# Comparative Analysis of Machine Learning Models for Emotion Classification in Speech Data

N. Siva[1], B. Venkata Sivaiah[2*], G. Sai Kumar[3], G. Jaya vardhan Raju[4],
V. Sushvitha[5], G. Chaithanya[6,] Sam Goundar[7]

[1] Assistant Professor, Department of Artificial Intelligence and Machine Learning,
Annamacharya Institute of Technology and Sciences, Rajampet.
[2] Assistant Professor, Department of Data Science, Mohan Babu University
(Erstwhile Sree Vidyanikethan Engineering College), Tirupati.
[3, 4, 5, 6] UG Scholar, Department of Computer Science and Systems Engineering,
Sree Vidyankethan Engineering College, Tirupati, India.
[7]RMIT University, Australia
[1] nsiva5809@gmail.com, [2 *]siva.bheem@hotmail.com
[3] gopalamsaikumar910@gmail.com,[4] vaardhaanraj@gmail.com,
[5] sushvitha2002@gmail.com,[6] chandu8466033179@gmail.com

***Abstract.*** Understanding emotions is critical to many fields, including psychology, medicine, and human-computer interaction. The study uses datasets from RAVDESS, SAVEE, CREMA, and TESS which cover a wide spectrum of emotions, including neutral, surprise, happiness, sadness, disgust, anger, and fear to thoroughly investigate machine learning algorithms for emotion identification in audio data. Long short-term memory (LSTM) networks, decision trees, and convolutional neural networks (CNNs) are the three different models that are investigated. Decision trees provide simple classification, LSTMs extract temporal correlations from the data, and CNNs are excellent at extracting features from audio signals. Performance indicators like recall, F1, precision, and accuracy score are used in performance evaluation. Significantly, the CNN model outperforms Decision Trees and LSTM networks with 72% and 77%, respectively, in emotion categorization accuracy, reaching a remarkable 91%. This work offers insightful information about how well different machine learning models perform when it comes to audio-based emotion recognition. These realizations will have a big impact on developing trustworthy emotion detection systems for emotional computing, human-robot interaction, and mental health assessment. Future studies could investigate ensemble approaches or hybrid models to improve emotion detection capabilities and progress the creation of increasingly intricate and accurate emotion recognition systems.

**Keywords:** Convolutional Neural Networks (CNN), Decision Trees, Emotion Recognition, LSTM Networks, Machine Learning Algorithms.

## 1 Introduction

In Regarding the field of speech emotion recognition (SER), recent years have witnessed significant advancements driven by the convergence of machine learning

K. R. Madhavi et al. (eds.), *Proceedings of the International Conference on Computational Innovations and Emerging Trends (ICCIET 2024)*, Advances in Computer Science Research 112,
https://doi.org/10.2991/978-94-6463-471-6_45

techniques, signal processing methodologies, and the growing availability of large-scale emotional speech datasets. Researchers have explored various machine learning models, including deep learning architectures like LSTM networks and CNN, in order to extract distinctive characteristics from speech signals and classify them into different emotional categories. While these efforts have shown promising results, challenges such as the subjective nature of human emotions, cultural influences on emotion expression, and the need for robust feature extraction methods persist. Additionally, choosing suitable machine learning models and optimizing hyperparameters remain crucial factors affecting the performance of SER systems. This study aims to contribute to the ongoing research in SER by conducting a comparative analysis of three prominent machine learning models: CNNs, Decision Trees, and LSTM networks. By evaluating these models using a diverse dataset comprising speech samples annotated with various emotional labels, insights into their effectiveness in recognizing and classifying emotions conveyed through speech are sought. By systematically comparing the performance of CNNs, decision trees, and LSTM networks across different emotional categories and dataset characteristics, strengths and limitations of each model are identified to provide practical guidance for researchers as well as practitioners in the field of affective computing. The motivation behind this research stems from the growing importance of SER in applications such as human-computer interaction, virtual assistants, mental health monitoring, and entertainment systems. Overall, this research aims to enhance understanding of machine learning models for SER and facilitate the development of more effective and empathetic human-machine interaction systems.

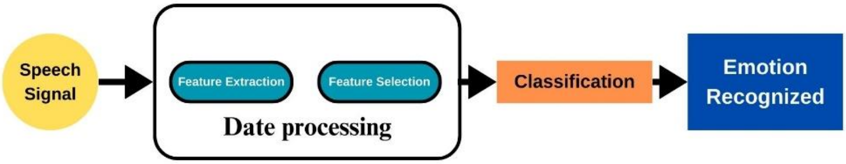


**Fig. 1.** Overview of the process of speech emotion recognition, starting with speech signals as input. Through feature extraction, classification, and interpretation of results, the system identifies emotions.

## 2 Literature Review

Recent Niharika S M and Soumya A., proposed a CNN-based voice Emotion Recognition model that can properly interpret emotions from voice data. The model captures important auditory properties but lacks other pre-processing techniques, such as data augmentation. Data augmentation approaches can improve accuracy by varied and robust training data, resulting in better emotion recognition performance. [1]

Majd Saloumi et al. developed a spoken Emotion Recognition (SER) model with a 1D-CNN architecture and MFCC designed exclusively for short spoken recordings. Their model, trained on the RAVDESS dataset, achieved an outstanding 83% accuracy by using data augmentation approaches. However, it is worth mentioning that expanding the dataset size has the potential to improve the model's accuracy even more, demonstrating the value of having a large and diverse training dataset when constructing more accurate SER models. [2]

K. Vamsi Krishna et al. investigated SER with Support Vector Machine (SVM), Multi-layer Perceptron (MLP), and a variety of audio characteristics such as MFCC, MEL, Chroma. This study implements a convolutional neural network (CNN) architecture for SER, emphasizing the integration of data augmentation techniques and extracting patterns and dependencies to improve performance. CNNs provide automated feature learning from raw audio data, allowing the model to recognize emotions in speech signals with greater robustness. [3]

Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi reflect on over a decade of research in automated emotion identification from speech, tracing the evolution of this field. Databases, modeling, annotation, analytic units, and prototypicality are all covered. They then discuss robustness, assessment, features, categorization, and system integration before moving on to automated processing. The thorough summary offers insights into this developing study area's past, present, and future. [4]

In a recent work, Shashidhar G. Koolagudi and K. Sreenivasa Rao examine the developing subject of emotion identification from speech. Important elements included in the review include different speech attributes, emotional speech corpora, and recognition models. In addition, it addresses the selection of classification models and lists important factors to take into account for next studies on emotion identification, especially with reference to India. [5]

# 3 Methodology

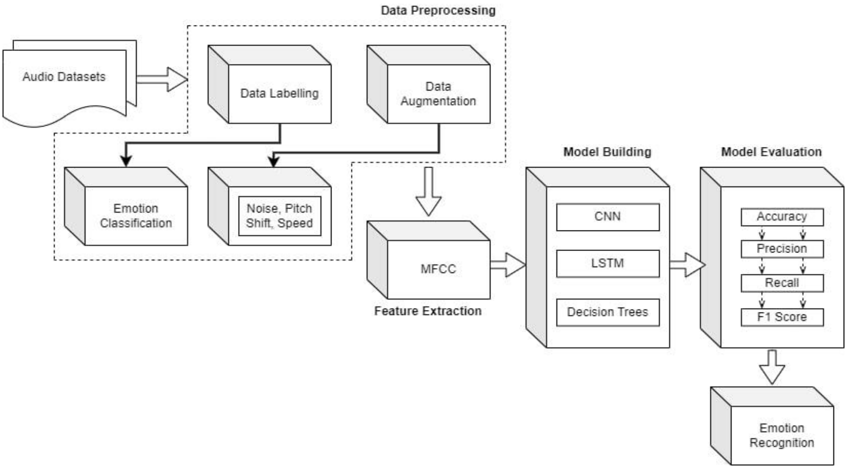


**Fig. 2.** Proposed Methodology

## 3.1 Datasets

The study utilized four widely used datasets: the TESS, RAVDESS, SAVEE, and CREMA-D. These datasets were chosen due to their diverse emotional content and the availability of labelled emotional categories. TESS comprises recordings of North American English speakers portraying seven emotional states, while RAVDESS contains speech samples from actors representing eight emotional categories. SAVEE consists of British English speakers expressing seven emotions, and CREMA-D features audio clips from actors portraying various emotions in a controlled environment.

## 3.2 Data Preprocessing

**Data Labelling.**
Data labelling for speech emotion recognition involves parsing the directory of audio files and extracting emotion labels from file names based on predefined conventions. Emotion indicators in file names are decoded and mapped to specific emotion categories. Each audio file is assigned the corresponding emotion label, ensuring a balanced distribution of emotions for effective model training. Quality assurance measures are implemented to verify labelling accuracy, including manual review and error rectification. The process is documented to maintain consistency across datasets.

**Data Augmentation.**
Data augmentation techniques are critical for improving the robustness and generalizability of machine learning models, especially in scenarios with limited training data. In the context of speech emotion recognition, data augmentation manipulates audio signals to generate a variety of training samples while retaining their underlying emotional content.

*Noise Injection.*
Random noise is introduced into the audio signal to simulate real-world environmental variations and improve model resilience to noise interference.

*Time Stretch.*
Alters the duration of an audio signal while maintaining its pitch and emotional characteristics. This technique uses temporal variations to mimic different speaking rates or durations of emotional expressions.

*Pitch Shifting.*
The pitch of the audio signal is changed, resulting in variations in vocal pitch without changing the emotional content. This technique simulates changes in vocal qualities or speaker identities.

*Time Shifting.*
The audio signal is shifted by a random time interval, resulting in phase shifts and time offsets. This technique represents temporal distortions and variations in the timing of emotional expressions.

*Speed Modification.*

The audio signal's speed is adjusted, either accelerated or decelerated, but its pitch remains unchanged. This technique simulates changes in speech rate or emotional intensity.

### 3.3 Feature Extraction.

In speech emotion detection, feature extraction includes converting raw audio data into a format that machine learning algorithms can analyze. Mel-frequency cepstral coefficients (MFCCs), a popular speech processing method, are commonly utilized for this purpose. MFCCs capture key aspects of the audio signal's spectral characteristics, such as pitch and timbre, making them ideal for representing speech. By combining MFCC feature extraction and data augmentation, the process generates a comprehensive feature set that captures the intricacies of speech signals, allowing for accurate emotion recognition.

## 4 Implementation

This paper delves into the domain of speech emotion recognition with a substantial dataset consisting of around 85,000 audio samples. Three distinct machine learning algorithms: CNN, LSTM and Decision Trees, each offering unique strengths for analyzing emotional content in speech signals. The implementation is carried out in Python, utilizing libraries such as TensorFlow and Keras within the Jupyter Notebook environment to facilitate efficient model development and evaluation.

**Convolutional Neural Network.**
CNN are particularly effective in extracting spatial and temporal patterns from sequential data, making them well-suited for analyzing audio signals. By leveraging convolutional layers, pooling operations, and nonlinear activation functions, CNNs can automatically learn hierarchical representations of the input data, enabling robust extraction of features.

**Long Short-Term Memory.**
LSTM networks, on the other hand, excel at capturing long-term dependencies and temporal dynamics within sequential data. These networks incorporate memory cells that can retain information over extended time periods, making them highly suitable for modelling time-series data like speech signals. By learning from the sequential nature of the input data, LSTM networks can effectively capture the temporal dynamics underlying emotional expressions in speech.

**Decision Trees.**
This provides a different approach to modelling by partitioning the input space into smaller regions based on feature values. Each partition represents a decision node, and the tree structure is built recursively by selecting the most informative features at each

node. Decision trees offer interpretability and simplicity, making them useful for understanding the relationships between input features and target labels.

## 4.1 Model Training

The dataset is separated into training and validation sets during the training phase of speech emotion recognition algorithms. The training technique consists of modifying the model parameters to minimize the loss function and improve performance on training data. This optimization is commonly carried out utilizing techniques such as stochastic gradient descent (SGD) or its derivatives. Throughout training, the model iteratively updates its parameters according to the gradients of the loss function with respect to these parameters. This modification process is repeated until the model converges to a point where additional training does not significantly improve performance on the validation set, or until a predetermined number of epochs are reached.

## 4.2 Model Evaluation

The evaluation of the developed models for speech emotion recognition involved the use of several key metrics, will later aid in recognizing speech samples.

**Accuracy.**

The capacity of a test to accurately identify weak and strong instances is known as accuracy. We should record the small percentage of true positive and true negative results in thoroughly reviewed instances in order to measure the exactness of a test. This might be expressed mathematically as:

$$Accuracy = \frac{(True\ positives + True\ Negatives)}{Total\ no\ of\ Test\ samples} (1)$$

**Precision.**

Precision quantifies the percentage of correctly classified samples or occurrences among the positives. Consequently, the accuracy may be determined by applying the subsequent formula:

$$Precision = \frac{(True\ positives)}{True\ Positivies + False\ positives} (2)$$

**Recall.**

A ML statistic called recall evaluates a model's ability to identify all relevant samples inside a given class. The ratio of correctly expected positive perceptions to actual positives provides information of a certain class.

$$Recall = \frac{(True\ positives)}{True\ Positivies + False\ Negatives} (3)$$

**F1 Score.**

It is a metric used in ML assessments to gauge a model's accuracy. It combines review scores and model precision. The accuracy measurement calculates the frequency with which a model predicts correctly throughout the whole dataset.

$$F1\,Score = 2X\frac{(Precision * Recall)}{(Precision + Recall)} \quad (4)$$

## 5 Results and Comparison

Upon conducting a comprehensive comparative analysis of CNN, LSTM networks, and decision trees for emotion prediction from speech data, several key insights and outcomes have emerged:

**Table 1.** Performance Evaluation Table

| Algorithm | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| CNN | 0.9181 | 0.9345 | 0.9024 | 0.9179 |
| LSTM | 0.7712 | 0.8759 | 0.7290 | 0.7644 |
| Decision Tree | 0.7290 | 0.7292 | 0.6810 | 0.7291 |

As shown in Table 1, The CNN model obtained the maximum accuracy of 91.81%, as well as good precision (93.45%) and recall (90.24%) rates, yielding a strong F1 score of 91.79%. This demonstrates CNN's extraordinary ability to reliably classify emotions from speech data. In contrast, the LSTM model achieved an accuracy of 77.12%, indicating somewhat worse performance than CNN. While LSTM achieved remarkable precision (87.59%), its recall (72.90%) and F1 score (76.44%) were significantly lower, indicating difficulties in understanding some emotional nuances in the dataset. Similarly, the Decision Tree model attained an accuracy of 72.90%, with recall, F1 Score and precision values ranging between 72 and 73%. Although decision trees provide interpretability and simplicity, their performance lags behind CNN and LSTM, demonstrating limits in dealing with the complexity of emotion.

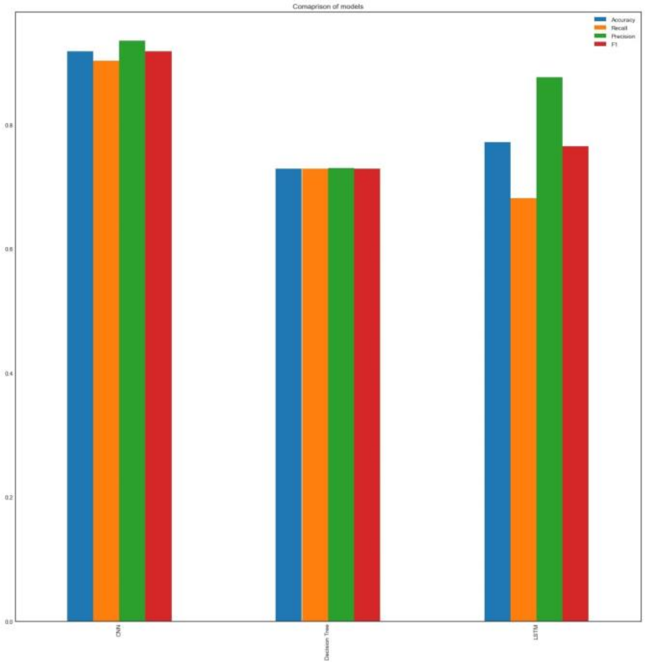


**Fig. 3.** Performance Comparison Graph

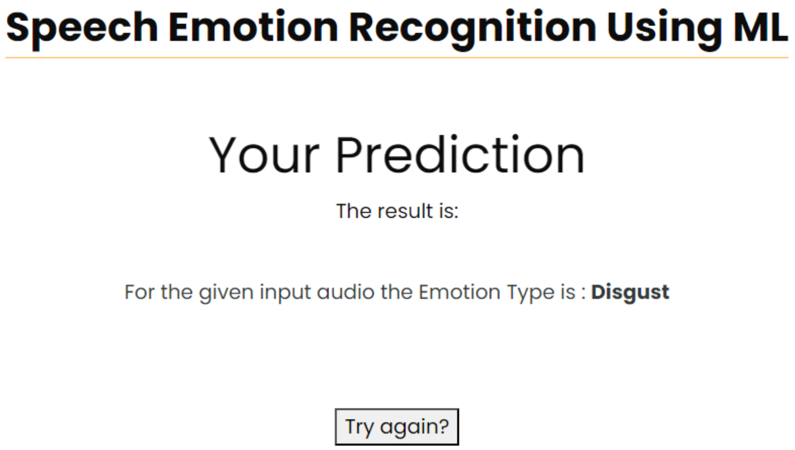
In terms of predicting emotions from speech data, the CNN model was used as the primary model. This selection was made based on its demonstrated superiority in accuracy and performance measures when compared to other models like LSTM and decision trees. The CNN model uses convolutional neural networks to analyze and extract information from audio inputs, allowing for accurate emotion classification.

**Comparison with existing methods.**

The approach presented in this study achieved an accuracy of 91% in emotion recognition and demonstrated the capability to accurately detect emotions in given speech samples. This performance represents a substantial advancement compared to the existing methods, which relied solely on a smaller number of datasets without utilizing data augmentation techniques and yielded lower accuracies.

The CNN model is chosen for its exceptional performance in accurately classifying emotions into the following categories:

Angry, Calm, Disgust, Fear, Happy, Neutral, Sad, Surprise



Speech Emotion Recognition Using ML

Your Prediction

The result is:

For the given input audio the Emotion Type is : **Disgust**

**Fig. 4.** Emotion prediction of a random audio

## 6 Conclusion

The study presented a comprehensive approach to SER using a combination of CNN, LSTM, and decision tree models. Through extensive experimentation and evaluation, several key findings emerged. Firstly, the CNN model outperformed the LSTM and Decision Tree models in emotion classification, with an accuracy of 91.81%. This highlights the effectiveness of CNN architectures in capturing relevant features from audio data for accurate emotion prediction. Secondly, while the LSTM and Decision Tree models showed respectable performance, their accuracies of 77.12% and 72.90%, respectively, were notably lower than those of the CNN model. This suggests that more complex temporal dependencies captured by LSTM may not always translate to better performance in speech emotion recognition tasks, and decision trees may struggle with capturing nuanced patterns in audio data. Furthermore, the study identified specific emotions where each model excelled. For instance, the CNN model demonstrated particularly high accuracy in classifying emotions such as anger, fear, and happiness, indicating its robustness in capturing the spectral features associated with these emotions.

## 7 Future Work

Future work could explore hybrid models combining acoustic, contextual, and linguistic features for more robust emotion classification. Additionally, investigating the generalizability of the models across different languages and cultures may offer useful insights on the universality of emotional expression in speech.

# 8 References

1. Niharika S M, Soumya A: Speech Emotion Recognition using Machine Learning. In: 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS). (2023)
2. Majd Saloumi et al: Speech Emotion Recognition using One-Dimensional Convolutional Neural Networks. In: 46th International Conference on Telecommunications and Signal Processing (TSP). (2023)
3. K. Vamsi Krishna et al: Speech Emotion Recognition using ML(SVM). In: 6th International Conference on Computing Methodologies and Communication (ICCMC). (2022)
4. Björn Schuller, Anton Batliner et al: T.: Emotion-Oriented Systems: The HUMAINE Handbook, Springer. (2011)
5. Koolagudi, S. G., & Rao, K. S. "Emotion Recognition from Speech: A Review." International Journal of Speech Technology, 15(2), 99-117 (2012)
6. Kasarapu Ramani, Lakshmi Haritha M. "Deep Learning and its Applications: A Real –World Perspective". In Deep Learning and Edge Computing Solutions for High Performance Computing, Springer. (2021)
7. B. Dinesh, P. Chilukuri, G. P. Sree, K. Venkatesh, M. Delli and K. R. Nandish, "Chat and Voice Bot Implementation for Cardio and ENT Queries Using NLP," 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA) (2023)
8. A V Sriharsha, Ms K Yochana, “Improving Efficiency of CNN using Octave Convolution”, Journal Article Open Access International Journal of Recent Technology and Engineering (IJRTE), Volume: 8, Issue: 6, Pages: 5412-5418. (2020)
9. Schuller, B., & Batliner, A. "META-INTERSPEECH: A Roadmap Towards the Audio-Visual Processing of Human Emotional Behavior." In Interspeech (2015)
10. Kim, S. H., & Kim, H. Y. "Emotion Recognition System Using Short-Term Monitoring of Physiological Signals." IEEE Transactions on Biomedical Engineering, 60(12), 3374-3383 (2013)
11. Schuller, B., et al. "Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge." Speech Communication, 53(9-10), 1062-1087. (2011)
12. Lotfian, R., & Cohn, J. F. "Analysis of Head-Pose-Induced Image Variations for Improved Emotion Recognition." In International Conference on Multimodal Interfaces. (2012)
13. Raju, S. Viswanadha, A. Vinaya Babu, G. V. S. Raju, and K. R. Madhavi. "W-Period Technique for Parallel String Matching." IJCSNS 7, no. 9 (2007): 162.
14. Jiang, W., Chen, T., & Sun, M. "Emotion Recognition from Speech: An Overview." Frontiers in Systems Neuroscience, 8, 106. (2014).
15. Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions." IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(1), 39-58. (2009)