Comprehensive Analysis of SVD-based Speech Emotion Recognition

Divya Sharma, Koushik Reddy K, Kusalanatha Reddy C, Badri Ram G, Pavan Kumar Reddy
Department of Electronics and Communication Engineering, New Horizon College of Engineering, Bangalore
E-mail : er.divyasharma@gmail.com, koushikreddy949@gmail.com, kushalnathreddy2002@gmail.com, galijerlabadriram5978@gmail.com, ankepallipavankumar138@gmail.com

Abstract*— Over time, there has been a growing interest in the domain of Speech Emotion Recognition (SER) utilizing both Matlab and Python interfaces. SER involves analyzing input speech to discern the underlying emotional state being conveyed. This process typically comprises several steps, such as feature extraction, feature matching, classification, and database integration. By employing algorithms, relevant features are extracted from the input speech and subsequently matched using various models. These models enable the analysis of specific characteristics within the speech signal, allowing the system to identify the emotional state expressed. Emotions recognized by the system encompass anger, boredom, anxiety, disgust, happiness, neutrality, and sadness. The main objective of this report is to present a survey on the application of the Singular Value Decomposition (SVD) algorithm in combination with diverse classifiers and different speech emotion databases. There is a plethora of audio features that can be utilized for extraction, and a wide range of classifiers are available, including the Hidden Markov Model (HMM), Convolutional Neural Network (CNN), Support Vector Machine (SVM), and Long Short Term Memory (LSTM). Significantly, the report will delve into these models and explore their implementation using various speech emotion databases, such as the Berlin and Tess databases, as well as other pertinent ones.

Keywords*— Hidden Markov Model (HMM), Convolutional Neural Network (CNN), Support Vector Machine (SVM), Long Short Term Memory (LSTM), feature extraction, classification, databases.

I. INTRODUCTION

Now a days Speech Emotion Recognition(SER) has been gaining interest and is one of the topics which has been continuously researched in speech processing. In the field of Human Computer Interaction (HCI) it is the most important topic. Recent technologies like artificial intelligence, IOT and Machine learning have scaled up communication and automation industries [1][2][3][4]. It research was started from the late fifties. From late fifties it indicated that the growth of publication papers are increasing every year. It was widely used in many applications and also applied in so many fields such as security, education, human computer interaction, teaching, entertainment and so on. It illustrates the valuable insights into an individual's emotional and mental state through a groundbreaking area of study known as automatic emotion recognition. The majority researchers are concerned in SER because it is a good source for affective computing. As now SER was emerging across various fields such as artificial intelligence and also an important topic in signal processing and pattern recognition. To find the underlying state of emotion of the individual from his input speech signal is the main objective of SER system. From Fig 1 SER has three main objectives for getting successful output one is to use a good database with many utterances and second one is choice of algorithm for extraction of features and the last one is classification for feature matching. In the database, numerous speeches from various individuals encompass a diverse range of emotional states. As a result, numerous researchers have made valuable use of this categorized emotional data for their studies and analyses.

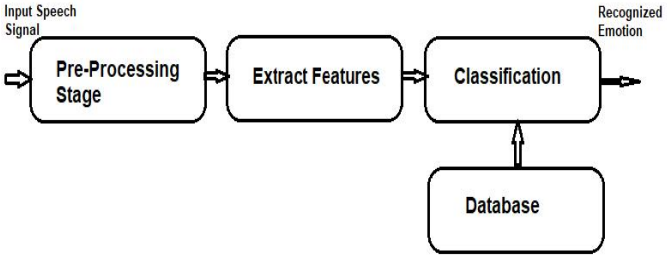

Fig.1. Speech Emotion Recognition System

By increasing the count of utterances, we can achieve greater accuracy in determining the emotional state of unknown input speech. Feature extraction is a significant goal within Speech Emotion Recognition (SER). Feature extraction involves capturing speech signal information such as pitch, frequency, and formant. In this study, the Singular Value Decomposition (SVD) algorithm is employed. Many researchers prioritize selecting the best feature sets to preserve maximum information. However, as the number of features increases, so does the dimensionality. The final objective is classification, where the raw data in the form of utterances is categorized into specific emotions. Various classification algorithms including HMM, CNN, SVM, LSTM have been proposed and utilized for SER. Utilizing both feature extraction and selection techniques improves learning performance, reduces complexity, and enhances storage efficiency.

II. LITERATURE SURVEY

Numerous studies have been conducted in the field of natural language processing, exploring its diverse applications and developing various classifiers to enhance accuracy and design systems for emotional state recognition. A review based on these have presented in [5], [6], and [7].

Numerous databases offer vast collections of audio files, among which the ravadness format stands out as a high-quality option. A review centered on this particular database has been presented in [8].

Numerous approaches have been introduced for speech emotion recognition (SER) systems, employing various techniques, with the feature extraction process being of utmost importance. Among the annual advancements in this field, one notable technique is the TEKO model, which presented in [9].

The study [10][11] introduces various models utilized in the ser system and outlines the general architecture for its implementation using these diverse models. Among the models discussed are HMM, neural networks, and other relevant approaches.

Several models exist for implementing speech emotion recognition (SER) systems and classification. Feature extraction techniques like SVD, LDA and combinations with various classifiers have been explored to enhance accuracy while using fewer acoustic features. The results and accuracies of these methods are discussed in [12].

Currently, numerous hybrid models have emerged, combining various techniques such as LSTM-CNN, LSTM-RNN, CNN-LSTM-DNN, among others, as classifiers and DNN also proposed. These hybrids have been studied for their recognition rates and accuracy using different databases [13].

This study [14] introduces various classifiers and assesses their accuracies across different databases. Specifically, the CNN_LSTM model trained on IEMOCAP achieved an accuracy of 95.89%. Meanwhile, the SVM model applied to EMOdb attained an accuracy of 95.3%, while the DNN model, also evaluated on EMOdb, reached an accuracy of 96.97%.

In the context of SER (Speech Emotion Recognition) systems, the primary objective is feature extraction, wherein data is obtained from audio files. Various algorithms are employed for this purpose. For instance, the study [15] introduces a novel approach called Unsupervised Feature Extraction Using Singular Value Decomposition (SVD).

Study [16] proposes about feature selection on ser from the speech and not only this and so many papers [17] published on speech emotion recognition based on factors that affect and also about the reduction of unwanted noise that increases the computational cost.

Study [18] proposes a study on ser in natural environment also the models that are used in that environment and their disadvantages in the context of the speaker, type of environment the speech signal is recorded.

III. PROPOSED MODEL

This section presents the proposed system. The SER flow/Block diagram of proposed model is shown in Fig. 2. The steps involved are discussed below.

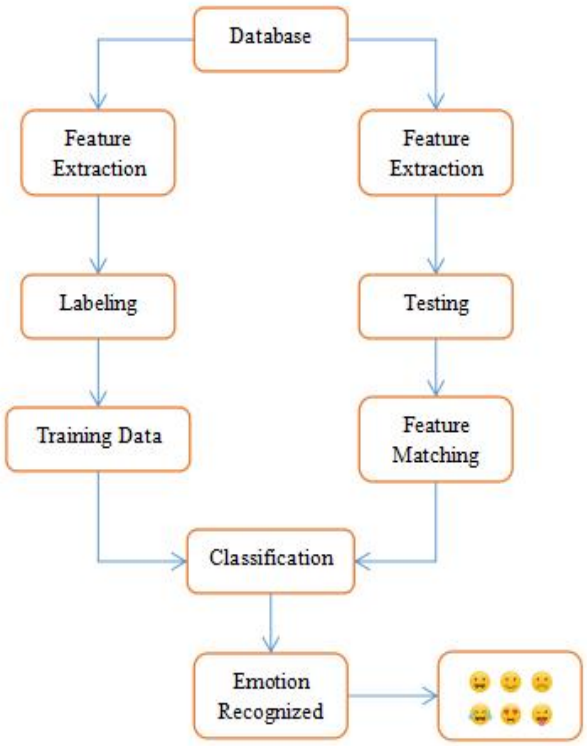

Fig.2. SER flow

A. *Selection of databases:*

In this paper, we utilized two prominent datasets, TESS (Toronto Emotional Speech Set) and the Berlin database, to conduct our design for Speech Emotion Recognition (SER). These datasets provided us with valuable raw audio data in wav file format, allowing us to analyze and study emotional speech. The TESS dataset offers a collection of professionally acted recordings, encompassing various sentences spoken in different emotional states. It covers a wide range of emotions, including anger, fear, happiness, pleasant, surprise, sadness, and neutral. The contributions from multiple speakers in TESS ensure diversity in terms of gender, age, and vocal characteristics, enhancing the dataset's representation.

Additionally, we incorporated the Berlin database, which features over 500 utterances from different speakers. This dataset includes diverse texts and emotions, with 10 distinct speakers exhibiting different genders and ages. The Berlin database, stored in wav file format, provides an extensive range of emotional expressions and serves as a valuable resource for evaluating the accuracy of our SER methodology. By leveraging these datasets, we were able to analyze and compare the performance of our SER models across different emotional states, genders, ages, and textual contexts. The aggregation of the TESS and Berlin datasets enriched our research, facilitating a comprehensive exploration of speech emotion recognition and its applications.

B. *Feature Extraction:*

We use wav file and svd algorithm used for feature extraction. At first we load audio files in the form of wav files and extract features from those files also using svd algorithm for feature extraction.

i) SVD algorithm:

Singular Value Decomposition (SVD) is a method used for decomposition of data matrix. It is a matrix factorization method expresses a data matrix as into orthogonal matrices based on singular values on its diagonal in decreasing order. To compute both singular values and vectors of original matrix

[U,S,V] = svd (original matrix)

The SVD theorem states that

Amxn =UmxmSmxnV'nxn

Let the matrix A is mxn matrix. It expresses an mxn matrix as

A = U*S*V'

Here, U, V are mxm and nxn orthogonal matrices which are left and right singular vectors for corresponding singular values. It is a eigenvalue method. This function removes extra rows or columns of zeroes from the diagonal matrix of singular values S, along with the columns in U or V that multiply those zeroes in the expression A = U*S*V'. By doing this removing of zeroes and columns can decrease storage necessity without comprising the accuracy also decreases execution time.

C. *Classification:*

Generally, now a days there are several types of models for SER system to extract the desired features from the audio files. Once the data is collected, it is fed into classifier. The primary function of the classifier is to analyze the speaker's emotional state by employing various algorithms. This process is conducted on a specific dataset. Optimal results and accuracy depend on the choice of the appropriate classifier. Some of the classifiers that we used are Hidden Markov model (HMM), Convolutional Neural Network (CNN), Support Vector Machine (SVM), Long Short Term Memory (LSTM).

i. Hidden Markov Model(HMM):

Generally, all the events that are upto the classification are not visible,there will be some hidden events. HMM is generally a Markov chain whose internal states are observed only through some probabilistic functions. The hidden states of the model helps in capturing the attributes of the data. The goal of HMM is to adapt a Markov chain by observing its hidden states. At the time of classification,speech signal is taken and the probability of every single speech signal in the model are calculated and compared it to the test sample.For better understanding of HMM model. Let's say that there are two friends M and N. M mood changes according to the breakfast he eats in the morning. N looks at M and observes how his mood changes daily. N does he not know the breakfast taken by M, but he predicts it by the mood of the M. Breakfast of M is a hidden event but it can be predicted by known events. This implies that it allows us to predict a sequence of hidden variables from the set of observable variables. It describe about sequence of events. It is also called as sequential generating probabilistic model.

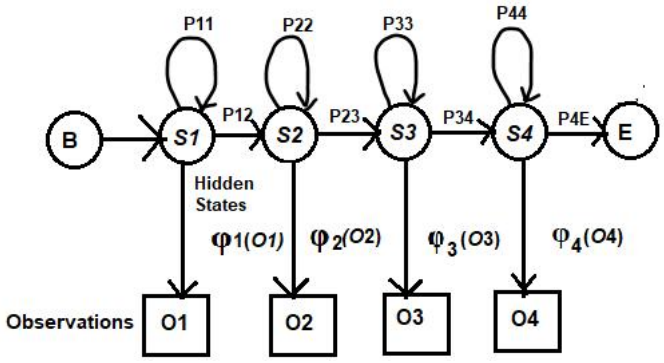

Fig 3: Hidden Markov Model (HMM)

HMM model tells about both observable events and hidden events which we don't observe them directly as we can see in Fig 3. This proposed methodology states that a database has to be selected first and then the utterances has to be converted into wav file format. Then we train the data and using SVD algorithm the features are to be extracted and every file will be labeled. Now using HMM we classify the test samples.

ii. Support Vector Machine(SVM):

Speech emotion recognition involves the categorization of spoken language into different emotional states. SVMs can be used in this context by applying them to acoustic attributes extracted from speech signals. These features, such as pitch, intensity, and spectral characteristics, can be used to differentiate between different emotional states. We use rbf function for plotting points in 3D as shown in Fig 4.

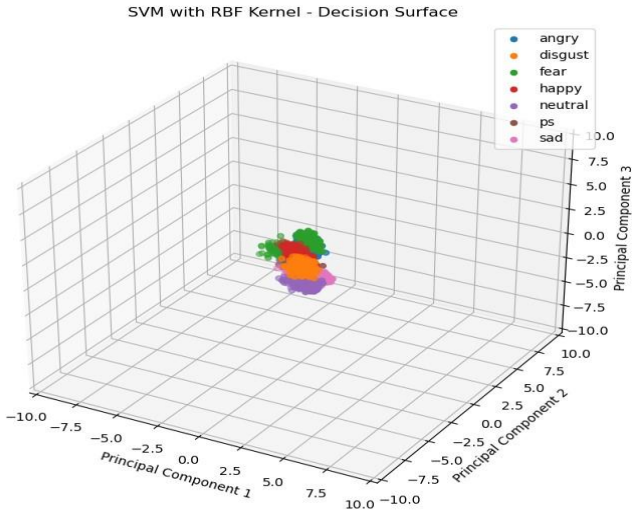

Fig 4: SVM with RBF Kernel

In SVM, the algorithm seeks to find a hyperplane that separates the acoustic features corresponding to different emotional states. This hyperplane can then be used to classify new speech signals into the appropriate emotional state. SVMs have been shown to be effective in speech emotion recognition tasks and have been used in various applications, such as detecting emotional distress in speech for mental health diagnosis or analyzing customer feedback in call center recordings.Overall, SVMs provide a powerful and efficient tool for speech emotion recognition, allowing for accurate classification of speech signals based on their acoustic features.

iii. Long Short Term Memory (LSTM):

Long short-term memory(LSTM) it refers to analogy that standard recurrent neural networks (RNN) have both long-term memory and short-term. The LSTM architecture main aim is to provide short-term memory for RNN that leads to lost of thousands of time steps. The network changes from ones per episode of training analogous to have physiological changes in long-term memories the output of previous step is used in the input of current step in recurrent neural network long short-term memory can be used for unsegmented,linked handwriting recognition and speech recognition. From Fig 5 the LSTM consists of four neural networks and numerous memory blocks known as cells in chain structure.

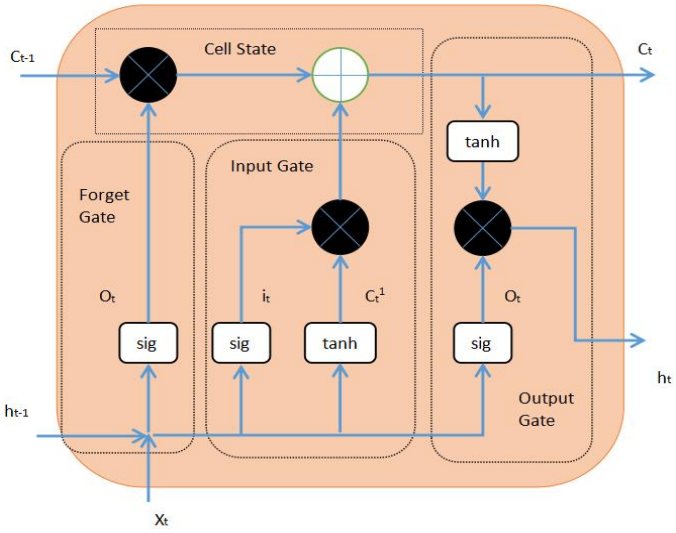

Fig 5: LSTM CELL

In LSTM there are three entrances

1. Input gate:

Each input value is used to change the memory and the sigmoid function determines to allow 1 or 0 value. The tanh function assigns weight to the data provided,determining their importance on a scale of -1 to 1.

2. Forget gate:

In forget gate it tells about the current and previous information are kept and thrown out the values are passed into a sigmoid function,which can only output value between 0 and 1. The value 0 means that all preceding information is forgotten and 1 means that all preceding information is kept.

3. Output gate:

The LSTM model's output is derived in the Output Gate. Depending on the context, it might be, for instance, a term that enhances the sentence's meaning

iv. Convolutional Neural Network(CNN):

In this paper we used CNN as it is widely used in image classification but also used in the natural language processing applications. A CNN consists of multiple layers of interconnected nodes and each layer performs a specific kind of data processing as we can observe from Fig 6. CNN's can also be used for speech recognition tasks, where they are typically applied to spectrograms or other time-frequency representations of audio signals. The layers in a CNN for speech typically include:

1. Input layer: This layer receives the spectrogram or other audio representation as input.
2. Convolutional layers: These layers apply convolutional filters to the input to extract features.
3. Pooling layers: These layers reduce the dimensionality of the features extracted by the convolutional layers, typically by taking the maximum or average value within a small region.
4. Fully connected layers: These layers take the pooled features as input and produce the final output, which is often a probability distribution over the possible speech transcriptions.

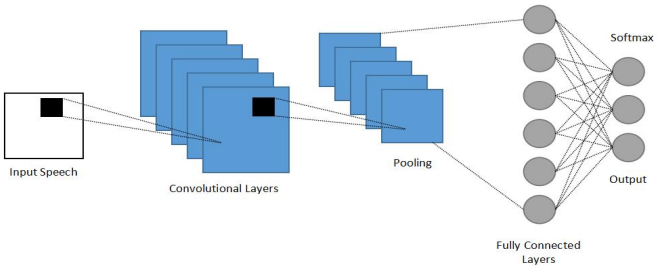

Fig.6 :Convolutional Neural Network(CNN)

Overall, a CNN for speech recognition can be seen as a series of layers that gradually extract more and more abstract features from the input spectrogram, culminating in a prediction of the most likely transcription for the spoken words.We use ReLu function as activation function as it is defined as

$$f(x) = \max(0, x)$$

IV. RESULTS & DISCUSSION

As now we discussed about the models and algorithms as for feature extraction we wav files, SVD algorithm and for classification we use CNN, LSTM, SVM and HMM models. CNN, LSTM, SVM and HMM models with SVD algorithm are designed and implemented on Jupyter Notebook. We extract attributes from audio files in database and train data. After labeling the trained data, a test speech sample is loaded from database which we trained and extract features from speech signal which will be stored in the workspace then compare it with the trained data by four models as classifier.

A. *Data Sheet*

Now, let us check the recognition rate of each model with TESS database:

TABLE 1: Recognition Rate of Emotions of RESS Database

Methods / Database	SVD	CNN	LSTM	HMM
TESS	97%	99%	98%	85%

For berlin database the accuracy is getting about 82% on average.From the TABLE 1 we can say that the TESS database achieve high accuracy than the berlin database. Here, we utilized two distinct databases featuring varied languages, utterances, and emotional expressions. When compared to TESS database the other database have less utterances with several emotion. By this we can say that the higher the utterances the more the accuracy increases.

B. Frequencies of Different Emotions

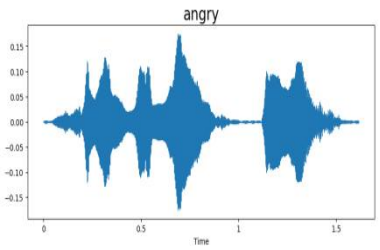

Fig. 7 Speech signal of Anger

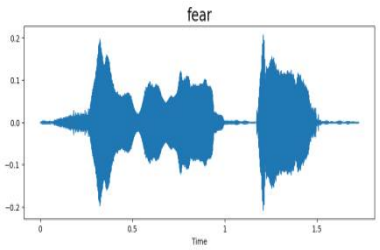

Fig. 8 Speech signal of Fear

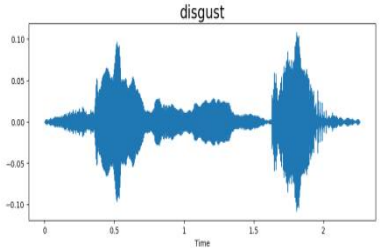

Fig.9 Speech signal of Disgust

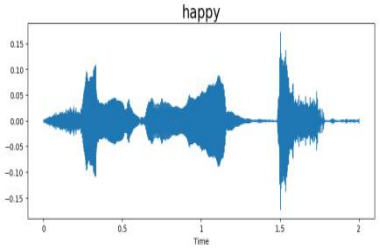

Fig.10 Speech signal of Happiness

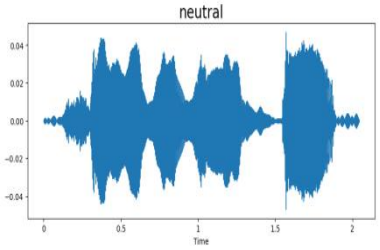

Fig.11 Speech signal of Neutral

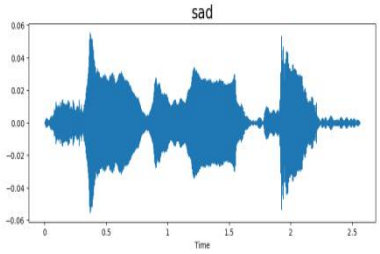

Fig. 12 Speech signal of Sadness

C. Spectogram of Different Emotions

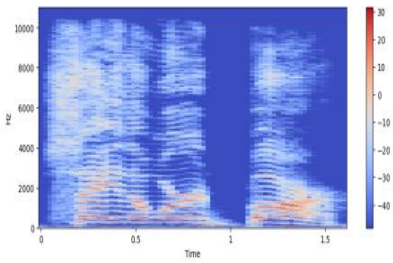

Fig. 13 Spectogram of Anger

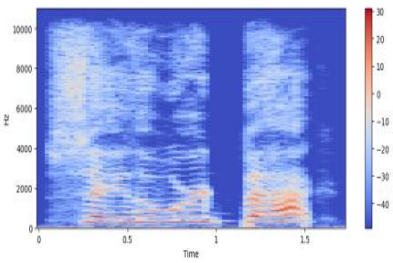

Fig.14 Spectogram of Fear

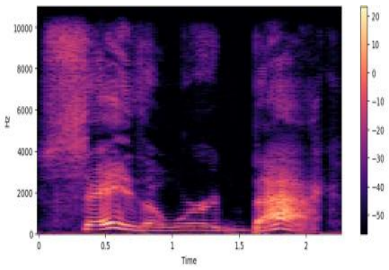

Fig. 15 Spectogram of Disgust

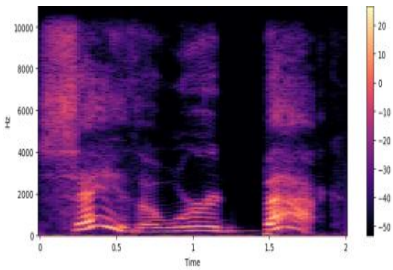

Fig. 16 Spectogram of Happiness

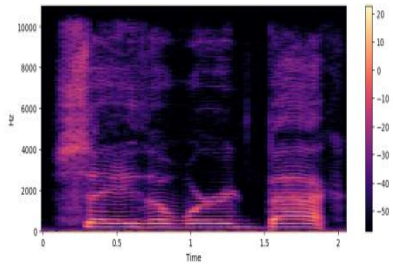

Fig. 17 Spectogram of Neutral

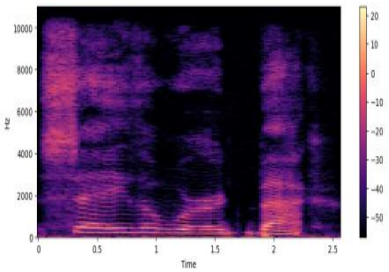

Fig.18 Spectogram of Sadness

The above Fig 7,8,9,10,11,12 were frequency speech signals and Fig 13,14,15,16,17,18 are spectrograms of different emotions from the Tess database. The features extracted in speech signal are pitch, frequency from the input speech signals. As now we discussed about how the results displayed after the program.

D. Graphical Analysis

1. HMM:

By using Hidden Markov model with svd we get to see that accuracy increases as the increases of utterances of different databases with decrease in loss. As the utterances are more then train data also increases then hidden states also increases with decrease in loss. We can check the recognition rate of emotions of sample databases with SVD by using table to compare which database yields more accuracy.

2. SVM:

Here, we are using svm classifier by using SVD with rbf as kernel as we can see from fig 19 that accuracy score is about 97%.

```
accuracy = accuracy_score(y_test, y_pred)
print('Accuracy:', accuracy)

Accuracy: 0.9714285714285714
```

Fig 19: SVM Accuracy

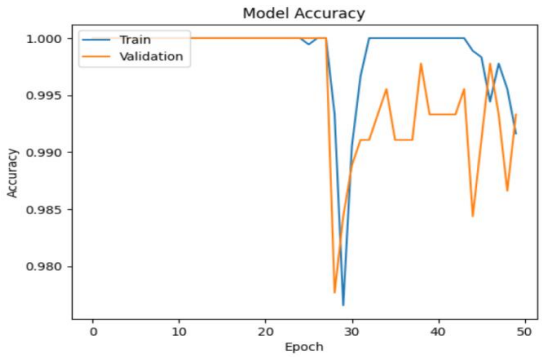

Fig 20 :Train Accuracy vs Val Accuracy(CNN)

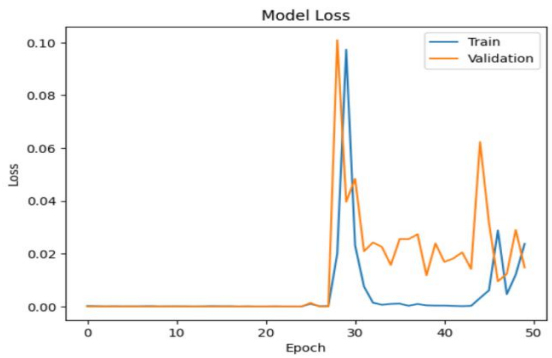

Fig 21 : Train Loss vs Val Loss(CNN)

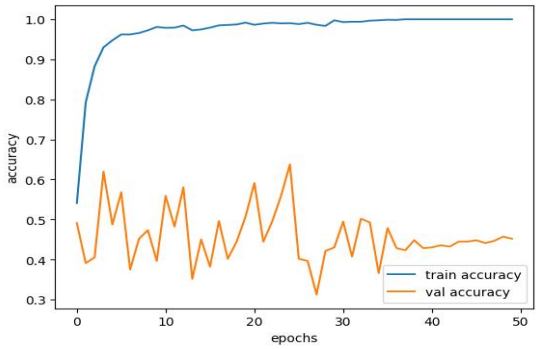

Fig 22 :Train Accuracy vs Val Accuracy(LSTM)

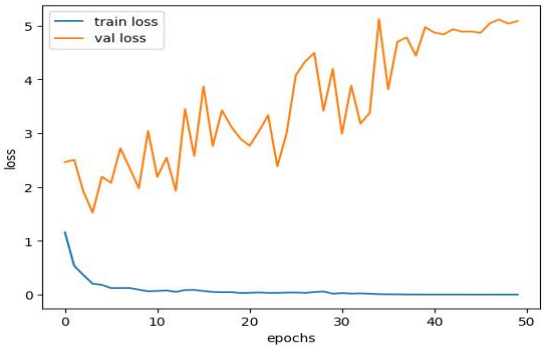

Fig 23 : Train Loss vs Val Loss(LSTM)

3. CNN:

We are using CNN classifier by using SVD here we are assigning of 50 epochs and for each epochs the we can see the increase in the accuracy and decrease of loss in the below graphs are train accuracy was increasing by the increase of epochs with val accuracy and also decease in train data loss as the val loss increases. From Fig 20 & 21 we observe that after some epochs at 28th epoch the accuracy decreased due to overfitting and again accuracy increases due to overfitting.

4. LSTM:

Here, we are using LSTM classifier which is extended of RNN model by using SVD here we are assigning of 50 epochs and from Fig 22 & 23 for each epochs we can find the increase in the accuracy and decrease of loss in the below graphs are train accuracy was increasing by the increment of epochs with with val accuracy and also decease in train data loss as the val loss increases.

V. CONCLUSION

In this research, we introduced a Speech Emotion Recognition (SER) system that utilizes the SVD algorithm in combination with various classifiers. Our approach involved using different databases containing up to seven emotions. Through the SVD process, we extracted features from two types of databases, and then used classifiers for feature matching. This investigation enabled us to evaluate how different classifiers and features influence the accuracy of the speech emotion recognition system. During feature selection, we focused on selecting highly discriminant features, which were then compared with the features obtained from the trained data and test samples. The results of our SER study indicated that databases with a higher number of utterances achieved higher accuracy in emotion recognition compared to others. To implement our approach, we employed the SVD method to extract features from audio files, and we experimented with several classification models, including HMM, SVM, CNN, and LSTM. As evidenced in the results and discussion section, we observed an improvement in accuracy with an increase in the number of utterances in the databases. In conclusion, this study provides a viable method for extracting audio signal features using SVD and demonstrates its effectiveness when compared to previous methods. The outcomes suggest that our proposed approach holds promise for speech emotion recognition.

REFERENCES

[1] D. Sharma, S. Jain and V. Maik, "Energy efficient clustering and optimized loadng protocol for iot," Intelligent Automation & Soft Computing, vol. 34, no.1, pp. 357–370, 2022.

[2] Application of artificial intelligence in energy efficient hvac System design: a case study P Adhikary, S Bandyopadhyay, S Kundu - ARPN journal of Engineering and Applied sciences, 2017.

[3] M. Dhivya and T. Parameswaran, “Smart scheduling on cloud for IoT-based sprinkler irrigation,” International Journal of Pervasive Computing and Communications, 2020, http://dx.doi.org/10.1108/IJPCC-03-2020-0013.

[4] D. Sharma, S. Jain and V. Maik, "Optimized tuning of loadng routing protocol parameters for iot," Computer Systems Science and Engineering, vol. 46, no.2, pp. 1549–1561, 2023.

[5] Mehmet Berkeehan Akcay, Kaya Oguz (2020). “Speech emotion recognition: Emotional models databases, features, prepossessing methods, supporting modalities and classifiers”.

[6] Teddy Surya Gunawan, Muhammad Fahreza Alghifari, Arman Morshidi, Mira Kartiwi. (2017). A Review on Emotion Recognition Algorithms using Speech Analysis.
[7] M. Maithri, U. Ragavendra, Anjan Gudigar, Jyothi Samanth, Prabal Datta Barua, Murugappan Murugappan, (2022). “Automated emotion recognition: Current trends and future perspectives”.
[8] Livingstone SR, Russo FA (2018). “The Ryerson Audio-Visual Database of Emotional speech emotion and Song (RAVDESS)”.
[9] Leila Kerkeni, Youseef Serrestou, Mbarki, Raoof, Ali Mahjoub, Cleder (2019).“ Automatic Speech Emotion Recognition”.
[10] Saliha Benkerzaz, Youssef Elmir, Abdeslem Dennai (2019). “A Study on Automatic speech emotion Recognition”.
[11] Harshavardhan GM, Mahendra Kumar Gourisaria, Manjusha Pandey, Siddharth Swarup Rautaray, (2020). “A comprehensive survey and analysis of generative models in machine Learning”.
[12] Palani Thanaraj Krishnan, Alex Noel, Vijayarajan (2021). “Emotion classification from speech signal based on empirical mode decomposition and non-linear features”.
[13] Nithya roopa S, Prabhakaran M, Betty.P,(2018). “Speech Emotion Recogntion using Deep Learning”.
[14] Raoudha YAHIA CHERIF, Abdelouahab MOUSSAOUI, Nabila FRAHTA, Mohamed BERRIMI(2021).”Effective speech emotion recognition using deep learning approaches for Algerian dialect”.
[15] Kourosh Modarresi (2015) “Unsupervised Feature Extraction Using Singular Value Decomposition”.
[16] J.Rong,Y.P.P.Chen, “Acoustic feature selection for automatic Emotion Recognition from speech”.
[17] Zhe Chen, Yanmei Zhang, Junbo Zhang, Rui Zhou, Zhen Zhong, Chaogang Wei, Yuhe Liu Jing Chen,(2021).”Cochlear Synaptopathy: A Primary Factor Affecting Speech Recognition Performance in Presbycusis”.
[18] Shah Fahad, Ashish Ranjan, Jainath, Akshay Deepak,(2021). “A survey of speech emotion recognition in natural environment”.
[19] S. S. Narayanan, (2005). “Toward detecting emotions in spoken dialogs,” IEEE Trans. Speech Audio Process.