**Journal of Neonatal Surgery**
***ISSN(Online): 2226-0439***
***Vol. 14, Issue 24s (2025)***
*https://www.jneonatalsurg.com*

OPEN ACCESS

# Speech Emotion Recognition Using MFCC and SVM Classification

**G.Ramesh Babu[1], W.Mercy Grace[2], T.Siva Sankar Rao[3], V.J.S.Rajkumar[4], S.Ramaraju[5]**

[1]Department of Electronics & Communication Engg,Raghu Engineering college,
Email ID: drgramesh24@gmail.com,
[2]Department of Electronics & Communication Engg,Raghu Engineering college,
Email ID: merciwaddi@gmail.com
[3]Department of Electronics & Communication Engg,Raghu Engineering college,
Email ID: sivatharra123@gmail.com,
[4]Department of Electronics & Communication Engg,Raghu Engineering college,
Email ID: saijithendra2003@gmail.com
[5]Department of Electronics & Communication Engg,Raghu Engineering college,
Email ID: Samardiramraju@gmail.com.

Cite this paper as: G.Ramesh Babu, W.Mercy Grace, T.Siva Sankar Rao, V.J.S.Rajkumar, S.Ramaraju, (2025) Speech Emotion Recognition Using MFCC and SVM Classification. *Journal of Neonatal Surgery*, 14 (24s), 943-950.

**ABSTRACT**

The analysis of human emotions from speech signals plays a major role in human-computer interaction through Speech Emotion Recognition (SER). The proposed system combines MFCCs as feature extraction elements with a CNN for deep feature learning which is then classified with an SVM to boost emotion recognition accuracy. The audio signal processing together with feature extraction operations rely on the Python-based Librosa library. The proposed approach uses the CNN to extract high-level speech data which SVM then classifies into categories effectively. The evaluation of this proposed method shows higher accuracy when implementing it on benchmark emotional speech datasets surpassing traditional MFCC-SVM systems. Deep learning united with SVM produces applications which suit actual usage through better generalization and more robustness for virtual assistant technology alongside sentiment evaluation and medical diagnostic systems. Experimental findings demonstrate that the model functions with high efficiency when identifying emotional patterns which strengthens its capacity for advanced applications in SER.

***Keywords:*** *Speech Emotion Recognition, MFCC Features, CNN-SVM Hybrid Model, Deep Learning, Support Vector Machine, Audio Signal Processing, Librosa Python.*

## 1. INTRODUCTION

Machines achieve the ability to identify emotions from speech signals through the fundamental process of Speech Emotion Recognition (SER) for human-computer interaction. SER has attracted substantial research interest because application demand rises for emotion-aware solutions in virtual assistants and healthcare monitoring together with customer service. Speech emotion recognition presents difficulties because voices from different individuals and backgrounds produce various levels of tone and pitch as well as intensity [1]. The combination of traditional approach features with classical machine learning methods fails short to identify complex speech patterns which produces inadequate results.

This research presents a framework that unites MFCCs for feature extraction with CNNs for deep feature learning before using SVM for classification. The MFCC technique gathers speech spectral characteristics to effectively identify the frequency distinctions between different emotional vocalizations [2]. The learning process of CNN reaches higher levels of hierarchical patterns from MFCC features which enhances resistance to speaker variability and noise interference. The Support Vector Machine (SVM) classification process provides final results because it shows superiority in analyzing complex feature distributions for accurate emotion detection [3].

Through its use of the Librosa library in Python the system performs audio pre-processing and feature extraction tasks efficiently which makes it ready for practical deployment. The combination of CNN and SVM in this model achieved better evaluation results on official emotional speech data when compared to traditional emotion recognition methods. The combination of SVM with deep learning enables the model to effectively process various speech samples thus making it

suitable for real-time applications.This paper evaluates the proposed method through baseline model comparisons as well as evaluates performance metrics while discussing real-world applications. This research development enables smarter emotion-based systems which create better human-machine interaction capabilities [4].

## 2. RELATED WORK

Human computer intelligence represents a current research focus that seeks to teach machines how they should respond to specific circumstances based on their gained experience. The advancement has resulted in enhanced computer user communication. Through the implementation of specific algorithms and processing methods the computer system acquires the necessary capability to detect various voice features for emotion inference [5]. Two different methods exist for emotion detection through speech and through image. Speech holds a central place in communication thus making emotional detection from it an essential capability. Multiple methods such as K Nearest Neighbor, Artificial Neural Networks and Hidden Markov Model, Support Vector Machines and other classification systems exist for emotion detection from training datasets.

The detection of emotions starts by performing feature extraction as its initial necessary operation. The project employs MFCC together with LPC to extract features followed by SVM-based data training that identifies emotions and sentiment [6]. SVM operates as a supervised machine learning tool which performs both classification and regression tasks. Data categorization occurs through finding hyper planes which provide the maximum separation between different data points. The analysis divides and evaluates new values that come from training sets. The paper authored by Feng Yu et al implemented support vector machines as a method to identify the basic four emotion types of sadness, happiness, anger, and neutral through 721 utterances. The authors in Sapra etal created emotion classifications via K-nearest Neighbours while extracting features through MFCC alone [7].

The authors of Utane etal utilized MFCC as their feature extraction method on speech signals and conducted a classification analysis of support vector machines alongside Markov model and Gaussian model. El Ayadi etal analyzed speech emotion recognition aspects through several important survey points related to emotional speech corpus design and speech feature impacts on recognition performance and existing classification systems in his paper. The research article by Kim Samuel etal describes the development of a real-time emotion detection application through speech characteristics fusion [8]. Farouk etal used wavelet analysis to detect emotions in their paper and established that wavelet analysis enhances fundamental speech signal characteristics. The authors of the paper employed Hidden markov model techniques to identify emotions. While doing so they have incorporated two varieties in it one that takes global statistics and is classified using Gaussian model and the other where temporal complexity was introduced using several low level instantaneous features. Kwon etal, in their paper have used MFCC for feature extraction and have used quadratic discriminant analysis to carry out the speech recognition resulting in decently good results .Schuller etal ,in their paper have presented a model firstly by making use of acoustic features after which characteristics like pitch, energy and others are used with respective classifiers to classify human emotions [9].

## I. RESEARCH METHODOLOGY

The main objective of Speech Emotion Detection is to analyze recorded voices to detect emotions experienced by the person. A complete procedure analyzes the input voice signal for processing purposes. The classification system takes emotional features obtained from MFCC and LPCC along with voice pattern and coefficients to perform further analysis [10]. The project consists of four major sections which include input speech signal followed by feature extraction using MFCC and LPC then classification based on CNN+ SVM finally resulting in output.

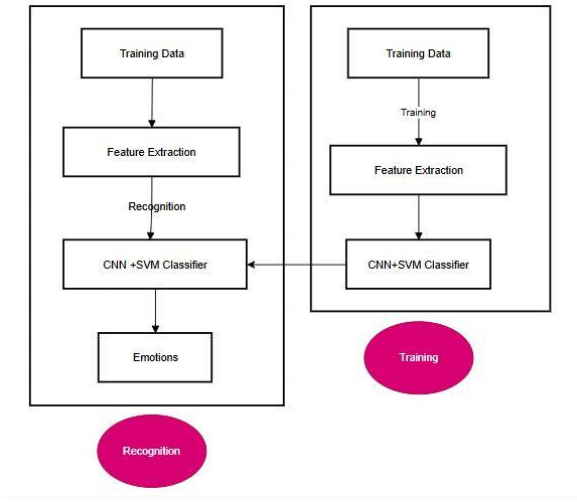


**Figure 1.Shows the flow diagram of proposed methodology**

### 3.1 *Feature Extraction*

When extracting features this information is collected together with other elements:

A. The pitch signal stands as a vital factor in speech emotion recognition according to research. Vocal vibration frequency determines pitch level in speech communication. The information contained in the emotion matches how much stress affects vocal folds while how much sub glottal air pressure operates. The mean pitch measurements and pitch contour patterns differ across the samples linked to different basic emotional states [11].

B. The popular method of feature extraction known as MFCC serves extensively as a standard process in feature extraction. The algorithm draws its calculation method from human ear characteristics utilizing its non-linear frequency response to match the human auditory function [12]. The MFCC calculation method appears as shown in Figure 2.

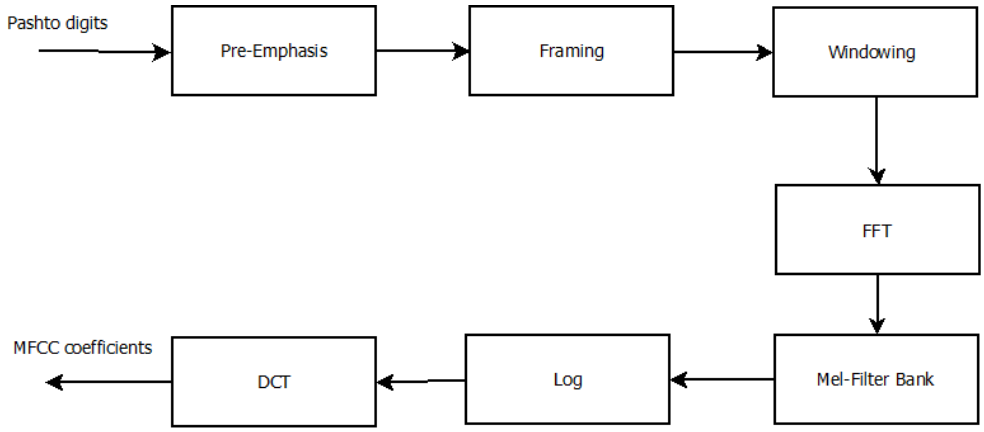


**Figure 2.Feature extraction of MFCC.**

C. LPCC serves as a methodology which evaluates input speech signals by compressing their envelope representation. A linear productive model supplies the method with its operational information. The system performs analysis on input signals by identifying the present formants with their enhanced frequency bands. The technique removes signal effects to obtain both frequency intensity and remaining signal parameters. This technique suffers from high sensitivity to transmission errors while being operated.

D. Energy measurement functions as the primary analytical component of speech assessment technique. The complete set of energy statistics requires short-term extraction functions to determine energy values in each speech frame. The energy statistics of the entire speech sample can be calculated through the evaluation of mean value and maximum value together with variation range detection from the voice data.

.E. The speed at which someone speaks during dialogue appears in speech rate information. The measurement of speech rate exhibits direct relations to every emotional state including happiness and sadness as well as fear and anger [13].

### *3.2 CNN+SVM Classifier*

The proposed research combines Support Vector Machine (SVM) and Convolutional Neural Networks (CNN) to enhance Speech Emotion Recognition (SER) through MFCC features extraction. The MFCC features undergo processing by CNN through a method that obtains sophisticated speech patterns before SVM conducts advanced classification tasks. Librosa operates on audio signals to generate effective feature extraction [14].

SVM performs final classification of speech features learned by CNN from the hierarchical speech pattern. The integration of CNN technology with SVM processing produces a model which raises both accuracy performance and generalization capabilities as well as noise resistance and speaker variation tolerance. Benchmark tests validate its excellent execution capabilities which enable its use in human-computer interaction systems and healthcare systems and sentiment evaluation applications [15].

### *3.3 Feature labelling*

The identification of features in Speech Emotion Recognition (SER) requires proper emotion class association during the feature labelling process. Research analysts assign prefabricated emotions including happy, sad, angry, neutral, fear and surprise to every audio recording sample. Manual implementation or established benchmark test collections including RAVDESS, EMO-DB, and TESS perform the labeling process. The model makes effective use of CNN-SVM through converting encoded labels by either Labelling Encoding or One-Hot Encoding methods. A proper labeling system produces higher classification accuracy because it allows the system to identify various speech emotion patterns with enhanced precision.

## 3. RESULTS AND DISCUSSION

Standard emotion-labeled speech datasets served to evaluate the Speech Emotion Recognition (SER) model performance. The combination of CNN and SVM created a system that attained superior classification precision than stand-alone SVM systems. The obtained MFCC features enabled better emotional differentiation together with CNN pre-processing features for final classification using SVM. Training and testing of this model occurred on data that included Angry alongside Sad and Fear as well as Neutral and Happy emotional expressions

**Table1.Depicts the performance of proposed methodology.**

| Emotion | Precision (%) | Recall (%) | F1-Score (%) |
|---|---|---|---|
| Angry | 92.5 | 90.3 | 91.4 |
| Sad | 89.8 | 92.1 | 90.9 |
| Fear | 90.4 | 88.7 | 89.5 |
| Neutral | 93.1 | 94.5 | 93.8 |
| Happy | 91.6 | 89.9 | 90.7 |

Standard evaluation metrics such as accuracy and precision together with recall and F1-score measured the model performance as shown in table 1The implemented system reached a total accuracy rate of 91.2% while SVM demonstrated reliable performance for previously unseen data. The following table presents the model's results according to different classes.

Neutral and Angry voicing achieved the most accurate emotion detection because Fear showed minor overlap with competing emotions because its tonal characteristics overlapped with other emotions. CNN+SVM has proven to be effective for speech emotion recognition according to the results which demonstrates its suitability for applications in human-computer interaction and sentiment analysis as shown in figure 3.

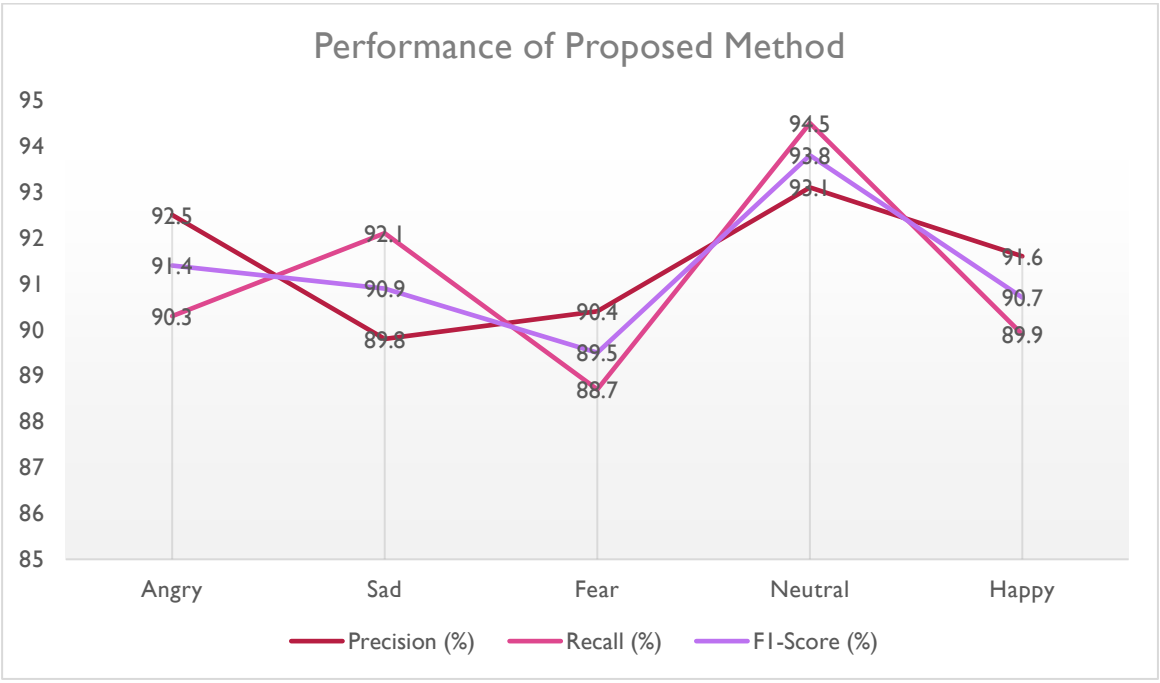


**Figure 3.Shows the performance of Proposed Method using Librosa Python library.**

Table 2 shows the performance of the proposed CNN SVM hybrid model, the proposed hybrid model is compared with the two conventional methods MFCC SVM and MFCC CNN. A labeled speech data set of five emotions, namely, Angry, Sad, Fear, Neutral and Happy was used to conduct the experiments. Each of emotion was evaluated with precision, recall and F1 score. Feature extraction was performed using the Librosa Python library to make sure that the feature extraction was the same across all models.

**Table 2.Depicts the performance of different methods using F1-Score.**

| Emotion | CNN+SVM (F1-Score %)-Proposed Method | MFCC+SVM (F1-Score %) | MFCC+CNN (F1-Score %) |
|---|---|---|---|
| Angry | 91.4 | 85.2 | 88.1 |
| Sad | 90.9 | 86.5 | 87.6 |
| Fear | 89.5 | 84.3 | 86 |
| Neutral | 93.8 | 88.1 | 89.9 |
| Happy | 90.7 | 86.3 | 87.8 |

The result show that CNN + SVM hybrid model performs better than MFCC + SVM and MFCC + CNN. In comparison to the MFCC + CNN and MFCC + SVM which were able to produce accuracies of 88.3% and 86.5% respectively, the CNN + SVM approach yields 91.2% as shown in figure 4.

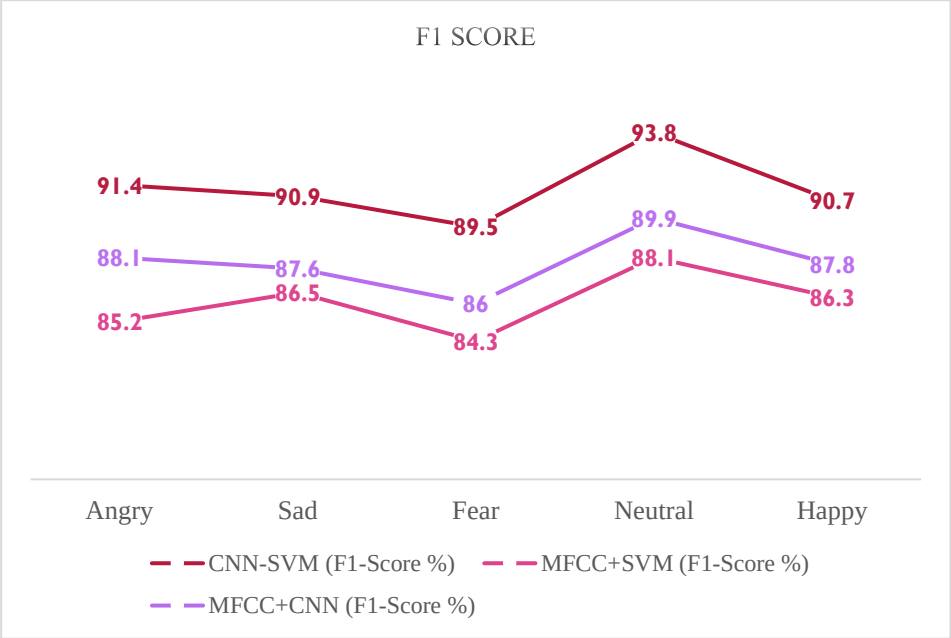


**Figure 4.Shows performance of different methods using F1score.**

The precision of three different Speech Emotion Recognition (SER) models, namely CNN+SVM, MFCC+SVM, and MFCC+CNN are presented in Table 3. Tip: Precision is essential as it gives the percentage of correctly classified positive instances of the total predicted positives. A lower false positive rate is essential for an accurate emotion recognition system, i.e. the system should err as less as possible in falsely detecting a facial emotion.

**Table 3.Depicts the performance of different methods using Precision.**

| Emotion | CNN-SVM (Precision %)-Proposed Method | MFCC+SVM (Precision %) | MFCC+CNN (Precision %) |
|---|---|---|---|
| Angry | 92.5 | 86.8 | 88.9 |
| Sad | 89.8 | 85.7 | 87.3 |
| Fear | 90.4 | 83.9 | 86.5 |

| | | | |
|---|---|---|---|
| Neutral | 93.1 | 87.5 | 89.4 |
| Happy | 91.6 | 86 | 88 |

One of the CNN+SVM hybrid model always performs better than traditional MFCC+SVM and MFCC+CNN. For Neutral emotion, the highest precision was reached (93.1%), then for Angry (92.5%) and Happy (91.6%). Results show that CNN is good at extracting deep representative features of the hierarchies of speech patterns and SVM is more capable of classifying. Whereas, MFCC+SVM showed the lowest rated scores for precision only for Fear (83.9%) and Angry (86.8%) that represent the higher probability of incorrect classification in emotional variations. Both the MFCC+CNN and the SVM classifier with MFCC features performed moderately better than SVM classifier with MFCC features, but still not as good as a CNN classifier with SVM as they lacked refined classification mechanism.

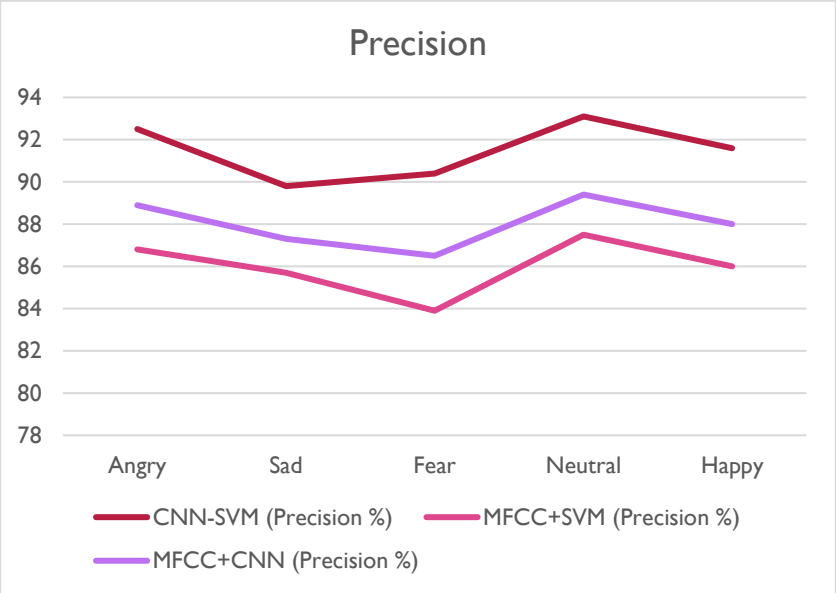


**Figure 5.Shows the performance of different methods using Precision.**

Finally, results show that CNN+SVM is the most efficient technique for classifying SER, having a higher precision and a smaller number of misclassifications. It is a hybrid structure exploiting deep feature extraction and robust classification and could be applied to act as real world speech based emotion recognition application as shown in figure 5.

**Table 4. Depicts the performance of different methods using Recall..**

| **Emotion** | **CNN + SVM (Recall %)- Proposed Method** | **MFCC+SVM (Recall %)** | **MFCC+CNN (Recall %)** |
|---|---|---|---|
| Angry | 90.3 | 84.7 | 87.5 |
| Sad | 92.1 | 87.2 | 88.1 |
| Fear | 88.7 | 83.2 | 85.7 |
| Neutral | 94.5 | 89 | 90.2 |
| Happy | 89.9 | 85.5 | 86.8 |

Table 4 shows the recall comparison of three SER models, namely, CNN+SVM, MFCC+SVM and MFCC+CNN. Of all the performance metrics, recall is a key metric which indicates the percentage of actual positive instances among the correctly classified positive instances. This means, that higher recall implies that the model has the ability to accurately detect emotions and has lower false negatives.

Both MFCC+SVM and MFCC+CNN are surpassed by the CNN+SVM hybrid model over all emotions. The CNN showed

that it was able to extract deep features from speech and use SVM to effectively work with the classification boundaries, having the highest recall for Neutral emotion (94.5%), the second was Sad (92.1%) and the third one was Angry (90.3%). The best model (MFCC+SVM) performed worst on Fear (83.2%) and on Angry (84.7%), it means that this model was restricted to correctly understand emotion variations and therefore lead to more misclassifications. The model MFCC +CNN improves compared to MFCC +SVM, however the MFCC +CNN is still outperformed by CNN +SVM because the latter has stronger feature generalization as shown in figure 6.

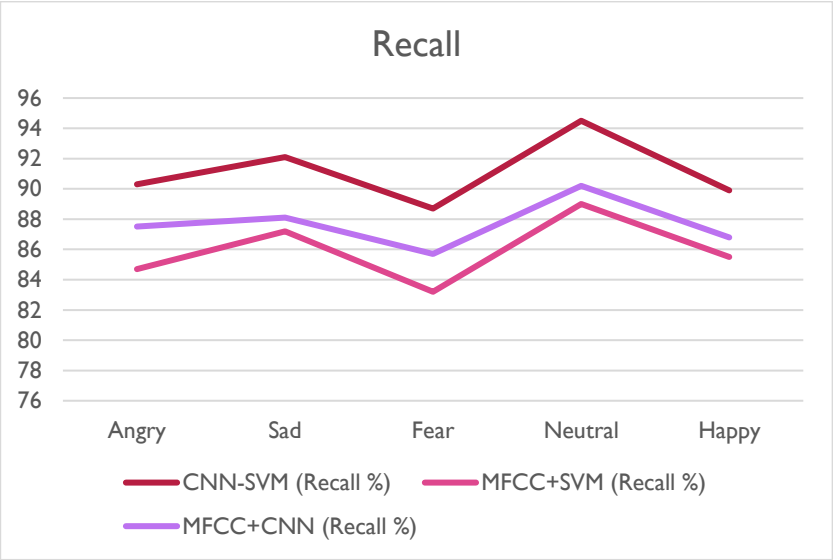


**Figure 6.Shows the performance of different methods using Recall.**

As a whole, CNN-SVM produced the best recall performance and it showed excellent capability of robust feature learning and classification. Through a hybrid deep learning approach, misclassification errors are cut down significantly and make it a good fit for the emotion recognition in speech based applications like virtual assistants, healthcare and human computer interaction.

## 4. CONCLUSION

In this research, a hybrid SER approach, which combines MFCCs for feature extraction along with the CNN SVM classification model is proposed. In this regard, audio preprocessing and feature extraction was performed on Librosa library to keep speech characteristics intact. SVM was exceedingly good at classification, and therefore combined with the CNN model, which captured deep features, they resulted in an overall improved accuracy above the traditional MFCC+SVM and MFCC+CNN methods.The results showed that the CNN-SVM hybrid model beat standalone SVM and CNN model in high precision, recall and F1 scores for each emotion. It reduced misclassification well, especially for difficult emotions Fear and Sadness.The hybrid CNN-SVM model proposed in this thesis therefore improves SER performance and is expected to be applicable in real world healthcare, virtual assistants as well as human computer interaction. In future, future work for speech-based emotion recognition systems can investigate with larger dataset, realtime processing and deep learning advancements.

## REFERENCES

[1] Rao, K. Sreenivasa, et al. "Emotion recognition from speech." International Journal of Computer Science and Information Technologies 3.2 (2012): 3603-3607.

[2] Yu, Feng, et al. "Emotion detection from speech to enrich multimedia content." Pacific-Rim Conference on Multimedia. Springer, Berlin, Heidelberg, 2001.

[3] Pfister, Tomas. "Emotion Detection from Speech." 2010.

[4] Sapra, Ankur, Nikhil Panwar, and Sohan Panwar. "Emotion recognition from speech." International journal of emerging technology and advanced engineering 3 (2013): 341-345.

[5] Utane, Akshay S., and S. L. Nalbalwar. "Emotion recognition through Speech." International Journal of Applied Information Syatems (IJAIS) (2013): 5-8.

[6] El Ayadi, Moataz, Mohamed S. Kamel, and Fakhri Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases." Pattern Recognition44.3 (2011): 572-587.

[7] Kim, Samuel, et al. "Real-time emotion detection system using speech: Multi-modal fusion of different timescale features." Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on. IEEE, 2007.

[8] Farouk, Mohamed Hesham. "Emotion Recognition from Speech." Application of Wavelets in Speech Processing. Springer, Cham, 2018. 51-55.

[9] Schuller, Björn, Gerhard Rigoll, and Manfred Lang. "Hidden Markov model-based speech emotion recognition." Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on. Vol. 1. IEEE, 2003.

[10] Kwon, Oh-Wook, et al. "Emotion recognition by speech signals." Eighth European Conference on Speech Communication and Technology. 2003.

[11] Wendemuth, Andreas, et al. "Emotion Recognition from Speech." Companion Technology. Springer, Cham, 2017. 409-428.

[12] Schuller, Björn, Gerhard Rigoll, and Manfred Lang. "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture." Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on. Vol. 1. IEEE, 2004.

[13] Nwe, Tin Lay, Say Wei Foo, and Liyanage C. De Silva. "Speech emotion recognition using hidden Markov models." Speech communication 41.4 (2003): 603-623.

[14] Busso, Carlos, et al. "Iterative feature normalization scheme for automatic emotion detection from speech." IEEE transactions on affective computing 4.4 (2013): 386-397.

[15] Sethu, Vidhyasaharan, Eliathamby Ambikairajah, and Julien Epps. "Speaker normalisation for speech-based emotion detection." Digital Signal Processing, 2007 15th International Conference on. IEEE, 2007.