Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Recognition Of Emotion Based On Speech Samples Using CNN Modelling

Akbar Ali [1], Vishal Sharma [2]

[1] *Research Scholar, Department of Computer Science and Engineering, Medi-Caps University, Indore*
akbar.ali@medicaps.ac.in
[2] *Assistant Professor, Department of Computer Science and Engineering, Medi-Caps University, Indore*
vishal.sharma@medicaps.ac.in

The objective of this study is to attempt to build an emotion recognition system through speech samples using deep learning techniques. Emotions are fundamental human trait, serving as a means of expressing thoughts and communicating intention. Emotion Recognition systems analyse audio signals to extract and predict the emotional state of a speaker. Emotions are generally classified as Anger, Happiness, Sadness, and Neutral. These systems rely on spectral and prosodic features to detect emotions. Mel-frequency Cepstral Coefficients (MFCC) are a significant spectral attribute, while prosodic attributes include frequency, loudness, and pitch. The frequency of an audio broadcast can be used to distinguish between various sounds and ascertain the gender of the speaker. The study shows that when Support Vector Machines (SVM) are used in Emotion Recognition to categorise and predict tasks, especially in identifying the speaker's gender. Emotions are identified utilising certain attributes through the utilisation of additional machine learning models such as Radial-Basis Function (RBF) and Back Propagation networks. The proposed model shows an accuracy of 72% reflecting reliability on CNN modelling.

Keywords: CNN, Speech Recognition, SVM

Introduction:

Emotions play a crucial role in all areas of human functions. They make an substantial impact on day to day performance. Usually, they are conveyed by spoken words, nonverbal signals, and bodily gestures. This is done by examining vocal patterns in speech. The speaker's voice output has an affect of physiological responses when they are experiencing various emotional states while speaking. For eg: Anger is related to increased muscle tension and change in breathing pattern. It affects the vocal folds, their oscillation, the architecture of the vocal tract, and the acoustic properties of speech.

Shahin et al [1] conducted research on identifying emotions in Arabic speech from Emirati speakers. The authors used a hybrid classifier termed cascaded GMM-DNN, which combines

Gaussian Mixture Models (GMM) and Deep Neural Networks (DNN). The study aimed to recognise emotions that are not dependent on specific texts or speakers.

Alonso et al [2] focuses on measuring the levels of activation in emotional speech. Some devices are specifically designed to detect alterations in the speaker's emotional state. This study uses activation recognition to monitor the emotional states of users by differentiating between two categories: high activation, represented by anger and happiness movements, and low activation, characterised by neutral, dull, and sad emotions. This study presents an innovative approach that involves dividing the data into segments based on time and then recognising the emotions associated with each segment. This methodology deviates from the conventional method employed in prior research, which believed that each examined phrase only encompassed a single emotional state. Temporal bifurcation allows for the identification of changes in emotions to continuously assess the speech stream. The proposed feature set comprises six features, including two prosodic features and four paralinguistic features. These features possess qualities of being strong, simple, and able to compute quickly, which makes them ideal for creating applications that operate in real-time. The results exhibited the efficacy of distinguishing between high and low degrees of activation in a corpus that was performed. This was accomplished using a hierarchical segmentation, feature extraction, and classification system. The approach created a metric called Emotional Temperature to measure the level of high or low activation. When compared to previous comparable studies, Emotional Temperature has exhibited a degree of accuracy that is equivalent to the overall research, while capitalising on the utilisation of a more compact and uncomplicated set of characteristics. Nevertheless, the primary endeavour to quantify activation has centred on examining spoken language in connection with gender, uncovering a diminished precision rate in female speech and a more pronounced variability in precision rate across various languages for this gender.

Gjoreski et al. [3] evaluated the Support Vector Machine (SVM) method trying the Leave-One-Subject-Out (LOSO) strategy. They performed a comparative analysis of the accuracy of the standard Support Vector Machine (SVM) with its default parameters and the enhanced SVM using Auto-WEKA. The results indicated that the Support Vector Machine (SVM) performed better when combined with Auto-WEKA, resulting in significantly higher accuracy compared to the traditional SVM.

Dai et al. [4] suggested that the two main ways to address data scarcity are to obtain more data or build technologies to handle small data sets. ERFTrees, a new feature selection method, extracts valuable features from small data sets. There are two benefits to employing this emotion recognition system. First, it removes irrelevant data and reduces training data dimensionality. Second, by lowering the input set, most machine learning algorithms that struggle with small data sets can improve recognition accuracy. Using Chinese (Mandarin) emotional data sets, this study found that the algorithm proposed in this study reduces dimensionality better than linear and non-linear methods like PCA, MDS, and ISOMap.

Subramanian et al. [5] created a Tamil language dataset for emotion recognition and was tested using traditional machine learning and deep learning classifiers on the Tamil Emotional dataset. The suggested TFS approach outperforms Indian English and Malayalam datasets with 97.96% accuracy, making it better for regional languages. A newly trained framework for speech emotion classification has been developed. The proposed framework primarily employs a preprocessing phase to mitigate the background noise and distortions in the input voice signal. Subsequently, the two recently introduced speech attributes pertaining to energy and phase have been incorporated with cutting-edge attributes for assessing the characteristics of speech emotion. The TFS algorithm has been developed to identify the most suitable features by utilising a statistical method.

Wu et al. [6] introduced a new method for person re-identification that utilises an improved deep feature representation. The authors suggest a system that utilises deep learning approaches to enhance the resilience and precision of identifying humans across various camera perspectives. The approach emphasises the extraction of deep features that remain unchanged despite alterations in illumination, position, and background clutter. The efficacy of the proposed technique is verified on multiple benchmark datasets, showcasing substantial enhancements compared to current methodologies. The paper enhances the field by providing a more dependable and effective answer to the complex issue of person re-identification.

Lee et al. [7] proposed a hierarchical binary decision tree method for the purpose of emotion identification. This approach breaks down the challenge of recognising emotions in multiple classes into a sequence of binary judgements, which enables a categorization process that is more organised and easier to understand. The decision tree is constructed using acoustic information derived from speech signals. The experimental findings suggest that the hierarchical technique enhances the accuracy of classification in comparison to flat multi-class classifiers. The paper makes a valuable contribution to the area by presenting an approach that is both scalable and successful for recognising emotions. This method has the potential to improve applications in affective computing and automated emotional analysis.

Basu et al. [8] conducted a comprehensive analysis of the progress made in speech-based emotion recognition. Their evaluation specifically examined the many techniques and methodology utilised in this area of research. The paper examines several techniques for extracting features, algorithms for classification, and the use of multiple databases for training and testing emotion detection systems. The review emphasises the advancements achieved in enhancing the precision and resilience of speech emotion recognition. It also acknowledges the persisting obstacles, including handling noisy data and addressing cross-cultural disparities in emotional display. This study offers a comprehensive analysis of the present patterns and future possibilities in the field of speech-based emotion identification.
Akçay et al. [9] covered various aspects of Emotion Recognition, including emotional models, databases, feature extraction techniques, pre-processing methods, supporting modalities, and classifiers. The authors present the current cutting-edge techniques in SER, emphasising the advantages and drawbacks of various methodologies. The analysis also examines potential areas for future research and the obstacles that must be overcome to progress in the subject.

This document provides a significant reference for scholars and practitioners by summarising the essential elements and approaches employed in SER, offering guidance for future advancements and enhancements.

Swain et al. [10] conducted a study on several databases, methods for extracting features, and classifiers that are employed in the field of voice emotion recognition. The research highlights the need of choosing suitable databases that provide a broad spectrum of emotional expressions and circumstances. The text also explores many characteristics, including prosodic, spectral, and wavelet-based characteristics, and their influence on the performance of emotion recognition. The authors conduct a comparative analysis of different classifiers, encompassing both conventional machine learning algorithms and deep learning methods, emphasising their respective advantages and drawbacks. This review offers significant insights into the essential elements of speech emotion recognition systems and provides guidance to researchers in making well-informed judgements in their studies.

Caldognetto et al. [11] examined the impact of emotional states on the articulation of speech sounds, with a specific emphasis on the pronunciation of labial sounds. The authors analyse the co-production of speech and emotions by studying the changes in phonetic targets throughout the expression of different emotions. The study demonstrates that emotions have a notable impact on the way labial sounds are produced, resulting in changes in the specific speech targets. These findings emphasise the significance of considering the emotional context in investigations related to phonetic and speech processing. The research enhances our comprehension of the interaction between emotion and speech output, providing valuable insights into the development of more precise and lifelike voice synthesis and recognition systems.

In their research, EI Ayadi et al. [12] examined different techniques for extracting features, classification systems, and databases that are often used in the field of Emotion Recognition. The efficiency of key features, including prosodic, spectral, and wavelet-based features, is assessed in the context of emotion recognition. The survey encompasses many categorization algorithms, encompassing both conventional machine learning techniques and more recent methodologies such as support vector machines and neural networks. In addition, the paper examines current databases that are frequently utilised for the purpose of training and assessing SER systems. This survey provides academics with a valuable resource by summarising the essential elements and approaches in SER, so offering guidance for future research and development.

Grimm et al. [13] presented the Vera Ammittag German audio-visual emotional speech database, which was created to facilitate studies on emotion recognition. The collection comprises audio-visual recordings of German speakers exhibiting diverse emotions, offering a comprehensive dataset for the development and evaluation of emotion identification algorithms. The study provides a comprehensive description of the database's design, collecting, and annotation methods, with a particular focus on its usefulness in research on recognising emotions using many modes of communication. The database is proven to be a

helpful instrument for developing the profession, providing a comprehensive resource that collects both auditory and visual indicators of emotional expression. This effort makes a substantial contribution to the availability of top-notch datasets for research on emotion recognition.

Jeon et al. [14] examine the difficulties and potential of recognising emotions in speech across different languages. They do this by comparing automated classification algorithms with human perception. The authors examine the efficacy of emotion detection systems when applied to speech in several languages and explore the extent to which these systems can exhibit cross-linguistic generalisation. The study reveals that although automatic categorization systems demonstrate potential, they frequently fail to match human performance, particularly in cross-lingual situations. These findings emphasise the necessity for emotion recognition models that are more resilient and flexible, capable of accommodating language variation. The research highlights the significance of considering linguistic and cultural disparities while creating emotion identification systems.

Lin and Wei [15] propose a combined method for voice emotion recognition that utilises Hidden Markov Models (HMM) and Support Vector Machines (SVM). The authors suggest a system that integrates the temporal modelling skills of HMM with the classification prowess of SVM to enhance the accuracy of emotion recognition. The study provides a comprehensive description of the feature extraction methodology, model training, and evaluation methodologies employed in their approach. The experimental results indicate that the hybrid HMM-SVM system surpasses conventional approaches, providing superior recognition rates for different emotional states. This study enhances the progress of emotion identification technology by offering a more efficient and dependable approach for categorising emotions in speech.

Current Speech Emotion Recognition systems have several identified limitations: The model requires extensive pre-processing to understand the input audio signal, which lacks consistency. The Speech Emotion Recognition system cannot comprehend audio files of different lengths, making them impractical and costly for commercial applications. The entire system remains unchanged. These devices are impractical for real-world use due to their insufficient performance in real-world situations, and they are not easily customisable due to their rigid design.

Proposed System and Methodology:

The proposed system is designed to preprocess audio data, extract relevant features, and train a machine learning model for classification tasks. The process begins with the collection and loading of audio data, which is stored in a CSV file and subsequently loaded into a Pandas DataFrame for exploration and preprocessing. We have taken RAVDESS dataset for our analysis. During the preprocessing phase, any missing values in the dataset are handled by dropping rows containing NaN values. Categorical labels are then encoded using the LabelEncoder to convert them into numerical format. The dataset is split into features (X) and

target labels (y), and the features are standardized using a StandardScaler to ensure they have zero mean and unit variance.

Feature extraction is a critical step in this system, where librosa, a library for audio and music analysis, is used to extract key audio features. Common features such as Mel-frequency cepstral coefficients (MFCCs), chroma features, and spectral contrast are computed for each audio recording. These extracted features form a feature matrix that serves as the input for the machine learning model.

The model development phase involves building a Sequential model using TensorFlow and Keras. The model architecture consists of layers like Conv1D for capturing temporal patterns in the audio data, Flatten for converting the 3D output from Conv1D to 1D, and Dense layers for classification. The model is compiled with a suitable loss function, such as categorical cross-entropy, and an optimizer like Adam, which helps in adjusting the weights to minimize the loss function during training.

For model training and evaluation, the dataset is split into training and testing sets using the train_test_split function. The model is trained on the training set while monitoring its performance on a validation set to prevent overfitting. After training, the model is evaluated on the testing set to assess its performance using metrics like accuracy, precision, and recall. If necessary, hyperparameter tuning is performed to enhance the model's performance by adjusting parameters such as the learning rate, number of epochs, and batch size.

Once the model demonstrates satisfactory performance, it is deployed to a production environment where it can be used to classify new audio recordings. Continuous monitoring of the model's performance is essential to ensure it maintains its accuracy over time. Periodically, the model may be retrained with new data to adapt to any changes in the audio characteristics or the classification requirements.

This systematic approach, encompassing data collection, preprocessing, feature extraction, model development, training, evaluation, and deployment, ensures the development of a robust audio-based machine learning model capable of accurately classifying audio recordings.

Implementation:

Data Preparation

We began by loading the RAVDESS dataset, which consists of speech audio files labelled with various emotions. To prepare the data for input into our Convolutional Neural Network (CNN), we first extracted Mel-Frequency Cepstral Coefficients (MFCC) features from the audio files. These MFCC features provide a compact representation of the audio signals, capturing the essential characteristics while reducing the dimensionality of the data.

Subsequently, we converted these MFCC features into 2-D log Mel-spectrograms, which serve as the input for our CNN model.

Feature Extraction

In our approach, we utilized MFCC to efficiently reduce the dimensionality of the dataset, facilitating quicker processing and improved model performance. By focusing on MFCC features, we were able to capture the critical aspects of the audio signals relevant to emotion recognition while minimizing the computational burden.

CNN Model Development

We designed a CNN architecture specifically tailored to process the 2-D log Mel-spectrograms generated from the MFCC features. The CNN model comprises several convolutional layers, followed by pooling layers and fully connected layers. This architecture is adept at capturing spatial hierarchies in the spectrograms, enabling effective learning of the underlying emotional patterns. We then trained the CNN model on the extracted features and evaluated its performance on a separate validation set to ensure its generalizability.

Short Pre-processing Steps

To enhance the model's ability to understand the input audio signal, we incorporated short and computationally inexpensive pre-processing steps. These steps, which included normalization and augmentation techniques, were essential in preparing the data for effective training and real-world application. Our pre-processing approach ensured that the system remained dynamic and adaptable, capable of handling non-static audio signals encountered in practical scenarios.

We implemented the entire pipeline, from data loading to model evaluation, ensuring a seamless workflow. The pipeline begins with loading the RAVDESS dataset, followed by extracting MFCC features and generating 2-D log Mel-spectrograms. These spectrograms are then fed into our CNN model for training. Throughout the implementation, we focused on maintaining the efficiency and scalability of the system, allowing for easy upgrades and configurations as needed. Our findings demonstrate that the combination of MFCC feature extraction and CNN modeling yields robust performance, achieving an average accuracy of about 72% in speech emotion recognition tasks. This approach underscores the potential of machine learning techniques in extracting and recognizing emotions from speech audio signals, providing valuable insights into verbal emotion conveyance.

Results:

The provided graph shows the accuracy and validation accuracy of a machine learning model over 20 epochs of training.

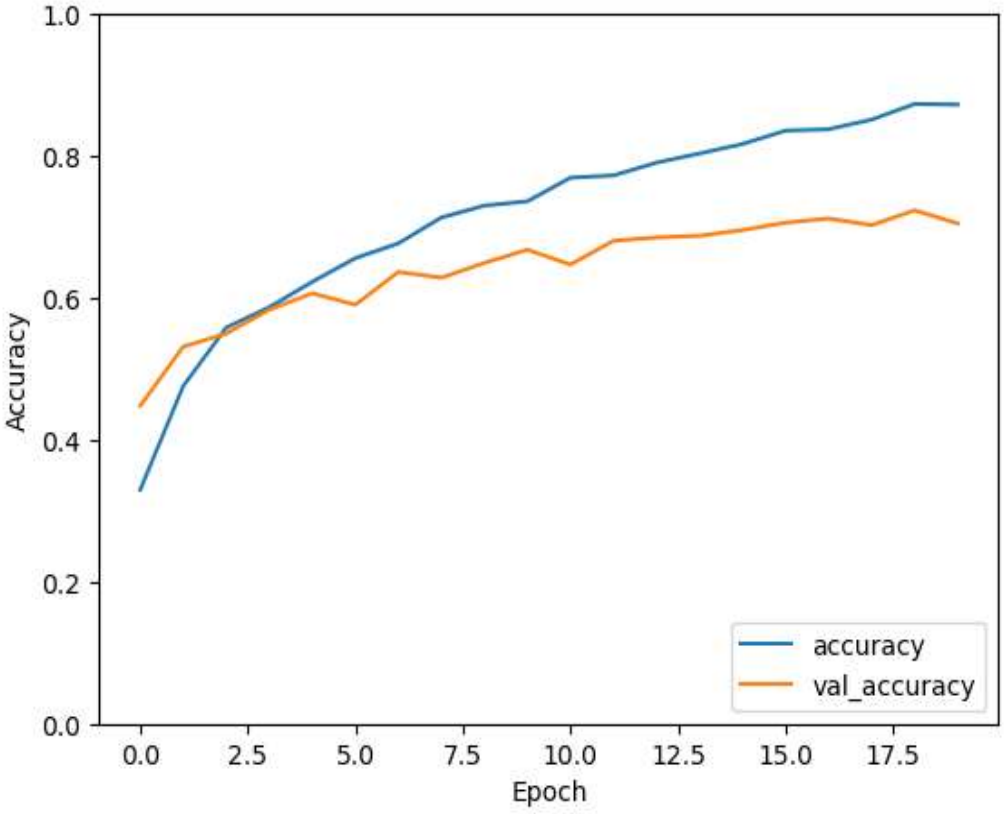

Fig 1: Training and Validation Accuracy

The graph shows that the training accuracy starts at a low value around 0.24 in the first epoch and increases steadily over the epochs. By the 20th epoch, the training accuracy reaches approximately 0.88, indicating that the model is learning and improving its performance on the training dataset.

The validation accuracy starts slightly higher than the training accuracy, around 0.43 in the first epoch, which is common if the validation set is easier or if the model is initially overfitting. It increases initially, reaching a plateau around the 7th epoch at about 0.61. After the 7th epoch, the validation accuracy shows fluctuations but generally remains within the range of 0.61 to 0.72. By the 20th epoch, the validation accuracy is around 0.72, which is the final value indicated.

Further the recommendation performance is figured out with the help of a confusion matrix. The provided image is a confusion matrix that illustrates the performance of a classification model on different categories. Each row represents the true labels, and each column represents

the predicted labels. The values in the matrix indicate the number of instances for each true-predicted label pair.

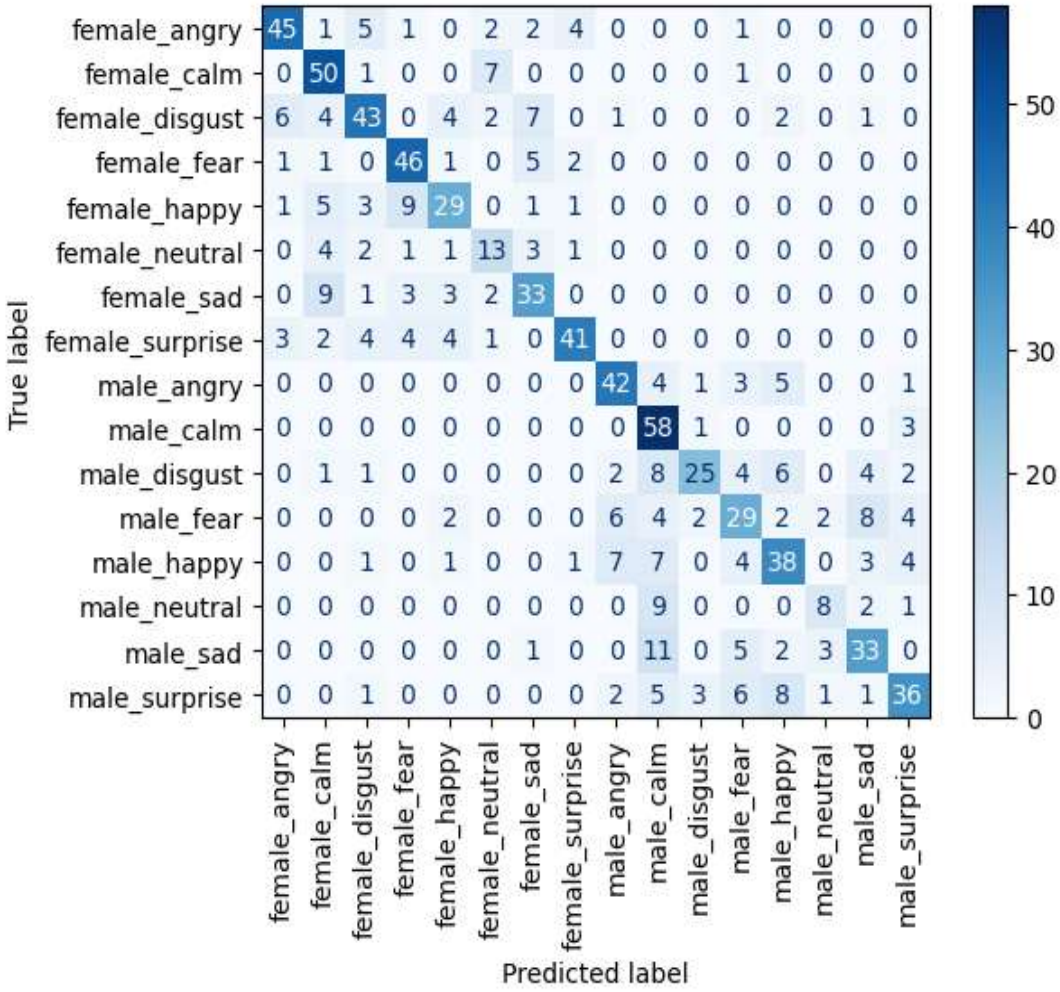

Fig 2: Confusion Matrix

The diagonal elements represent the number of correctly classified instances for each category while Off-diagonal elements indicate misclassifications. Categories like "female calm," "female fear," and "male surprise" show high accuracy with minimal misclassifications. Some categories, especially those with subtle differences like "female neutral" vs. "female sad" or "male happy" vs. "male neutral," show higher misclassification rates. There are fewer instances where female emotions are misclassified as male emotions and vice versa, indicating the model is somewhat effective in distinguishing between genders.

The model can be improved by improving the model's ability to differentiate between similar emotions by using techniques such as data augmentation, feature engineering, or more complex models. If certain categories have fewer samples, consider collecting more data to balance the dataset, which can help in reducing misclassification. Cross-validation is done to to ensure the model's performance is consistent across different subsets of the data.

The provided bar chart compares the accuracy of an existing system and a proposed system across various emotions. Each emotion has two bars: one for the existing system (blue) and one for the proposed system (orange). The proposed system shows significant improvements in the accuracy for detecting "female fear" (+0.57), "female sad" (+0.59), "female happy" (+0.29), and "male fear" (+0.38). There are minor improvements in categories like "female calm" (+0.02) and "male surprise" (+0.03). The proposed system shows decreased performance in categories like "female disgust" (-0.07), "male calm" (-0.54), "male happy" (-0.19), and "male neutral" (-0.19). For "female angry" and "female neutral," the proposed system shows a moderate improvement, indicating consistent performance enhancements.

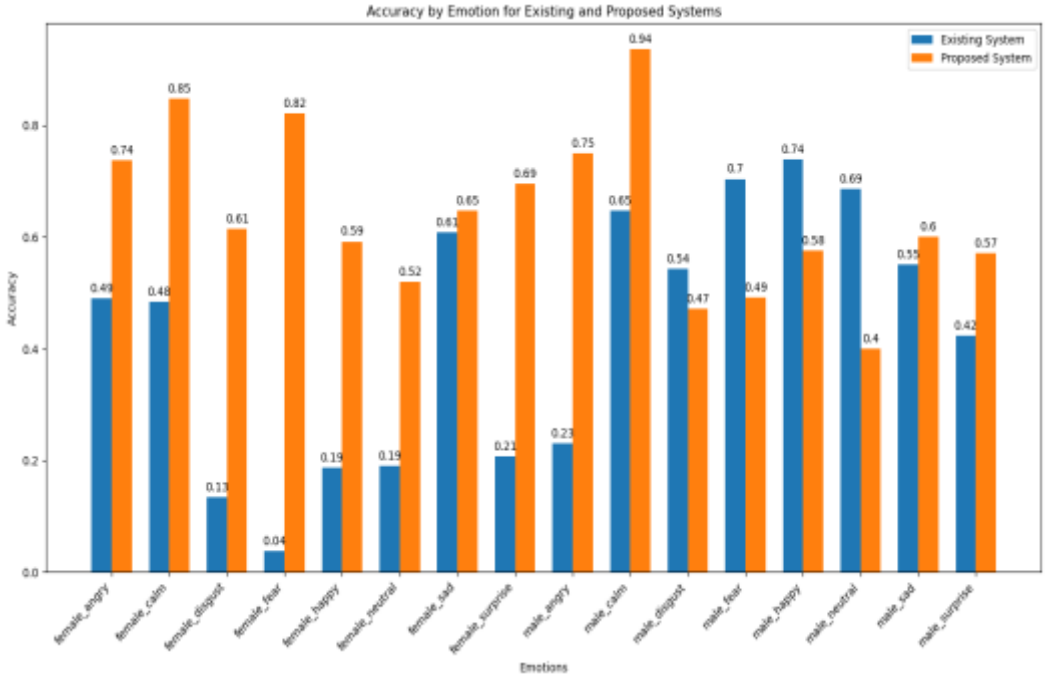

Fig 3: Accuracy of Existing and Proposed System

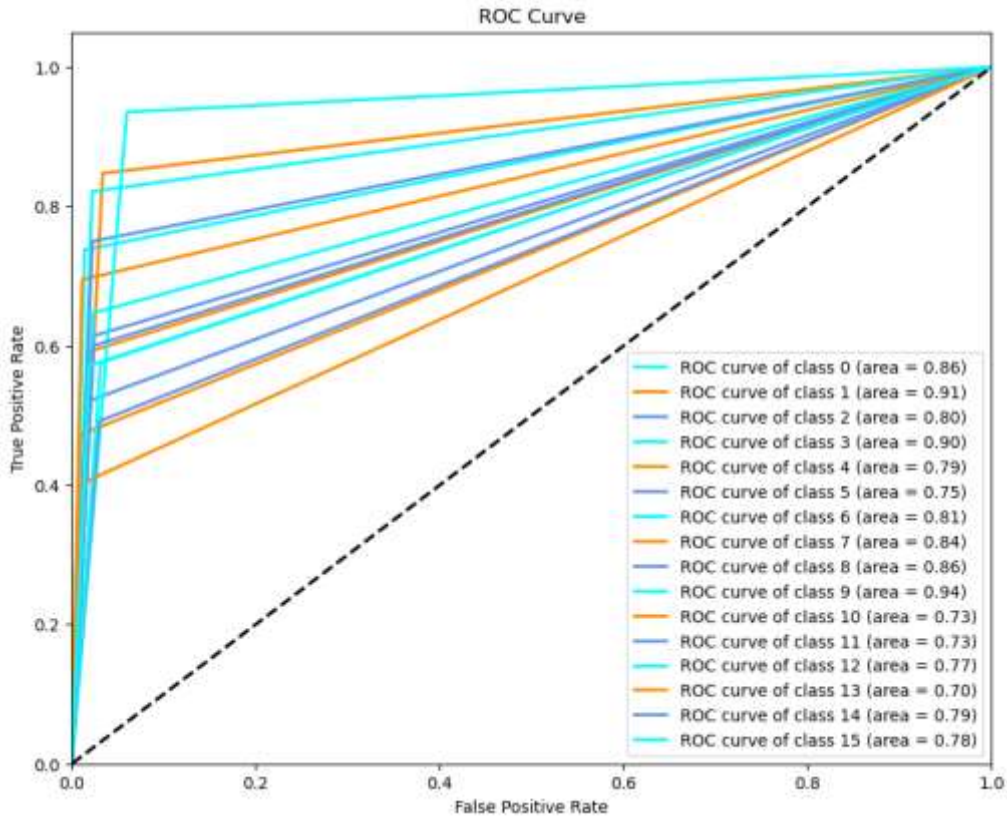

Fig 4: ROC Curve

The figure shows a Receiver Operating Characteristic (ROC) curve plot for a multi-class classification problem. It shows that the y-axis represents the True Positive Rate (TPR), also known as sensitivity or recall. The x-axis represents the False Positive Rate (FPR), which is 1-specificity. The diagonal dashed line from (0,0) to (1,1) represents a random classifier. Any classifier with curves above this line performs better than random guessing. Each ROC curve represents the performance of the classifier for a specific class. The plot shows curves for 16 different classes, with each curve labelled accordingly.

The legend includes the Area Under the Curve (AUC) for each class's ROC curve. A higher AUC indicates better performance. The Best Performing Class is Class 9 i.e. “male angry”, with an AUC of 0.94. Similarly the Worst Performing Class is Class 13 i.e. “male normal”, with an AUC of 0.70. Most classes have AUC values above 0.7, indicating good performance. Classes 1, 3, and 9 have especially high AUC values above 0.9, showing excellent classification performance for these classes.

High AUC (close to 1) shows that the model has a good measure of separability. It can distinguish between the positive class and the negative class effectively. Low AUC (close to 0.5) shows that the model has no class separation capacity, equivalent to random guessing.

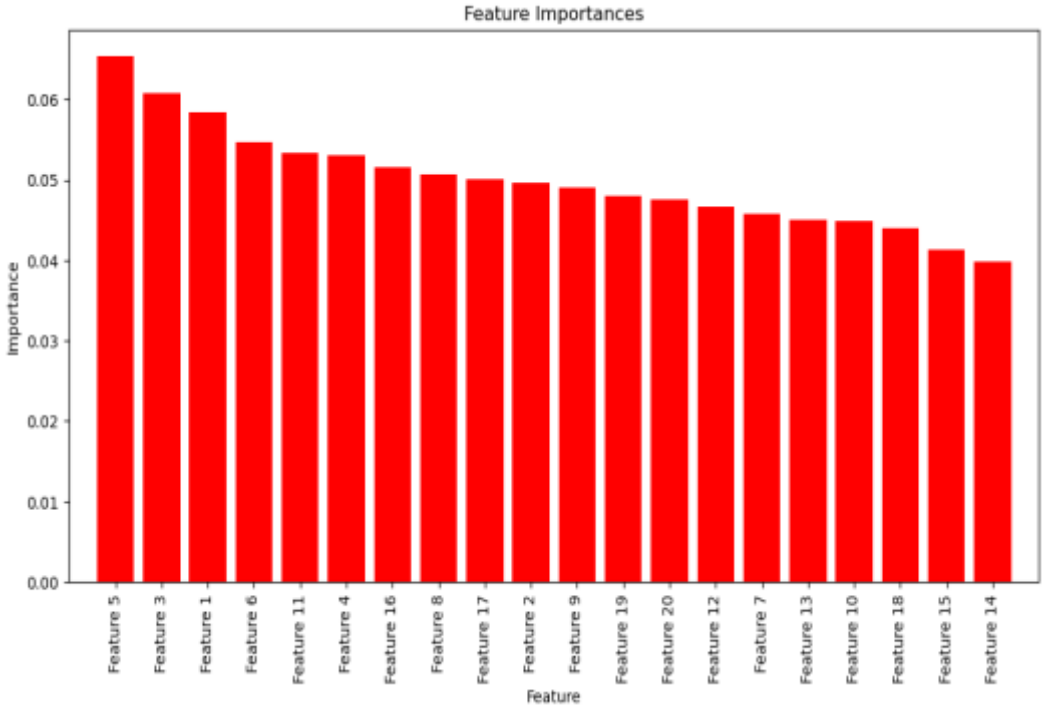

Fig 5: Feature Importances

Feature Importance

This measures the contribution of each feature to the model's predictions. Higher values indicate more important features. The ordinates show importance of each feature while the abscissa shows different features, labelled as Feature 1, Feature 2, etc. Feature 5 has the highest importance, around 0.065, along with Feature 3 and Feature 1, which also have high importance values, above 0.055. The Least Important Feature is Feature 14, which has the lowest importance, around 0.035. These features significantly impact the model's predictions and should be carefully considered in analysis and further model refinement.

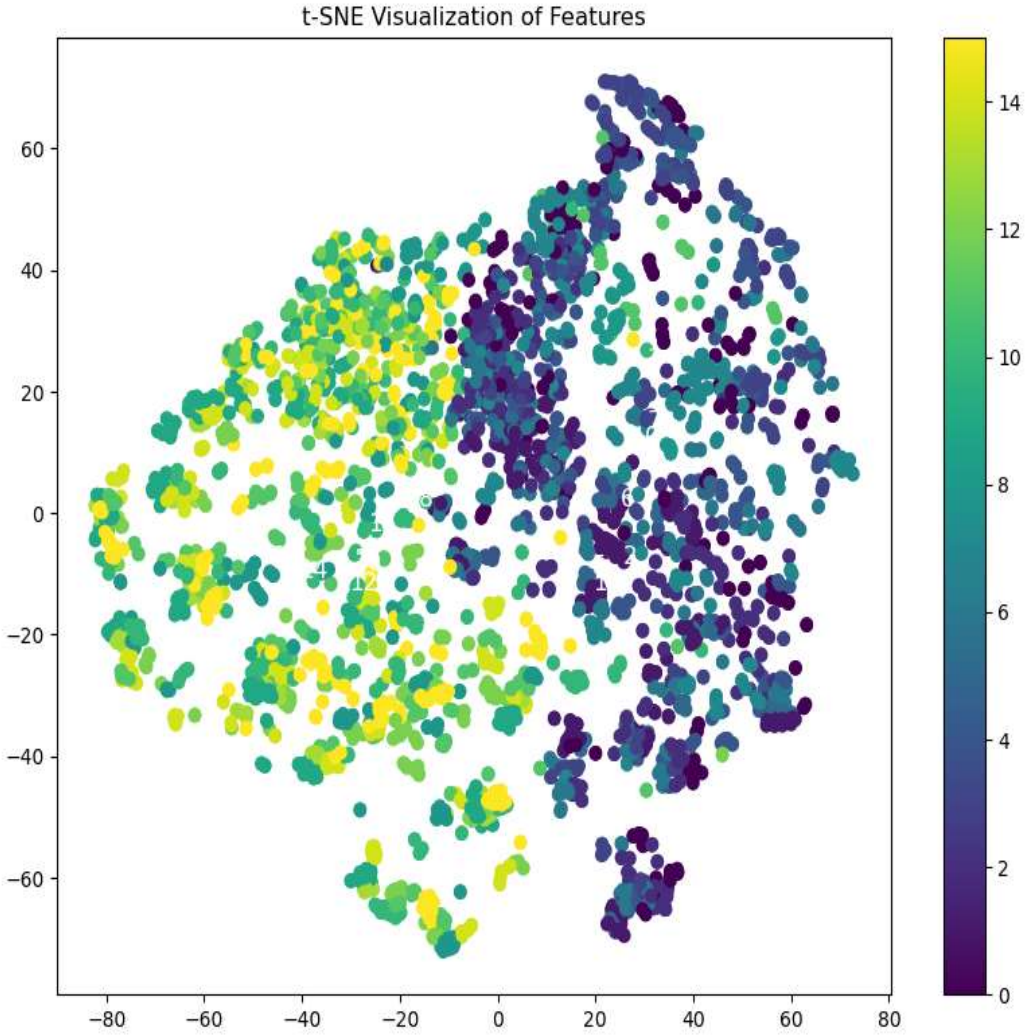

Fig 6: t-SNE Visualization Curve

Table 1: Performance of various predictions

Emotion	Precision	Recall	F1-Score	Support
Female Angry	0.80	0.74	0.77	61
Female Calm	0.65	0.85	0.74	59

Female Disgust	0.69	0.61	0.65	70
Female Fear	0.72	0.82	0.77	56
Female Happy	0.64	0.59	0.62	49
Female Neutral	0.48	0.65	0.55	40
Female Sad	0.63	0.65	0.64	51
Female Surprise	0.82	0.69	0.75	59
Male Angry	0.70	0.75	0.72	73
Male Calm	0.55	0.94	0.69	66
Male Disgust	0.78	0.47	0.59	53
Male Fear	0.55	0.49	0.52	59
Male Happy	0.60	0.58	0.59	66
Male Neutral	0.57	0.40	0.47	25
Male Sad	0.63	0.60	0.61	50
Male Surprise	0.71	0.57	0.63	63

This table summarizes the precision, recall, F1-score, and support for each emotion category.

MFCC shape after extraction: (20, 110)

MFCC shape after normalization: (20, 110)

MFCC shape after transpose: (110, 20)

MFCC shape after padding/trimming: (20, 20)

MFCC shape after reshaping: (1, 20, 20)

MFCC shape after expanding: (1, 20, 1)

Conclusion:
Machine learning and deep learning techniques build a reliable model for identifying emotion based on speech samples. The model shows a significant improvement in both training and validation accuracy from the start to the end of training, indicating effective learning. There is a noticeable gap between training and validation accuracy from around the 6th epoch onward. The training accuracy continues to rise, while the validation accuracy plateaus and fluctuates. This suggests that the model might be starting to overfit the training data, learning patterns that do not generalize well to the validation data. The final validation accuracy of 0.72 suggests that the model has a reasonably good performance but still has room for improvement. Further steps like regularization, tuning hyperparameters, or using more data could potentially enhance its performance. Following steps are used for improving the model:

1. **Regularization**: Techniques such as dropout, L2 regularization, or early stopping can help mitigate overfitting.

2. **Hyperparameter Tuning**: Adjusting learning rates, batch sizes, or network architecture might improve performance.

3. **Cross-Validation**: Using cross-validation can provide a more robust estimate of the model's performance and stability.

There is a scope for investigating the reasons for decreased performance in categories like "male calm," "male happy," and "male neutral." It could be due to overfitting, underfitting, or issues in the training data. It can be ensured by checking whether the dataset is balanced across all emotions to provide the model with sufficient examples of each category.

Conflict of Interest:
The authors confirm that there is no conflict of interest.

Data Availability Statement
The authors confirm that the data supporting the findings of this study are available within the article or its supplementary materials.

Declaration of generative AI and AI-assisted technologies in the writing process
During the preparation of this work the author(s) used Quillbot in order to frame the sentences in sophisticated manner. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References:

1) Shahin A, Nassif B, Hamsa S. Emotion Recognition Using Hybrid Gaussian Mixture Model and Deep Neural Network. IEEE Access 2019;7:26777–87. https://doi.org/10.1109/ACCESS.2019.2901352.

2) J. B. Alonso, J. Cabrera, M. Medina and C. M. Travieso, New approach in quantification of emotional intensity from the speech signal: Emotional temperature, Expert Syst. Appl., vol. 42, pp. 9554-9564, Dec. 2015.
3) Gjoreski M, Gjoreski H, Kulakov A., Machine Learning Approach for Emotion Recognition In Speech, Informatica., vol. 37, pp. 501-505, 2013.
4) Dai W, Han D, Dai Y, Xu D. Emotion Recognition and Affective Computing on Vocal Social Media. Inf Manag 2015;52(7):777–88.
5) Subramanian, R., Aruchamy, P. An Effective Speech Emotion Recognition Model for Multi-Regional Languages Using Threshold-based Feature Selection Algorithm. Circuits Syst Signal Process **43**, 2477–2506 (2024). https://doi.org/10.1007/s00034-023-02571-4.
6) S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, W.-S. Zheng, an enhanced deep feature representation for person re-identification, in: Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on, IEEE, 2016, pp. 1–8.
7) S Lee C-C, Mower E, Busso C, Lee S, Narayanan S. Emotion recognition using a hierarchical binary decision tree approach. Interspeech 2009;53:320–3.
8) Basu, S.; Chakraborty, J.; Bag, A.; Aftabuddin, M. ‘‘A Review on Emotion Recognition using Speech”. In Proceedings of the International Conference on Inventive Communication and Computational Technologies (ICICCT 2017), Coimbatore, India, 10–11 March 2017.
9) Akçay, Mehmet, B.; O˘ guz, K. ‘‘Speech emotion recognition: Emotional models, databases, features, pre-processing methods, supporting modalities, and classifiers”. Speech Commun. 2020, 166, 56–76.
10) Swain M, Routray A, Kabisatpathy P. Databases, features and classifiers for speech emotion recognition: a review. Int J Speech Technol 2018;21:93–120. https://doi.org/10.1007/s10772-018-9491-z.
11) Caldognetto, E. M., Cosi, P., Drioli, C., Tisato, G., & Cavicchio, F. (2004). Modifications of phonetic labial targets in emotive speech: Effects of the co-production of speech and emotions. Speech Communication, 44, 173–185.
12) EI Ayadi M, Kamel MS, Karray F (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 1(44), 572–587.
13) Grimm, M., Kroschel, K., & Narayanan, S. (2008). The Vera Ammittag German audio-visual emotional speech database. In International conference on multimedia and expo, pp. 865–868.
14) Jeon, J. H., Le, D., Xia, R., & Liu, Y. (2013). A preliminary study of cross-lingual emotion recognition from speech: Automatic classification versus human perception. In Interspeech, Layon, France, pp. 2837–2840.
15) Lin, Y.-L., & Wei, G. (2005). Speech emotion recognition based on HMM and SVM. In: Fourth International conference on machine learning and cybernetics, Guangzhou, pp. 4898–4901.

Prof. Akbar Ali is currently working as Assistant Professor in Department of Mechanical Engineering, Medi-Caps University. With academic experience of over 17+ years, he has done his research in Solar desalination pertaining to the field of Thermal Engineering, in he was specialized and completed his Masters in 2012. Prof. Ali completed his graduation in BE Mechanical Engineering in 2007 from Jawaharlal Institute of Technology Borawan. He has authored a text book on Heat Transfer, and published and presented many research papers. He

also works in the domain of Data Science and Machine Learning and currently pursuing PG in Artificial Intelligence.

Prof Vishal Sharma is currently working as Assistant Professor in Department of Computer Science and Engineering, Medi-Caps University. With academic experience of 15+ years, he is currently pursuing his research from VIT, Bhopal. His areas of expertise are Machine Learning, NLP and Blockchain Architecture.