by feature engineering and machine learning, with the first stage causing a delay in decoding time. They described a CNN model that extracts emotion from raw speech input without using feature engineering. This approach outperforms traditional feature-based SVM classification, with an average accuracy of 65.7% across six emotion categories, a 4.5% increase. A separate, CNN-based sentiment analysis module distinguishes sentiments from speech recognition results, with an F-measure of 82.5 on human-machine dialogues when

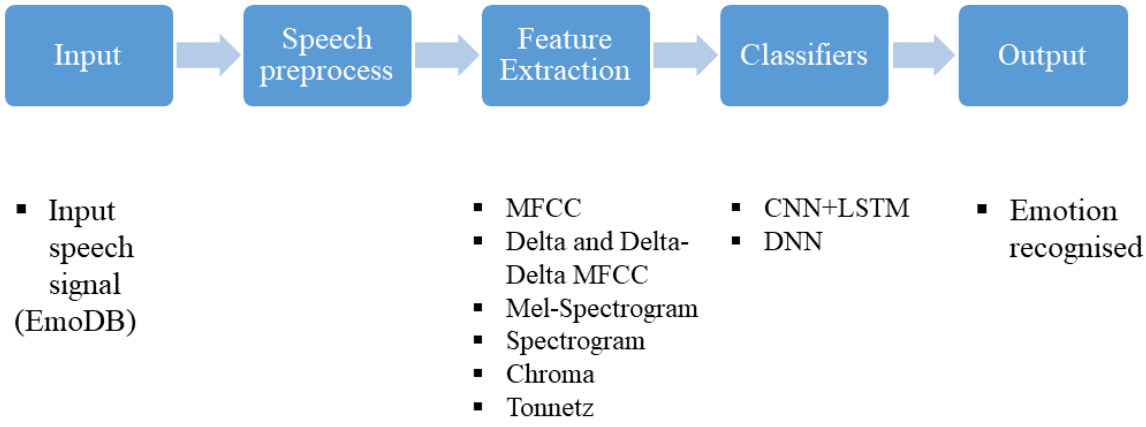

Fig. 1: Block diagram of Speech Emotion Recognition

trained using out-of-domain data [4]. Ashima Yadav and Dinesh Kumar Vishwakarma have proposed A Multilingual Framework of CNN and Bi-LSTM for Emotion Classification and suggested a language-independent deep learning framework for voice emotion classification. They developed a novel mix of 1D CNN and Bi-LSTM units to extract MFCC-based and deep high-level features. The proposed system uses CNN to extract local information from signals, while the Bi-LSTM layer models the signal's long-term contextual dependencies. They verify their suggested architecture against two multilingual datasets, EmoDB for German and RAVDESS for English [5]. S. Kim et al. proposed a CNN-based SER model that takes the mel-spectrogram as input and employs multiple convolutional and pooling layers and fully connected layers for emotion classification. They achieved competitive performance on benchmark datasets such as the IEMOCAP and SAVEE datasets [2]. Fushiki's paper on estimating prediction error using k-fold cross-validation provides valuable insights and practical guidance for researchers and practitioners in machine learning, statistics, and data mining. The proposed technique has since become widely adopted for assessing model performance and guiding model selection in various application domains [6]. Mujaddidurrahman et al. [7] proposed this literature on SER. The study implemented a 2D-CNN model with Log-Mel spectrogram inputs for audio emotion recognition, demonstrating superior performance to existing deep learning models. Evaluation of the EMODB database showcased the model's efficacy in accurately capturing diverse emotional states from speech data, addressing the complexity of differentiating emotions in human-machine interaction. Aida-Zade, Xocayev, and Rustamov [8]proposed on Speech recognition using Support Vector Machines. The study focused on building an acoustic model for Azerbaijani speech recognition, utilising Support Vector Machines (SVM) with MFCC and LPC features. Through experimentation, SVM with polynomial and radial basis kernels outperformed a Multilayer Artificial Neural Network, emphasising the efficacy of SVM techniques in this context. Hossain and Muhammad proposed an emotion identification system based on deep learning, utilising audio-visual big data. They employed Mel-spectrograms for speech signals fed into a 2D CNN followed by ELMs, and for video signals, they utilised a 3D CNN also followed by ELMs. Finally, SVM was used for emotion categorisation, integrating outputs from speech and video processing streams [9]. Mao et al. [3] proposed a Convolutional Neural Network (CNN) approach for speech emotion recognition, focusing on learning affect-salient features. The CNN learns simple features in lower layers and affect-salient, discriminative features in higher layers through two separate learning phases. They employ a sparse auto-encoder (SAE) for local invariant feature learning and salient discriminative feature analysis (SDFA) to identify affect-salient features, introducing a novel objective function for SER emphasising feature saliency, orthogonality, and discrimination. From this Literature review, we got to know that there are several techniques, mainly SVM with an accuracy of 44.4% with the Berlin database, CNN with an accuracy of 75.94% with the Emo-DB dataset, 82.31% for eight emotions and 79.42% for five emotions on the IEMOCAP and Emo-DB datasets, respectively, using a CNN-Transformer architecture for capturing spatial and sequential features,82% on the RAVDESS dataset using a combined CNN-LSTM architecture and 74% using a CNN-Transformer encoder architecture.

The rest of the paper is organised as follows: Section II explains the datasets used in our work for SER. Section III describes the methodology (pre-processing, feature extraction, proposed model, and experimental setup) used to classify the emotion. Results, discussion, and comparison of our method with others are discussed in Section IV. The conclusion is explained in Section V.

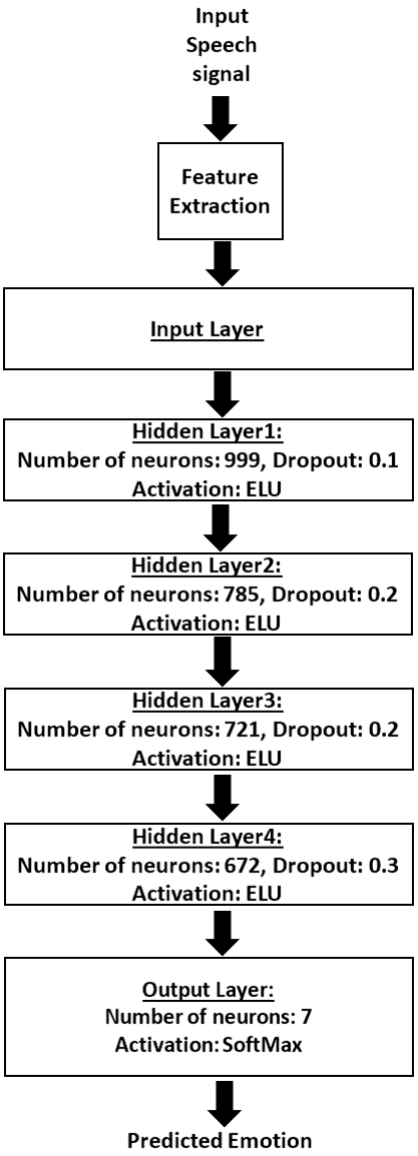

Fig. 2: Proposed DNN model for SER

II. Database

The EmoDB dataset, also known as the Berlin Database of Emotional Speech, is a freely available German emotional database1. It was created by the Institute of Communication Science at the Technical University in Berlin, Germany1. The dataset comprises 535 utterances recorded by ten professional speakers, five males and five females. The EmoDB dataset encompasses seven emotions: anger, boredom, anxiety, happiness, sadness, disgust, and neutral [10]. The data was initially recorded at a 48-kHz sampling rate and then downsampled to 16-kHz1 [10]. This dataset is commonly used for classification problems related to emotional speech [11] and has been leveraged in various research studies for tasks such as speech emotion recognition [10]. Each utterance is given a name based on the same system. For instance, the audio file 03a01Fa.wav contains speaker 03's speech of text a01 with the emotion "Freude" (Happiness).

III. Methodology

In this chapter, we are going to discuss the methodology of Speech Emotion Recognition. Firstly, we are going to discuss speech processing, for this speech processing, we used a data set of the different audio signals that are recorded with professional speakers [12]. Begins by importing necessary libraries and defining functions for energy and RMS calculations. It then loads the EmoDB dataset, a collection of German emotional speech, and normalizes the audio signals. Several features, including Mel spectrogram, MFCCs, spectrogram, delta of MFCCs, chroma feature, tonal centroid features (tonnetz), spectral contrast, and RMS energy, are extracted from these signals. These features, commonly used in speech and audio processing, are appended to a list along with the filename and emotion label. Finally, this data is written into a CSV file, which can be used as input for machine learning models for emotion classification. The effectiveness of these features can vary depending on the task and dataset, so experimentation and normalisation are recommended. Using a Deep Neural Network (DNN) model to classify emotions based on the EmoDB dataset. The dataset is divided into training and testing sets, and the unique values of the target variable are identified. The class weights are computed to handle class imbalance. The features are then standardized using the StandardScaler. The DNN model is built with four layers, each followed by batch normalization and dropout for regularization. The model is compiled with the Adam optimizer, sparse categorical cross entropy as the loss function, and sparse categorical accuracy as the metric. The model's initial weights are saved and then loaded back to ensure reproducibility. The model is trained for 1200 epochs with a batch size of 64, using early stopping and model checkpoint callbacks. The best model is saved based on the validation sparse categorical accuracy. uses a Convolutional Neural Network (CNN) combined with Long Short-Term Memory (LSTM) layers to classify emotions based on the EmoDB dataset. The dataset is divided into training and testing sets, and the unique values of the target variable are identified. The class weights are computed to handle class imbalance. The features are then standardized using the StandardScaler. The model is built with several layers: four Conv1D layers each followed by MaxPooling1D, BatchNormalization, and Dropout for regularization, two LSTM layers, and three Dense layers. The model is compiled with the Adam optimizer, sparse categorical cross entropy as the loss function, and sparse categorical accuracy as the metric. The model's initial weights are saved and then returned to ensure reproducibility. The model is trained for 900 epochs with a batch size 64, using early stopping and model checkpoint callbacks. The best model is saved based on the validation sparse categorical accuracy. In the current project, we used the 5-fold cross-validation strategy.

A. Pre-Processing

The preprocessing stage involves several steps. First, the audio files are loaded using the Librosa library. The audio signals are normalised to ensure they fall within a standard range. This is crucial as it helps reduce the dataset's variance and makes the features more comparable. Here, I used the Scipy library for signal processing.

B. Feature Extraction

Now, the audio sample feature is extracted using the Librosa library. 'Librosa' is the Python package for audio analysis and music analysis. It provides 'building blocks that are necessary to create information of music of retrieval systems. It is a Python package for music analysis. It is used in Python to extract features from audio files and process them. It starts with loading each audio file and normalising the signal. This normalised signal extracts a variety of features that are commonly used in audio processing. These include the Mel spectrogram, MFCCs, and signal spectrogram. It also calculates the delta and delta-delta MFCCs, which capture the change in MFCCs over time. Other features extracted include chroma (relating to the 12 different pitch classes), tonnetz (harmonic relations between different pitches), spectral contrast (difference in amplitude between peaks and valleys in a sound spectrum), and RMS energy. These features are then averaged over time to create a single feature vector for each audio file. This feature vector and the label for the emotion present in the audio file are then written to a CSV file. This results in a preprocessed dataset where each instance corresponds to an audio file and consists of the extracted features and the emotion label.

C. Cross Validation

Cross-validation is useful for assessing a statistical model's efficiency. A training and validation set are the two subsets created by splitting the available data into them. The validation set is used to evaluate the model's performance after it has been trained on the training set. This procedure guarantees a comprehensive assessment. The typical techniques for cross-validation are :

1) Validation Set Methodology
2) Cross-validation with leave-P-out
3) Do not include one cross-validation
4) Cross-validation in K-fold
5) K-fold stratified cross-validation

1) K-fold cross-validation: One method that is frequently used to assess how well prediction models work is K-fold cross-validation. The training data is divided into k folds or equal-sized chunks. The model is trained on k-1 folds for each iteration, and its performance is assessed on the remaining fold. To ensure that every fold functions as training and testing data, this step is performed k times. When analysing the model's generalizability, the average performance over all folds offers a more thorough evaluation than testing it on a single holdout set.

The steps for k-fold cross-validation are:

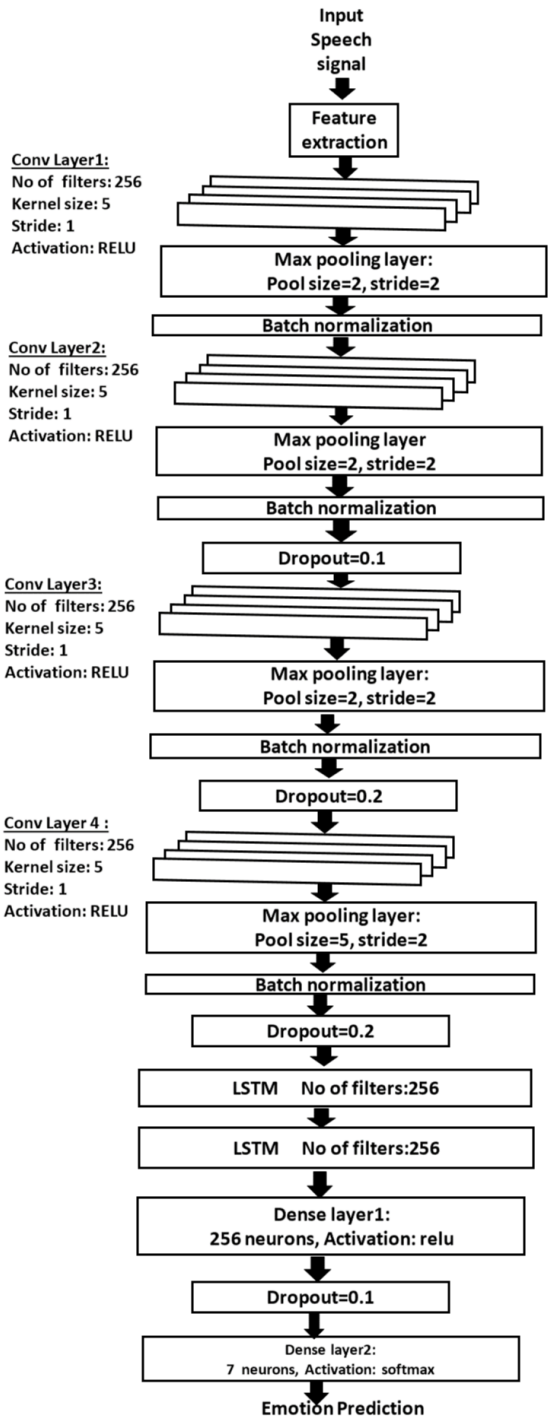

Fig. 3: Proposed CNN+LSTM model for SER

- The input dataset is created into k groups
- Among the k groups, one group is selected as the test dataset
- Rest of the groups are considered as the training dataset
- Now the model is fit with the training dataset, and the

performance of the model is evaluated on the test dataset

For example, consider the 5-fold cross-validation:

- Five folds are used to divide the dataset. The first fold in the first iteration is for the test data; the remaining folds are for training. The test data for the second iteration comes from the second fold, and the remaining folds are used for training. All of the folds will go through this procedure once more.

D. Proposed CNN+LSTM And DNN Classifier

In our study, we used a CNN+LSTM and DNN model to analyse emotion classification performance using the speech signal. The block diagram of the proposed CNN+LSTM and DNN classifier is shown in Fig.**??** and Fig.**??**, respectively. The input to the proposed Deep Neural Network (DNN) model is a feature vector equal to 260. The proposed DNN model includes multiple dense layers. The first dense layer contains 999 neurons, followed by a Batch Normalization layer and a Dropout layer of 0.1. The subsequent dense layers contain 785, 865, and 672 neurons, each followed by a Batch Normalization layer and a Dropout layer with rates of 0.2, 0.2, and 0.3, respectively. The output of the final Dropout layer is given to the output layer, which contains seven neurons corresponding to the seven different emotions in the dataset. All the layers except the output layer use the Exponential Linear Unit (ELU) activation function. The SoftMax activation function is used in the output layer to find the probability score of each emotion and to predict the emotions of the speech signal.

The input to the proposed Deep Neural Network (DNN) model is a feature vector of size 260 extracted from the EmoDB dataset. The proposed DNN model includes multiple dense layers. The first dense layer contains 999 neurons and uses the Exponential Linear Unit (ELU) activation function, followed by a Batch Normalization layer and a Dropout layer with a rate of 0.1. The subsequent dense layers contain 785, 865, and 672 neurons, each followed by a Batch Normalization layer and a Dropout layer with rates of 0.2, 0.2, and 0.3, respectively. The output of the final Dropout layer is given to the output layer, which contains seven neurons corresponding to the seven different emotions in the dataset. All the layers except the output layer use the Exponential Linear Unit (ELU) activation function. The SoftMax activation function is used in the output layer to find the probability score of each emotion and to predict the emotions of the speech signal. This architecture allows the model to learn complex patterns in the high-dimensional input data.

E. Experimental Setup

In our work, we first obtained the feature extracted CSV file of all the speech files, which are of '.wav' format, by extracting the Mel spectrogram, MFCCs, spectrogram, delta MFCCs, chroma, tonnes, spectral contrast, and RMS energy feature of each speech file by using the 'Librosa' library. From the Extracted CSV file, replace categorical labels with numerical ones and separate the features and target variables. The data is then split into training and validation sets using stratified 5-fold cross-validation, ensuring each fold has the same proportion of each label. Each fold's training and validation datasets are saved as separate CSV files. The model is trained on features extracted from audio files. The class weights are computed to handle class imbalance during model training. The features are then standardised using StandardScaler from sklearn. The model is compiled with the Adam optimiser, sparse categorical cross-entropy loss, and sparse categorical accuracy metric. The model's initial weights are saved and then returned to ensure reproducibility. The model is trained using the fit method with a batch size and epochs. During training, the model's best weights are saved whenever there is an improvement in validation accuracy, and training is stopped early if the validation accuracy does not improve for 50 consecutive epochs. The model is trained with class weights to handle class imbalance. The training history is stored for further analysis.

IV. Results and Discussion

We found the validation accuracy using a five-fold strategy shown in Table I. From the accuracy plots of five-fold, it was observed that the highest validation accuracy was 81% and the lowest testing accuracy was 73%. Using this 5-fold cross-validation technique, we obtained an average test accuracy of 77% by using the CNN+LSTM model. It was observed that the highest validation accuracy of 90% and the lowest testing accuracy of 84%. Using this 5-fold cross-validation technique, we obtained an average test accuracy of 87% by using the DNN model.

TABLE I: Accuracy of Five-Fold for EmoDB dataset

Fold	**Validation Accuracy in %**	
	CNN+LSTM	**DNN**
Fold-1	76.63	90.65
Fold-2	75.70	85.05
Fold-3	73.83	84.98
Fold-4	81.31	85.98
Fold-5	79.44	88.79

TABLE II: Classification report for CNN+LSTM model using Emodb dataset

	precision	**recall**	**f1-score**	**support**
anger	0.80	0.84	0.82	19
boredom	0.71	0.83	0.77	18
disgust	0.80	0.73	0.76	11
fear	0.80	0.86	0.83	14
happiness	0.71	0.62	0.67	16
neutral	0.73	0.69	0.71	16
sadness	0.92	0.85	0.89	13
accuracy			0.78	107
macro avg	0.78	0.77	0.78	107
weighted avg	0.78	0.78	0.77	107
Average classification accuracy=77.57%				

The Classification report for the EMO-DB dataset using the proposed features and using two models is shown in Tables

TABLE III: Classification report for DNN model using Emodb dataset

	precision	recall	f1-score	support
anger	0.94	0.79	0.86	19
boredom	0.93	0.78	0.85	18
disgust	0.91	0.91	0.91	11
fear	0.82	1.00	0.90	14
happiness	0.80	0.75	0.77	16
neutral	0.79	0.94	0.86	16
sadness	0.86	0.92	0.89	13
accuracy			0.86	107
macro avg	0.86	0.87	0.86	107
weighted avg	0.87	0.86	0.86	107
Average classification accuracy=86%				

II and III. From the CM, it was observed that emotions like $anger$, $disgust$, and $sadness$ are recognized with the highest recognition rate of 80%, 80%, and 92%, respectively, and emotions like $angry$, $boredom$, and $disgust$ were recognized with the highest recognition rate of 94%, 93%, and 91% respectively.

A. Comparison with the state-of-the-art models

In this section, I compare my results with results from state-of-the-art methods. Table IV compares my performance with the state-of-the-art methods for the EmoDB dataset.

TABLE IV: PERFORMANCE COMPARISON OF SER WITH STATE-OF THE-ART METHODS ON THE EmoDB DATASET

Author	**Feature**	**Model**	**Accuracy**
Lukose et al. [13]	MFCC	GMM	76.31%
Liu et al. [14]	Formants	SVM+RBF	78.66%
Ancilin et al. [15]	MFMC	SVM	81.5%
Ozseven et al. [16]	MFCC, pitch, Formant, band width	SVM	82.8%
Proposed	**MFCC, Mel spectrogram, spectrogram, delta of MFCCs, chroma, tonnetz**	**CNN+LSTM**	**77.57%**
Proposed	**MFCC, Mel spectrogram, spectrogram, delta of MFCCs, chroma, tonnetz**	**DNN**	**86%**

V. CONCLUSION

the comparative study of the Deep Neural Network (DNN) model and the Convolutional Neural Network combined with Long Short-Term Memory (CNN+LSTM) model, using a 5-fold cross-validation technique, revealed a significant difference in performance. The DNN model achieved an impressive average accuracy of 86% across the 5 folds, outperforming the CNN+LSTM model which achieved an average accuracy of 76%. The feature extraction techniques employed, including Mel Spectrogram, MFCC, Delta MFCC, Tonnetz, Chroma, among others, were instrumental in achieving these results. The data for these models was sourced from a CSV file. This analysis underscores the effectiveness of the DNN model in this specific context, although the CNN+LSTM model also demonstrated a respectable performance. Future work could explore potential improvements to the CNN+LSTM model or investigate the impact of other feature extraction techniques.

Speech emotion recognition (SER) has gained significant traction in recent years, fueled by the transformative power of deep learning. This technology has revolutionized the way we analyze and interpret human emotional cues conveyed through speech, enabling machines to comprehend and respond to the emotional nuances of human communication. With its ability to extract intricate patterns from complex data, deep learning has propelled SER to new heights of accuracy and robustness, paving the way for a myriad of promising applications in various domains.

REFERENCES

[1] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011, sensing Emotion and Affect - Facing Realism in Speech Processing. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639311000185

[2] S. Kim, S. J. Guy, K. Hillesland, B. Zafar, A. A.-A. Gutub, and D. Manocha, "Velocity-based modeling of physical interactions in dense crowds," *The Visual Computer*, vol. 31, pp. 541–555, 2015.

[3] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE transactions on multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[4] D. Bertero, F. Siddique, C.-S. Wu, Y. Wan, R. Chan, and P. Fung, "Real-time speech emotion and sentiment recognition for interactive dialogue systems," 01 2016, pp. 1042–1047.

[5] A. Yadav and D. K. Vishwakarma, "A multilingual framework of cnn and bi-lstm for emotion classification," in *2020 11th international conference on computing, communication and networking technologies (ICCCNT)*. IEEE, 2020, pp. 1–6.

[6] T. Fushiki, "Estimation of prediction error by using k-fold cross-validation," *Statistics and Computing*, vol. 21, pp. 137–146, 2011.

[7] A. Mujaddidurrahman, F. Ernawan, A. Wibowo, E. A. Sarwoko, A. Sugiharto, and M. D. R. Wahyudi, "Speech emotion recognition using 2d-cnn with data augmentation," in *2021 International Conference on Software Engineering Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)*, 2021, pp. 685–689.

[8] K. Aida-zade, A. Xocayev, and S. Rustamov, "Speech recognition using support vector machines," in *2016 IEEE 10th international conference on application of information and communication technologies (AICT)*. IEEE, 2016, pp. 1–4.

[9] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," *Information Fusion*, vol. 49, pp. 69–78, 2019.

[10] D. H. Rudd *et al.*, "Emodb dataset (berlin database of emotional speech)," 2022. [Online]. Available: https://paperswithcode.com/dataset/emodb-dataset

[11] "Emodb dataset — kaggle," 2024. [Online]. Available: https://www.kaggle.com/datasets/piyushagni5/berlin-database-of-emotional-speech-emodb

[12] "Language-independent hyperparameter optimization based speech emotion," 2022. [Online]. Available: https://link.springer.com/article/10.1007/s41870-022-00996-9

[13] S. Lukose and S. S. Upadhya, "Music player based on emotion recognition of voice signals," in *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*. IEEE, 2017, pp. 1751–1754.

[14] Z.-T. Liu, A. Rehman, M. Wu, W.-H. Cao, and M. Hao, "Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence," *Information Sciences*, vol. 563, pp. 309–325, 2021.

[15] J. Ancilin and A. Milton, "Improved speech emotion recognition with mel frequency magnitude coefficient," *Applied Acoustics*, vol. 179, p. 108046, 2021.

[16] T. Özseven, "Investigation of the effect of spectrogram images and different texture analysis methods on speech emotion recognition," *Applied Acoustics*, vol. 142, pp. 70–77, 2018.