Journal of Information Systems Engineering and Management
2025, 10(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

An Improved Convolutional Neural Network For Speech Detection

[1]*Dr. Prolay Ghosh,
Assistant Professor, Department Of Information Technology, Jis College Of Engineering, Kalyani, Nadia, West Bengal, India
Email: Prolay.Ghosh@Jiscollege.Ac.In
Https://Orcid.Org/0000-0001-9267-5766

[2]Ms. Tanusree Saha,
Assistant Professor, Department Of Information Technology, Jis College Of Engineering, Kalyani, Nadia, West Bengal, India
Email: Tanusree.Saha@Jiscollege.Ac.In

[3]Dr. Debashis Sanki,
Assistant Professor, Department Of Information Technology, Jis College Of Engineering, Kalyani, Nadia, West Bengal, India
Email: Debasish.Sanki@Jiscollege.Ac.In

[4]Shibraj Basak,
Department Of Computer Application, Jis College Of Engineering, Kalyani, Nadia West Bengal, India
Email: Shibrajbond@Gmail.Com

*Corresponding Author: [1]*Dr. Prolay Ghosh,
Email: Prolay.Ghosh@Jiscollege.Ac.In

ARTICLE INFO

Received: 16 Nov 2024

Revised: 28 Dec 2024

Accepted: 20 Jan 2025

ABSTRACT

The detection of emotions from speech is the aim of this paper. Speech consists of anger, joy and fear have very high and wide range in pitch, whereas Speech consists of sad and tired emotion have very low pitch. Speech Emotion detection technology can recognize human emotions to help machines better for understanding intentions of a user to improve the human-computer interaction. Classification model named Convolutional Neural Network (CNN) based on mainly Mel Frequency Cepstral Coefficient (MFCC) feature to detect emotion have been presented here. Different approaches have been discussed and compared to find best CNN model using different combinations of parameters. The models have been trained to distinguish eight different emotions such as calm, neutral, angry, sad, happy, disgust, fear, surprise. The proposed work shows that CNN 3 Layer model with RMSprop optimizer when trained with 80 Epochs works best among other CNN models for the RAVDESS dataset.

Keywords: Emotion Detection, CNN, RMSProp, RAVDESS, Machine Learning.

Journal of Information Systems Engineering and Management
2025, 10(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

INTRODUCTION:

Emotion detection is the process of identifying human emotions using various processes. Different methods have been introduced to detect emotion. Siri, Cortana are very intelligent assistance. Emotion detection can help to improve computer and human interaction by accepting emotions.

Identifying human emotions Accurately, is a difficult task because of their complexness. Some emotions can have different expressions, and some emotions can have similar expressions. Emotions depends on character, culture, gender, situation and locality of a person. Sometimes, it can be hard for a real person to detect emotion from speech of someone. In the case of emotion detection by computers, it is even hard to detect true emotion of a user because of the emotion's complexity. There are different types of systems that use information like audio, text, image, video to detect human emotions, but sometimes not all information is available to use. For example, Call centers can use speech recognition to detect emotions. Detecting emotions from speech is a very tough work. A person's speaking style, volume, intonation, speed, words, etc., greatly affect the detection of emotion in speech. Also character, culture, gender, situation and locality, etc. are the other factors that can affect the detection of human emotion from speech. Fortunately, several methodologies have been developed to detect human emotions from speech. These methodologies are categorized by their advantages that make them superior to others. From analyzing different works on this topic, it was observed that CNN works better with RAVDESS dataset by comparing different algorithms. In this paper, different approaches have been discussed and compared to find best CNN model using different combinations of parameters.

Remaining paper is sectioned as follows: Section II. states the related work. The methodology part includes dataset, feature extraction, algorithms, are introduced in Section III. Section IV shows System design. The results are shown in Section V and conclusions in Section VI.

I. RELATED WORK:

The paper "Emotion recognition and affective computing on vocal social media"[1] highlights the advancements and challenges in the field of emotion recognition and affective computing within the context of vocal social media. The review emphasizes the importance of accurately analyzing and understanding the emotional content conveyed through voice recordings in social media platforms.

The literature review of the paper showcases the evolution of emotion recognition techniques, ranging from traditional approaches based on acoustic features to more recent advancements utilizing machine learning algorithms and deep learning models. Feature extraction and representation methods have also evolved, incorporating time-domain, frequency-domain, and spectrogram-based features, as well as leveraging the power of recurrent neural networks and attention mechanisms.

The availability of labeled datasets specifically designed for vocal social media has played a crucial role in training and evaluating emotion recognition models. However, challenges such as dataset diversity, cultural biases, and ethical considerations still need to be addressed to ensure comprehensive and inclusive datasets.

The paper also emphasizes the significance of cross-cultural and multilingual emotion recognition, given the variations in vocal expressions and cultural norms.

Overall, the paper highlights the promising developments in emotion recognition and affective computing on vocal social media, paving the way for improved understanding of human emotions, personalized user experiences, and applications in areas such as mental health, social interactions, and sentiment analysis.

The paper "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech"[2] explores the significance of considering speaker-sensitivity in emotion recognition, specifically focusing on both acted and spontaneous speech. The study highlights the limitations of traditional approaches that treat emotions as independent and neglect the influence of individual speakers.

By incorporating a ranking-based approach, the paper addresses the inherent subjectivity of emotion perception and recognition. The research demonstrates that speaker-sensitive emotion recognition

models outperform conventional methods by considering the unique vocal characteristics and individual differences among speakers.

The findings of the study emphasize the importance of personalized emotion recognition systems that account for speaker-specific information, such as vocal style, intonation, and other contextual cues. These speaker-sensitive models enable a more accurate and nuanced understanding of emotional expression, enhancing the overall performance of emotion recognition systems.

Furthermore, the paper showcases the utility of the ranking-based approach in distinguishing between acted and spontaneous speech. By incorporating ranking techniques, the study contributes to the development of emotion recognition models that can effectively differentiate genuine emotional responses from acted or exaggerated expressions.

The research presented in the paper has significant implications for various applications, including human-computer interaction, virtual agents, and affective computing. By considering the speaker-sensitive nature of emotional expression, these systems can provide more tailored and personalized experiences for users.

Overall, the paper emphasizes the importance of considering speaker-sensitivity in emotion recognition, highlighting the advantages of a ranking-based approach for both acted and spontaneous speech. The findings pave the way for further advancements in emotion recognition systems that account for individual differences and enhance the accuracy and robustness of emotion analysis in various real-world scenarios.

The paper "Shape-based modeling of the fundamental frequency contour for emotion detection in speech"[3] focuses on the significance of shape-based modeling of the fundamental frequency contour (F0) for emotion detection in speech. The study highlights the role of F0 contour in conveying emotional information and proposes a novel approach that captures the shape characteristics of F0 for accurate emotion detection.

The research demonstrates that traditional approaches that solely rely on mean or statistical measures of F0 may not fully capture the dynamic and nuanced variations in emotional speech. By incorporating shape-based modeling techniques, such as using spline curves or statistical moments, the paper presents a more comprehensive and robust method to capture the temporal dynamics and contour patterns of F0.

The findings of the study indicate that shape-based modeling of F0 provides enhanced discrimination between different emotional states. The approach not only improves the accuracy of emotion detection but also provides insights into the underlying expressive characteristics of speech.

The proposed method holds significant potential for various applications, including affective computing, human-computer interaction, and emotional speech synthesis. By accurately capturing the shape information of F0, these applications can effectively recognize and respond to the emotional states of users, leading to more personalized and engaging interactions.

The paper highlights the importance of shape-based modeling of the fundamental frequency contour for emotion detection in speech. The research showcases the advantages of considering the dynamic and temporal characteristics of F0 and presents a novel approach that improves the accuracy and discriminative power of emotion detection systems. The findings contribute to the development of more sophisticated and effective emotion recognition techniques, with potential applications in various domains requiring the understanding and analysis of emotional speech.

The paper "Spoken Emotion Recognition Using Deep Learning"[4] explores the application of deep learning methods for recognition of emotions in spoken language. The study highlights the benefit of deep learning models in capturing and analyzing the complex features present in speech data, leading to improved emotion recognition performance.

The research showcases different deep learning methods, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid models, such as CNN-RNN combinations. These models have proven to be highly successful in automatically learning discriminative features from speech signals and accurately classifying emotions.

The findings of the study indicate that deep learning methods outperform traditional methods in terms of recognition accuracy, robustness, and generalization. The ability of deep learning models to capture

hierarchical representations and temporal dependencies in speech data contributes to their superior performance in recognizing and distinguishing different emotional states.

The paper also discusses the importance of large-scale labeled datasets for training deep learning models. Researchers have developed and utilized emotion-specific datasets to train and evaluate these models, enabling more reliable and comprehensive emotion recognition systems. The research highlights the potential applications of spoken emotion recognition using deep learning, including affective computing, human-robot interaction, and sentiment analysis. These applications benefit from the accurate identification and understanding of emotions conveyed through spoken language, leading to more personalized and responsive systems.

The paper demonstrates the effectiveness of deep learning techniques for spoken emotion recognition. The advancements in deep learning models have revolutionized the field by providing powerful tools to analyze and interpret the emotional content of speech. The findings contribute to the growing body of research in affective computing and pave the way for the development of more sophisticated and context-aware systems that can understand and respond to human emotions in spoken language.

The paper "Speech emotion recognition: Features and classification models"[5] provides an overview of the features and classification models employed in speech emotion recognition. The study emphasizes the importance of accurate emotion recognition from speech signals and highlights the advancements in feature extraction and classification techniques.

The research outlines various acoustic features used in speech emotion recognition, including prosodic features, spectral features, and cepstral features. These features capture different aspects of speech, such as pitch, intensity, spectral content, and voice quality, which are crucial in distinguishing various emotional states.

The paper discusses several classification models commonly utilized in speech emotion recognition, including traditional machine learning algorithms such as SVM, KNN and decision trees, as well as more advanced techniques like artificial neural networks (ANN) and deep learning models. These models leverage the extracted features to classify emotions accurately.

The findings of the study suggest that combining multiple features and employing ensemble models can improve the performance of speech emotion recognition systems. Additionally, the use of deep learning models, such as CNNs and RNNs, has shown promising results in capturing complex patterns and temporal dependencies within speech signals.

The paper also highlights the importance of labeled datasets for training and evaluating speech emotion recognition models. Datasets such as the Berlin Emotional Speech Database (Emo-DB) and the Emotional Prosody Speech and Transcripts (Emo-Prosody) database have facilitated the development and benchmarking of emotion recognition systems.

The paper emphasizes the significance of accurate speech emotion recognition and provides insights into the features and classification models employed in this domain. The advancements in feature extraction techniques, coupled with various classification algorithms, have contributed to improved emotion recognition performance.

The paper "Automatic speech emotion recognition using modulation spectral features"[6] focuses on the utilization of modulation spectral features for automatic speech emotion recognition. The study highlights the significance of capturing temporal variations and dynamic characteristics in speech signals to improve the accuracy of emotion recognition systems.

The research introduces modulation spectral features, which provide a comprehensive representation of the modulations present in speech signals. By analyzing the modulations across different frequency bands, the proposed approach captures the temporal dynamics associated with different emotional states.

The findings of the study demonstrate that modulation spectral features outperform traditional spectral features in emotion recognition tasks. The ability to capture the temporal variations and modulations enhances the discriminative power and robustness of the emotion recognition models.

The paper also explores the effectiveness of different classification algorithms in conjunction with modulation spectral features. Various machine learning algorithms, including support vector machines

(SVM), random forests, and artificial neural networks (ANN), are evaluated for their ability to classify emotions accurately using the proposed features.

The results indicate that combining modulation spectral features with appropriate classification models leads to improved performance in speech emotion recognition. The fusion of dynamic features and machine learning techniques enables the systems to capture nuanced emotional cues and accurately classify different emotional states.

The research presented in the paper has implications for applications such as affective computing, human-robot interaction, and emotion-aware systems. By leveraging modulation spectral features, these systems can better understand and respond to the emotional states of individuals, leading to more personalized and engaging interactions.

The paper emphasizes the effectiveness of modulation spectral features for automatic speech emotion recognition. The findings highlight the importance of capturing temporal dynamics and modulations in speech signals and demonstrate the superior performance of modulation spectral features compared to traditional spectral features.

II. METHODOLOGY:

Dataset:

This proposed work uses RAVDESS Dataset. This dataset contains 1440 files, recorded by 24 actors, uttering two lexically equivalent sentences with a neutral North American accent. This includes different classes of emotion sad, disgust, surprise, calm, happy, fearful, and angry. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression [8].

Data Visualization:

Visualization of data provides a greater understanding of the problem and a potential type of solution. Techniques for visualizing the data include classes distribution, instance counts inside each class, the distribution of the dataset, the connection between the characteristics, and dataset clustering. Data visualization functions are available in the Python and R languages.

Data Preparation:

It's time to get the data ready for processing once data analyzations are done using different visualizations. The procedures involve in data preparation include correcting difficulties with quality, standardization, and normalization. The data are initially checked for problems including missing values, same type of data, outliers, erroneous data. The dataset did not include any incorrect, identical, or incomplete data.

Data Augmentation:

Fresh synthetic data samples can be generated by adding minor changes to our training set, this technique known as data augmentation. Noise input, time altering, pitch and speed changes, and pitch and time switching can be used to provide syntactic data for audio. Making the model resistant to these changes will increase its generalizability. The labels from the initial training are preserved when adding the changes for this to work.

Feature Extraction:

Feature extraction is a critical step in researching and identifying connections between numerous things. Since it is already known that the provided audio data cannot be directly processed by the models, they must be converted into a format that can be easily understood, and feature extraction is done to do this. The three axes of the audio signal are time, amplitude, and frequency which represent its three dimensions. In this project, only 5 features are extracted and they are Chroma_stft, Zero Crossing Rate, MelSpectogram, MFCC and RMS (root mean square) value and to train the model.

Modelling:

Different combinations of parameters of Convolutional Neural Network (CNN) are introduced in this paper to train and test the model. The dataset is split into training and testing data in 3:1 ratio.

Journal of Information Systems Engineering and Management
2025, 10(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

III. SYSTEM DESIGN FLOWCHART:

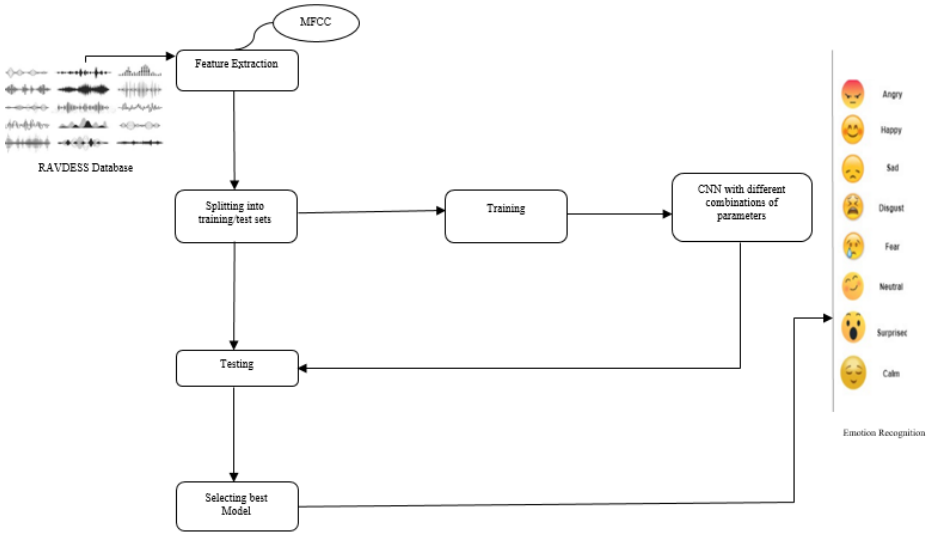


Fig.1: System Design Flowchart

IV. RESULT ANALYSIS:

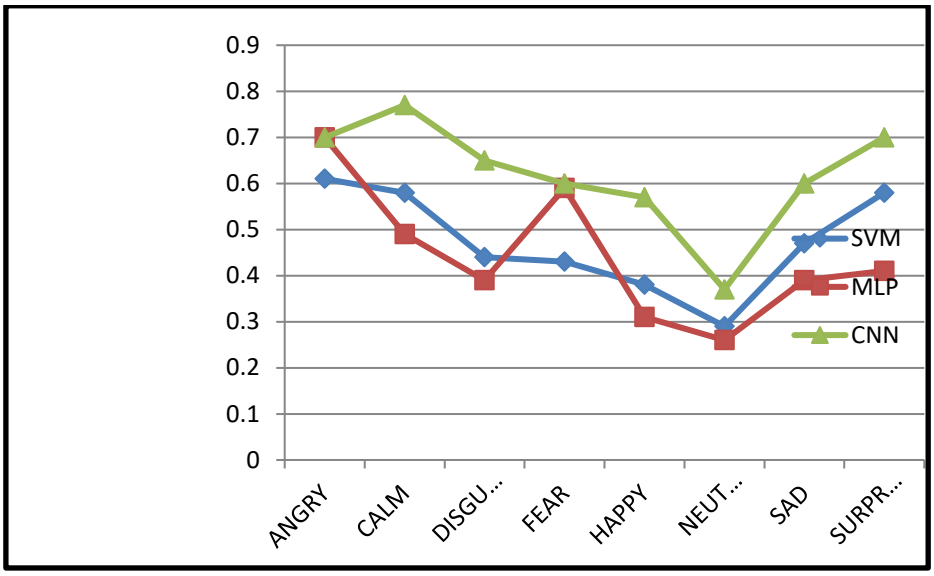
In this paper, Firstly, Different deep learning methods have been compared to find the best methods for speech emotion recognition. Table 1 shows accuracy on each of the emotion classes on different deep learning methods. As Table 2 shows that CNN performs the best on dataset, different combinations of parameters of Convolutional Neural Network (CNN) have been tested to find the best model combination, for RAVDESS Dataset. Table 3 shows accuracy on each of the emotion classes obtained by the base and the best CNN models. Table 4 shows the average accuracy from different combinations of parameter on CNN models. By comparing them, it has been found that, CNN 3 Layer model with RMSprop optimizer when trained with 80 Epochs works best for the RAVDESS dataset. Figure 2 shows the Confusion Matrix for best CNN model. Figure 3 shows the best CNN Model Loss and Accuracy plot against epochs for Training and Testing Data.

CLASS	SVM	MLP	CNN
ANGRY	0.61	0.70	0.70
CALM	0.58	0.49	0.77
DISGUST	0.44	0.39	0.65
FEAR	0.43	0.59	0.60
HAPPY	0.38	0.31	0.57
NEUTRAL	0.29	0.26	0.37
SAD	0.47	0.39	0.60
SURPRISE	0.58	0.41	0.70

Table 1: Accuracy on each of the emotion classes

Journal of Information Systems Engineering and Management
2025, 10(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

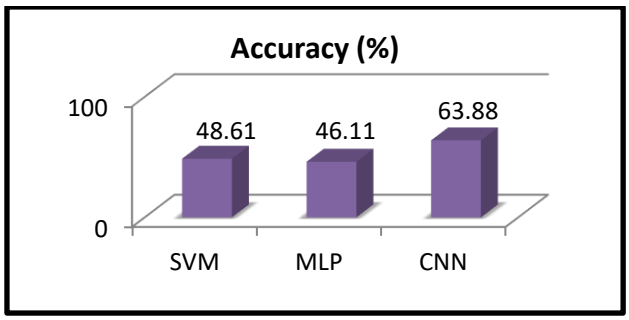
Research Article



Graph 1: Accuracy on each of the emotion classes

Classification Model	Accuracy (%)
SVM	48.61
MLP	46.11
CNN	63.88

Table 2: Accuracy of different classification models



Graph 2: Accuracy of different Classification models

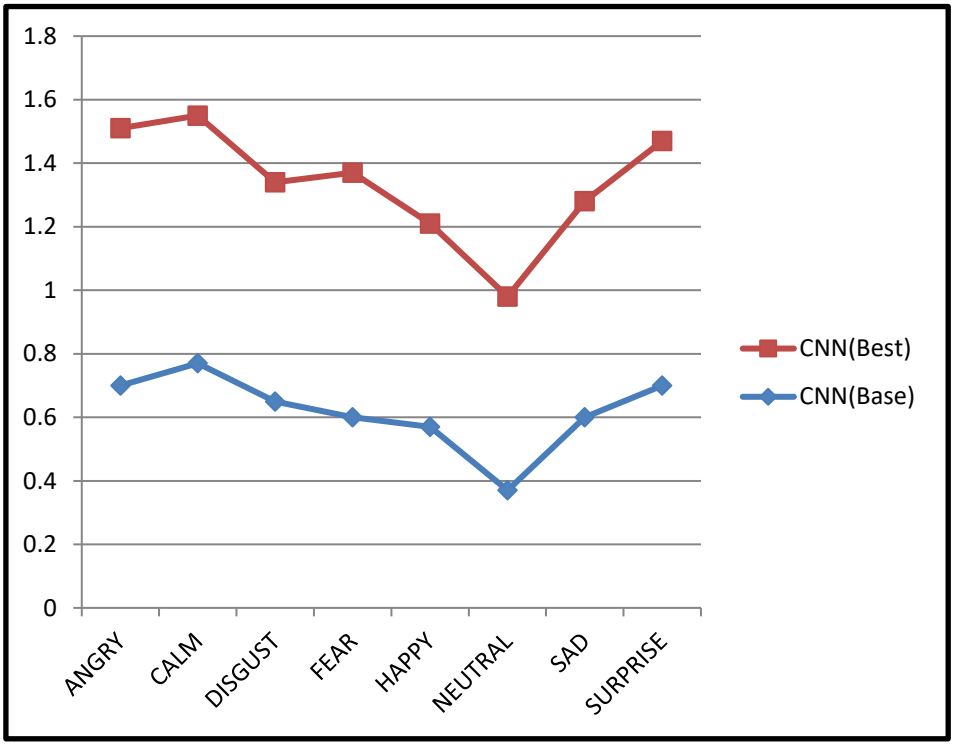
CLASS	CNN(Base)	CNN(Best)
ANGRY	0.70	0.81
CALM	0.77	0.78
DISGUST	0.65	0.69
FEAR	0.60	0.77
HAPPY	0.57	0.64

Journal of Information Systems Engineering and Management
2025, 10(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

NEUTRAL	0.37	0.61
SAD	0.60	0.68
SURPRISE	0.70	0.77

Table 3 Accuracy of different CNN models



Graph 3 Accuracy of different CNN models

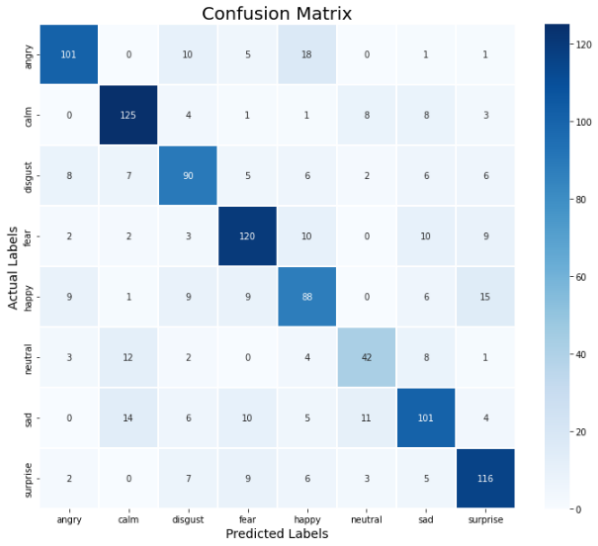


Fig.2: Confusion Matrix of best CNN Model

Journal of Information Systems Engineering and Management
2025, 10(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 4: Accuracy of different CNN model

CNN Hidden Layer	Optimizer	Epochs	Accuracy (%)
3	Adam	40	62
3	RMSprop	40	65
4	RMSprop	40	61
3	RMSprop	70	70
3	RMSprop	80	72
3	Adam	70	69

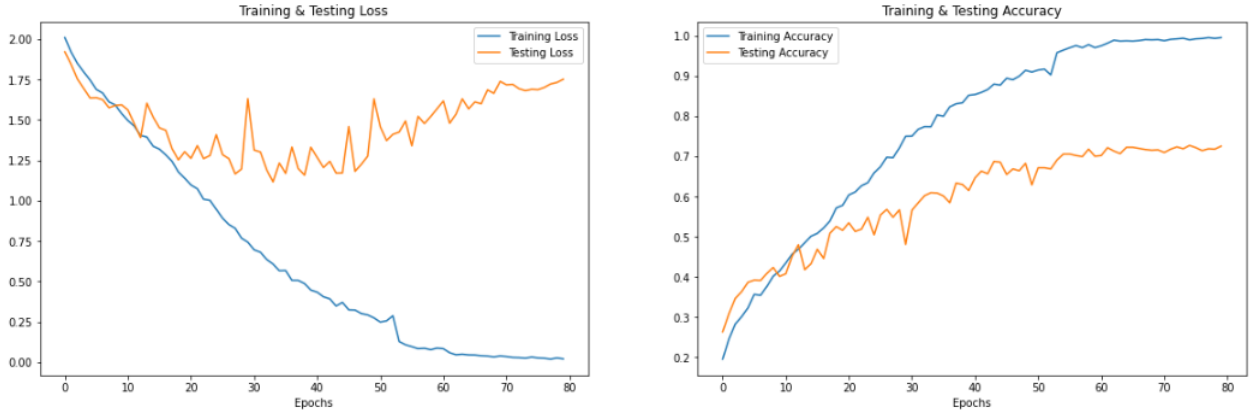


Fig.3: Best CNN Model Loss and Accuracy plot against epochs for Training and Testing Data

V.

VI. CONCLUSION:

This Paper, clearly shows through table 1 and graph 1 CNN has the best accuracy in detection of different class of emotion. From table 2 and graph 2 it can be observed that in CNN 63.8% accuracy has been obtained which is higher than SVM and MLP. But in this research higher accuracy for CNN has been targeted by using different combination of parameters like no of hidden layers, optimizer and Epochs. It has been shown that CNN 3 Layer model with RMSprop optimizer when trained with 80 Epochs works best for the RAVDESS dataset. This machine learning method can be used to extract emotion from human speech data accurately. This system can be useful for different fields like Call Centre for marketing or reporting, voice-based virtual assistants or chatbots etc. To improve the accuracy of this model different combinations of parameters in CNN model can be implemented.

VII. REFERENCES:

[1] W. Dai, D. Han, Y. Dai, and D. Xu, “Emotion Recognition and Affective Computing onVocal Social Media,” Inf. Manag., Feb. 2015.

[2] H. Cao, R. Verma, and A. Nenkova, “Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech,” Comput. Speech Lang., vol. 28, no. 1, pp. 186– 202, Jan. 2015.

[3] J. P. Arias, C. Busso, and N. B. Yoma, “Shape-based modeling of the fundamental frequency contour for emotion detection in speech,” Comput. Speech Lang., vol. 28, no. 1, pp. 278–294, Jan. 2014.

[4] Albornoz, E.M., Sánchez-Gutiérrez, M., Martinez-Licona, F., Rufiner, H.L., Goddard, J. (2014). Spoken Emotion Recognition Using Deep Learning. In: Bayro-Corrochano, E., Hancock, E. (eds) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2014. Lecture Notes in Computer Science, vol 8827. Springer, Cham.

[5] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, “Speech emotion recognition: Features and classification models,” Digit. Signal Process., vol. 22, no. 6, pp. 1154–1160, Dec. 2012.

[6] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, “Emotion recognition using a hierarchical binary decision tree approach,” Speech Commun., vol. 53, no. 9–10, pp. 1162– 1171, Nov. 2011.

[7] J.-H. Yeh, T.-L. Pao, C.-Y. Lin, Y.-W. Tsai, and Y.-T. Chen, “Segment-based emotion recognition from continuous Mandarin Chinese speech,” Comput. Human Behav., vol. 27, no. 5, pp. 1545–1552, Sep. 2011

[8] S. Wu, T. H. Falk, and W.-Y. Chan, “Automatic speech emotion recognition using modulation spectral features,” Speech Commun., vol. 53, no. 5, pp. 768–785, May 2011.

[9] C.-H. Wu and W.-B. Liang, “Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels,” IEEE Trans. Affect.Comput., vol. 2, no. 1, pp. 10–21, Jan. 2011.

[10] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, “Emotion recognition using a hierarchical binary decision tree approach,” Inter speech, vol. 53, pp. 320–323, 2009.

[11] J. Rong, G. Li, and Y.-P. P. Chen, “Acoustic feature selection for automatic emotion recognition from speech,” Inf. Process. Manag., vol. 45, no. 3, pp. 315–328, May 2009.

[12] M. M. H. El Ayadi, M. S. Kamel, and F. Karray, “Speech Emotion Recognition using Gaussian Mixture Vector Autoregressive Models,” in 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP ’07, 2007, vol. 4, pp. IV– 957–IV–960.25.

[13] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, “Primitives-based evaluation and estimation of emotions in speech,” Speech Commun., vol. 49, no. 10–11, pp. 787–800, Oct.2007.

[14] S. S. Narayanan, “Toward detecting emotions in spoken dialogs,” IEEE Trans. Speech Audio Process., vol. 13, no. 2, pp. 293–303, Mar. 2005.

[15] T. L. Nwe, S. W. Foo, and L. C. De Silva, “Speech emotion recognition using hidden Markov models,” Speech Commun., vol. 41, no. 4, pp. 603–623, Nov. 2003.

[16] Iqbal, Aseef & Barua, Kakon. (2019). A Real-time Emotion Recognition from Speech using Gradient Boosting. 1-5. 10.1109/ECACE.2019.8679271.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[17] Chaudhary AN, Sharma AK, Dalal JY, Choukiker LE. Speech emotion recognition. J Emerg Technol Innov Res. 2015;2(4):1169-71.

[18] Xu Dong An and Zhou Ruan 2021 J. Phys.: Conf. Ser. 1861 012064.

[19] Logan, Beth. (2000). Mel Frequency Cepstral Coefficients for Music Modeling. Proc. 1st Int. Symposium Music Information Retrieval.

[20] B. Schuller, G. Rigoll and M. Lang, "Hidden Markov model-based speech emotion recognition," 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)., Hong Kong, China, 2003, pp. II-1, doi: 10.1109/ICASSP.2003.1202279.

[21] Joshi, Aastha. “Speech Emotion Recognition Using Combined Features of HMM & SVM Algorithm.” (2013).

[22] Keshi Dai, Harriet J. Fell, and Joel MacAuslan. 2008. Recognizing emotion in speech using neural networks. In Proceedings of the IASTED International Conference on Telehealth/Assistive Technologies (Telehealth/AT '08). ACTA Press, USA, 31–36.

[23] Steven R. Livingstone, & Frank A. Russo. (2019). RAVDESS Emotional speech audio [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/256618.