IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, June 2025
Impact Factor: 7.67

Speech Emotion Recognition Using Feedforward Neural Network

Dr. Deepali Sale[1], Anand Bhagat[2], Pranit Bhalekar[3], Rohit Gorde[4], Mahendra Gayakwad[5]
Associate Professor, Department of Computer Engineering [1]
Students, Department of Computer Engineering [2,3,4,5]
Dr. D. Y. Patil College of Engineering and Innovation, Pune, India.

Abstract: *Speech Emotion Recognition (SER) has garnered significant attention in recent years due to its applications in human-computer interaction, healthcare, customer service automation, and affective computing. This paper presents the design and implementation of a real-time Speech Emotion Recognition (SER) system using Feedforward Neural Networks (FNNs) for detecting and classifying emotions from speech signals. By utilizing the RAVDESS dataset, effective pre-processing techniques, and a well-structured FNN architecture, the system is trained to accurately recognize eight distinct emotions. The study outlines the methodology, implementation details, and real-time deployment of the system within a web application, highlighting its practical feasibility and potential applications*

Keywords: Speech Emotion Recognition (SER), Feedforward Neural Networks (FNN), Deep Neural Networks (DNN), Affective Computing, Human–Computer Interaction (HCI), Real-time Systems

I. INTRODUCTION

Speech is a fundamental medium for human communication, conveying not only linguistic content but also emotional and intentional cues. With the growing prevalence of intelligent systems capable of processing speech, recognizing the emotional state of a speaker has become an important research focus. Speech Emotion Recognition (SER) systems aim to automatically identify emotions such as happiness, sadness, anger, and fear from speech signals. Accurate emotion recognition can significantly enhance human-computer interaction (HCI) by enabling systems to respond in a more natural, intuitive, and empathetic manner.

Traditional SER approaches largely relied on handcrafted feature extraction combined with classical machine learning algorithms, including Support Vector Machines (SVM) and Hidden Markov Models (HMM). While effective in some cases, these methods often struggled with the inherent variability and complexity of speech data, limiting their generalizability and accuracy across diverse datasets.

The advent of deep learning has transformed SER research, introducing models capable of automatically learning complex features from raw or minimally processed speech inputs. Among these models, Feedforward Neural Networks (FNNs) consisting of multiple fully connected layers have demonstrated potential in capturing intricate speech patterns relevant for emotion classification.

This paper focuses on investigating the application of FNNs in the domain of SER. We analyze their architectural design, performance strengths, and limitations relative to other deep learning techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Additionally, we discuss the challenges associated with deploying FNN-based SER systems in real-time scenarios, considering factors such as computational efficiency and latency.

II. LITERATURE SURVEY

In recent years, Speech Emotion Recognition (SER) has garnered significant attention owing to advances in deep learning and its diverse applications in human-computer interaction, healthcare, and education. Various neural network architectures have been explored to capture the emotional nuances in speech signals.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, June 2025
Impact Factor: 7.67

Zhang et al. [1] employed a Convolutional Neural Network (CNN) utilizing Mel-frequency cepstral coefficients (MFCCs) as input features, achieving substantial improvements over traditional machine learning methods. Despite these gains, the CNN's high computational demand posed challenges for real-time deployment. Our project addresses this limitation by designing a lightweight CNN architecture that balances computational efficiency with performance.

Chen and Wang [7] leveraged Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) to model temporal dependencies in emotion-rich speech using the EMO-DB dataset, resulting in notable accuracy improvements. However, RNN-based models are often computationally intensive, limiting their real-time applicability. To mitigate this, our approach integrates optimized architectures aimed at reducing processing time while retaining the temporal feature extraction benefits of RNNs.

Liu et al. [8] proposed a hybrid CNN-LSTM model that combined spatial and temporal feature extraction on the CREMA-D dataset. Although this hybrid model achieved promising results, its generalizability was constrained by the relatively small dataset size. Our work tackles this limitation through data augmentation and parameter optimization to improve both robustness and efficiency.

Khan et al. [9] introduced a lightweight CNN model tailored for real-time emotion detection with reduced computational complexity. However, this came at the cost of decreased accuracy compared to larger, more complex models. Our project seeks to enhance accuracy by refining feature selection and tuning model parameters without compromising computational efficiency.

Smith et al. [10] developed a noise-robust Deep Neural Network (DNN) capable of recognizing emotions in noisy conditions. While effective under certain noise scenarios, the model's generalization to diverse acoustic environments was limited. To improve robustness, our system incorporates advanced pre-processing techniques to handle a broader range of noise types.

These studies highlight the progress deep learning has brought to SER but also underscore persistent challenges related to real-time processing, noise robustness, and model generalization. Our research contributes to this field by proposing a lightweight, noise-resilient model optimized for efficient and accurate emotion recognition across varied acoustic settings.

III. MOTIVATION

Human emotions play a vital role in communication, significantly influencing the meaning and context of spoken language. As the demand for intelligent and responsive systems increases, incorporating emotional understanding into human-computer interaction has become essential. Traditional machine learning approaches for Speech Emotion Recognition (SER), including Support Vector Machines (SVM) and Hidden Markov Models (HMM), have shown limited effectiveness in capturing the complex, non-linear relationships inherent in speech data. These methods depend heavily on handcrafted features, which are often inconsistent due to variability among speakers, recording environments, and emotional expressions.

Deep Neural Networks (DNNs), particularly Feedforward Neural Networks (FNNs), provide a promising alternative by learning hierarchical features directly from raw data, thereby eliminating the need for manual feature engineering. The architecture of FNNs enables the detection of subtle and complex patterns within speech signals, resulting in improved emotion classification accuracy.

With the increasing importance of real-time systems and affective computing in applications such as virtual assistants, customer service, and healthcare, there is a pressing need for SER systems that are both accurate and computationally efficient. This research is motivated by the goal to address these challenges by leveraging DNN-based models to develop robust, scalable, and real-time capable Speech Emotion Recognition systems.

IV. OBJECTIVE

The primary objective of this project is to design and implement a Speech Emotion Recognition (SER) system utilizing Deep Neural Network (DNN) based Feedforward Neural Networks (FNNs). Specifically, this research aims to:

- Investigate the capability of FNNs to extract meaningful features from speech signals and accurately classify emotional states.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, June 2025
Impact Factor: 7.67

- Evaluate the performance and robustness of the proposed model across multiple datasets, including scenarios with real-world noise and variability, to ensure generalization.
- Optimize the model architecture and computational efficiency to enable real-time emotion recognition, facilitating its integration into interactive applications.
- Explore potential future enhancements, such as the incorporation of multimodal emotion recognition techniques (e.g., combining speech with facial expression analysis) to improve accuracy and broaden applicability across diverse domains.

V. METHODOLOGY AND ARCHITECTURE

1. Data Collection

We utilized the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset, which includes labeled audio samples representing multiple emotions like anger, happiness, sadness, fear, and calm. The dataset was selected for its balanced quality and standardized emotion representation.

2. Data Pre-processing

To prepare the data for model training, we performed the following pre-processing steps:

- Feature Extraction: Extracted Mel-Frequency Cepstral Coefficients (MFCCs) from each audio file, which effectively capture emotional cues in speech.
- Standardization: The features were scaled to a standard range to improve model convergence and performance.
- Label Encoding: Emotion labels were converted into one-hot encoded vectors for compatibility with the neural network.

3. System Architecture

The overall system, as illustrated in the flowchart figure 5.1: Flowchart Diagram Representing SER shows, how application is divided into the following components:

- Frontend Web Interface: Built using HTML/CSS and JavaScript, it allows users to either upload or record audio for emotion analysis.
- Backend Server (Flask): Handles request routing, invokes the prediction pipeline, and returns the result.
- Emotion Recognition Model: A trained Deep Neural Network (DNN) that accepts MFCC features and outputs predicted emotion.
- Visualization Module: Displays the predicted emotion output clearly on the results screen.

4. Model Architecture

We implemented a Sequential Deep Neural Network (DNN) with the following specifications refer to Figure 6.2: for Architecture of SER

- Input Layer: Accepts 40-dimensional MFCC vectors.
- Hidden Layers: Two fully connected dense layers with 128 and 64 neurons respectively, both using ReLU activation, followed by Dropout layers (rate = 0.5) to prevent overfitting.
- Output Layer: A dense layer with 8 neurons (one for each emotion) and a softmax activation function.

5. Model Compilation and Training

- Loss Function: Categorical Cross-Entropy
- Optimizer: Adam
- Evaluation Metric: Accuracy
- Training Configuration: The model was trained for 50 epochs with a batch size of 32, using 80/20 train-test split for validation.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, June 2025
Impact Factor: 7.67

6. Model Evaluation

The model was tested on unseen samples from the test set. Metrics such as accuracy and confusion matrix analysis were used to evaluate its classification performance across different emotions.

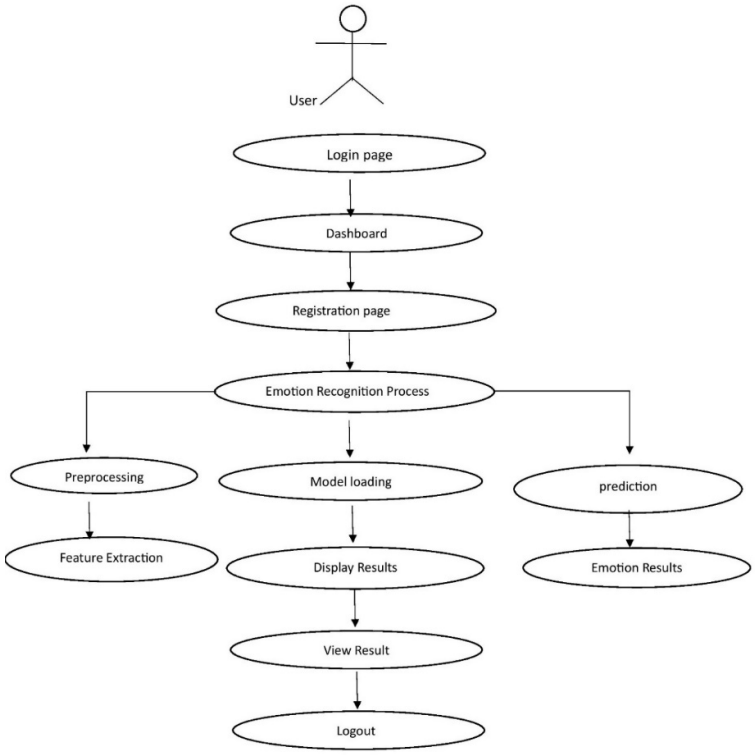

Fig. 5.1: Flowchart Diagram Representing SER

7. Deployment

The complete system was deployed as a web application using Flask as the backend framework. The flow includes:

- Uploading or recording an audio file
- Real-time feature extraction using librosa
- Feeding features to the model for prediction
- Returning the detected emotion to the frontend
- Displaying results to the user via a dynamic results page

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, June 2025
Impact Factor: 7.67

VI. ARCHITECTURE

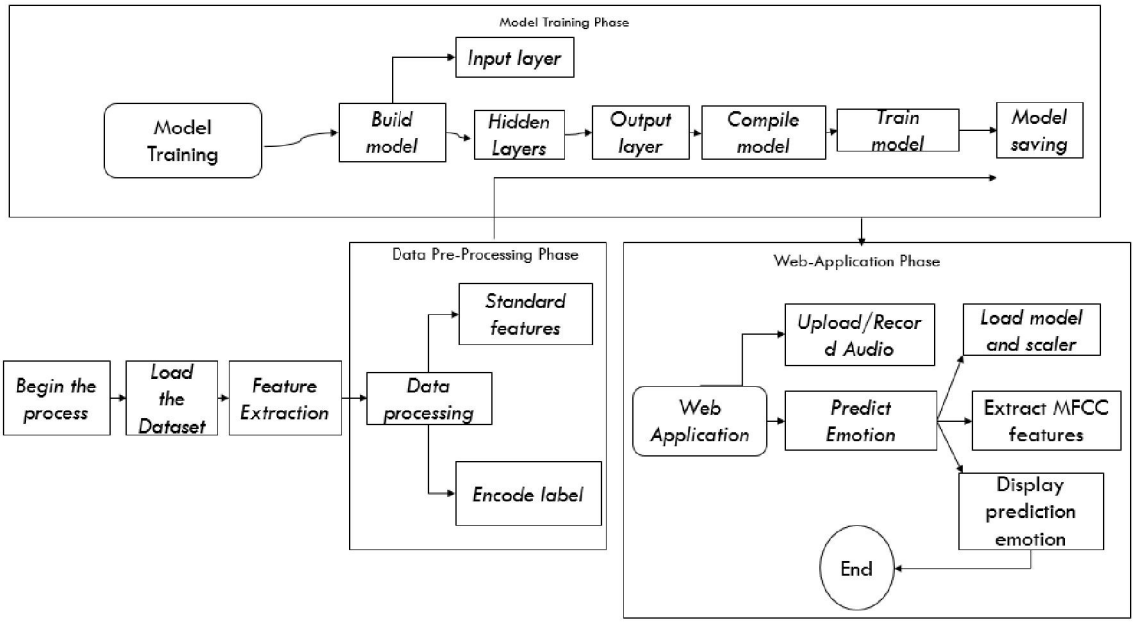

Fig. 6.1: Architecture of SER

VII. IMPLEMENTATION AND OUTPUT

The implementation of our Speech Emotion Recognition system, named EMOVIA, consists of both frontend and backend components integrated into a web-based interface. The application is developed using Python, Flask, Keras, HTML, and CSS, enabling users to interact with the model seamlessly. The system includes a registration and login feature for users, audio input functionality, and a backend model for real-time emotion prediction.

1. Registration and Login Interface

To ensure personalized usage and secure access, the system begins with user authentication:

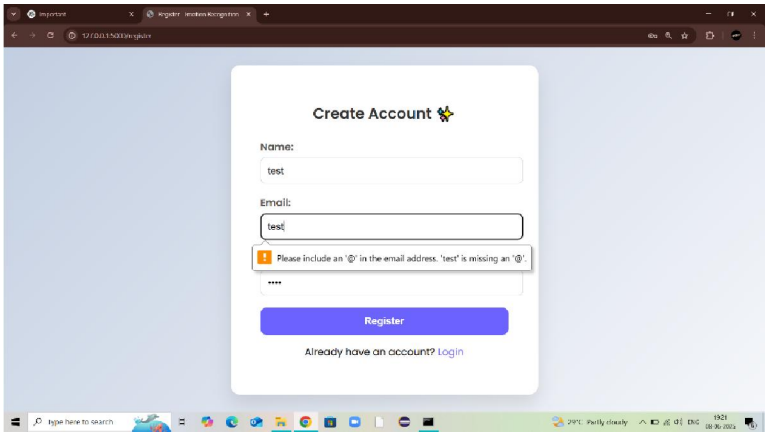

Fig. 7.1: Register Page

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, June 2025
Impact Factor: 7.67

a) Register Page

The Register Page provides a secure and intuitive interface where new users can create an account by entering their name, email address, and a password. This functionality helps ensure authenticated access and lays the groundwork for maintaining personalized emotion analysis records. Upon successful registration, users are redirected to the login page to access the system. You can see the layout and functionality of this interface in Figure 7.1

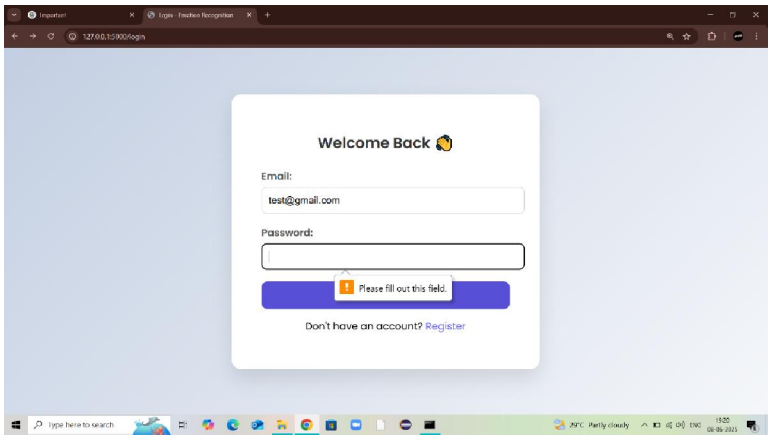

Fig. 7.2: Login Page

b) Login Page

The Login Page allows existing users to securely access the EMOVIA system by entering their registered email and password. The backend validates these credentials using basic authentication logic to ensure secure access. Upon successful login, users are redirected to the Home Page, where they can utilize the core functionalities of the platform such as uploading or recording audio for emotion prediction.You can see the interface of the login functionality in Figure 7.2

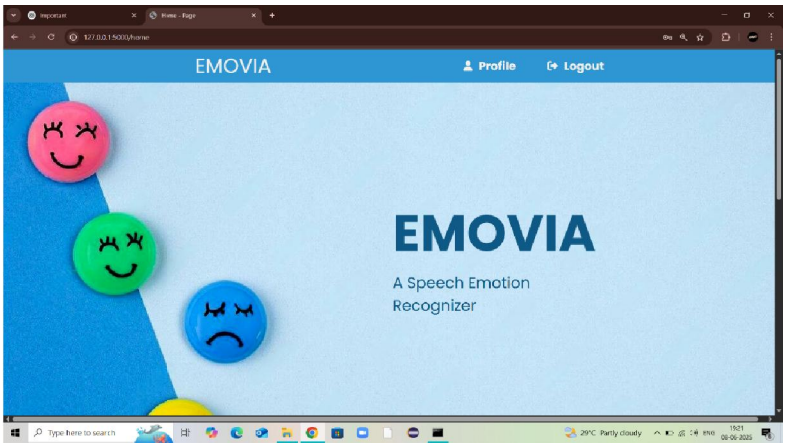

Fig. 7.3: Home Page

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, June 2025
Impact Factor: 7.67

2. Home Page

The Home Page acts as the central dashboard of the EMOVIA application. It greets users upon successful login and presents a clean, intuitive interface that allows navigation to the system's core services, including uploading an audio file or recording one in real-time. The user interface is deliberately designed to be minimalistic and accessible, ensuring ease of use even for individuals with no technical background. You can see the layout and functionality of the Home Page in Figure 7.3

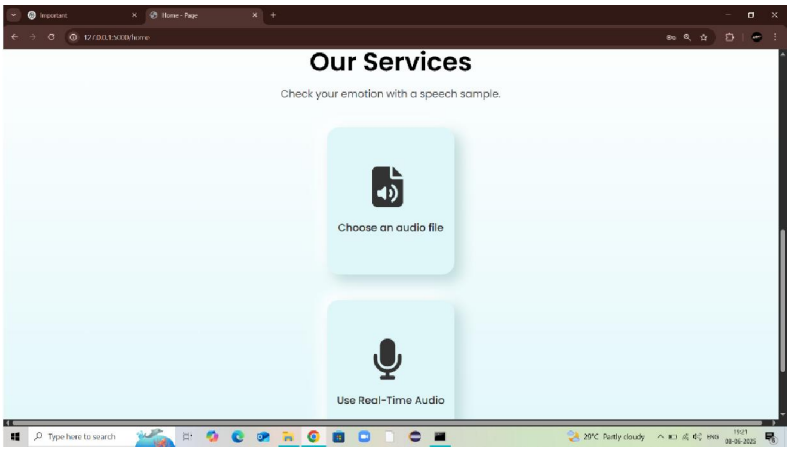

Fig. 7.4: Services Page

3. Services Page

On navigating to the Services Page, users are presented with two primary options for emotion recognition:

• Upload Audio File – This feature enables users to upload .wav files from their local device, which are then processed for emotional analysis.

• Record Audio – Users can utilize browser-based functionality to record their voice in real time. Once the recording is complete, the audio is automatically forwarded to the backend for processing.

The backend performs feature extraction using MFCCs, passes the extracted data to the trained Feedforward Neural Network (FNN) model, and returns the predicted emotion label to the user. The interactive layout of these service options is illustrated in Figure 7.4.

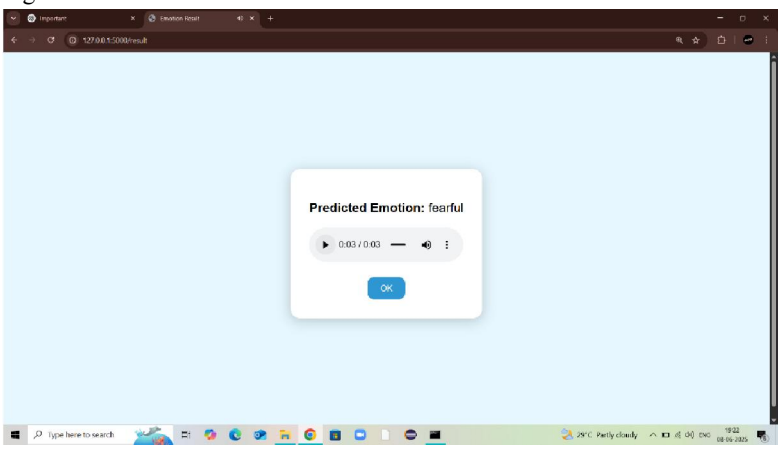

Fig. 7.5: Output Page

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, June 2025
Impact Factor: 7.67

4. Output Page

Once the model processes the input audio, the Output Page displays the predicted emotion—such as *"Predicted Emotion: Fearful"*—directly to the user. The result is shown in a clear and prominent manner, accompanied by a visual indicator (e.g., highlighted label or color-coded result) to enhance interpretability. This real-time feedback loop ensures users receive instant and actionable insight based on their speech input, completing the end-to-end interaction with the system. An example of the output interface is shown in Figure 7.5

VIII. PROJECT FEASIBILITY AND SCOPE

1. Feasibility Study

The feasibility of the Speech Emotion Recognition (SER) system is evaluated based on several aspects to ensure practical implementation, usability, and sustainability.

a) Technical Feasibility

The system leverages widely-used, open-source technologies such as Python, Tensor Flow/Keras, Flask, and Librosa. These tools are well-documented and community-supported, ensuring maintainability and scalability. The FNN (Feedforward Neural Network) model employed in this project is lightweight and performs efficiently even on moderate computational resources (e.g., laptops or cloud-based platforms). This makes the system technically feasible for academic, research, or small-scale production use.

b) Operational Feasibility

The user interface has been developed with simplicity and accessibility in mind. From registration to emotion detection output, the system is intuitive, requiring no prior technical expertise. Users can easily interact with the web interface to record or upload audio and receive real-time emotion predictions, ensuring that the application is operationally viable for educational or customer experience scenarios.

c) Economic Feasibility

The use of open-source libraries, freely available datasets (e.g., RAVDESS), and no requirement for proprietary APIs makes the solution highly cost-effective. The project can be deployed using free-tier cloud services or personal machines, ensuring minimal economic investment.

2. Scope of the Project

The current scope of the EMOVIA project focuses on identifying emotions from speech audio data using deep learning. It supports classification into eight basic emotional categories: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. The system is integrated into a full-stack web application, making it accessible and user-friendly.

In the broader sense, the project opens doors to several future enhancements, including:

- Integration with chat-bots or virtual assistants for sentiment-aware responses.
- Use in mental health monitoring tools to assess emotional well-being over time.
- Multilingual support for emotion detection in regional languages.
- Deployment as a mobile app for real-time emotional insight on the go.
- Expansion to audio-visual emotion recognition using video inputs.

IX. RESULT AND ANALYSIS

The performance of the proposed Speech Emotion Recognition (SER) system was evaluated using a test set comprising unseen audio samples from the RAVDESS dataset. The evaluation focused on four primary performance metrics: Accuracy, Precision, Recall, and F1-Score to comprehensively assess the model's classification ability.

The Feedforward Neural Network (FNN) model achieved a test accuracy of 81.90%, indicating that the system correctly identified emotions in approximately 82 out of every 100 test samples. The evaluation results demonstrated that the model effectively captured the emotional characteristics from speech signals using MFCC features.

The average performance metrics obtained during testing are summarized in Table 1.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, June 2025
Impact Factor: 7.67

Table 1: Result Metrics

Metric	Value(%)
Accuracy	81.90
Precision	82.50
Recall	81.20
F1-Score	81.80
Specificity	90.00

These results indicate that the model maintains a reliable balance between detecting the intended emotion classes (as reflected in precision and recall) and avoiding misclassifications (as reflected in specificity). The F1-score of 81.80% confirms the model's robustness in handling both false positives and false negatives. The system's ability to consistently achieve over 80% accuracy while maintaining high precision and recall values confirms its suitability for real-time deployment in interactive applications such as mental health monitoring tools, virtual assistants, and customer support systems.

X. CONCLUSION

In this research, we successfully designed, developed, and deployed a real-time Speech Emotion Recognition (SER) system using a Feedforward Neural Network (FNN) model, integrated into a web application named EMOVIA. Leveraging the publicly available RAVDESS dataset and effective pre-processing techniques such as Mel-frequency cepstral coefficients (MFCC) extraction, the system was trained to accurately classify eight distinct human emotions from speech signals.

The proposed model achieved a test accuracy of **81.90%**, demonstrating the capability of lightweight FNN architectures to perform efficient and reliable emotion recognition suitable for real-time applications. This performance validates the feasibility of deploying SER systems even in environments with limited computational resources.

The system was extended with a user-friendly web interface that allows users to register, log in, and interact via speech upload or live recording to receive immediate emotional analysis. This practical implementation showcases the system's potential for various real-world applications including mental health monitoring, human-computer interaction, virtual assistants, and enhanced customer experience.

Future work may focus on expanding the system's capabilities by incorporating multi-lingual support, mobile platform integration, and multimodal emotion recognition through the fusion of audio and facial expression analysis. Overall, this research establishes a solid foundation for advancing affective computing and developing intelligent, emotionally-aware human-machine interfaces.

REFERENCES

[1]. Zhang, X., et al., "Emotion Detection Using CNN," 2020.
[2]. "Feature Extraction and Classification of Emotion Recognition Using Deep Learning," *Journal of Computational Science*, vol. 40, pp. 101073, 2020.
[3]. "Emotion Recognition from Speech Signals Using DNN and SVM," *Applied Sciences*, vol. 10, no. 15, 2020.
[4]. "Speech Emotion Recognition Using Deep Learning Techniques: A Review," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 6, pp. 167-178, 2020. doi:10.14569/IJACSA.2020.0110621.
[5]. "Emotion Recognition from Speech Signals Using DNN and SVM," *Applied Sciences*, vol. 10, no. 15, 2020.
[6]. "A Novel Approach for Speech Emotion Recognition Based on DNN and GMM," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 512-525, 2021.
[7]. Chen, L., and Wang, H., "Temporal Feature Extraction with LSTM for SER," 2021.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, June 2025
Impact Factor: 7.67

[8]. Liu, Y., et al., "Hybrid CNN-LSTM Model for Speech Emotion Recognition," 2022.
[9]. Khan, R., et al., "Real-time Emotion Detection with Lightweight CNN," 2023.
[10]. Smith, J., et al., "Noise Robust SER Using Deep Neural Networks," 2023.
[11]. Kumar, A., & Singh, R. "A Comprehensive Review of Speech Emotion Recognition Systems." 2024.
[12]. Dr.Deepali Sale, Anand Bhagat, Pranit Bhalekar, Rohit Gorde, Mahendra Gayakwad "Speech Emotion Recognition Using Feedforward Neural Network" From IJARSCT 2024