Impact Factor (SJIF): 6.126 E-ISSN: 2583-1615

International Journal of Research in Academic World

Received: 05/February/2025 **IJRAW:** 2025; 4(SP3):37-43 **Accepted:** 19/March/2025

Speech Emotion Mapping Using Deep Learning

[*1]G Anitha, [2]Aiswarya Lakshmi MK and [3]Geethika P

[*1]Assistant Professor, Department of Data Analytics (PG), PSGR Krishnammal College for Women, Coimbatore, Tamil Nadu, India.

[2, 3]PG Student, Department of Data Analytics (PG), PSGR Krishnammal College for Women, Coimbatore, Tamil Nadu, India.

Abstract

Speech Emotion Recognition (SER) is crucial for improving human-machine interactions, enabling machines to understand and react to human emotions. This study delves into deep learning-driven SER approaches, employing sophisticated feature extraction methods such as Mel Frequency Cepstral Coefficients (MFCC), Mel Spectrogram, and Chroma Features. A comprehensive machine learning framework is established using classifiers like Support Vector Machines (SVM), Random Forest (RF), Multi-Layer Perceptron (MLP), k-Nearest Neighbors (KNN), and Naïve Bayes (NB). The Toronto Emotional Speech Set (TESS) dataset is utilized to train and validate these models, ensuring a broad spectrum of emotional variations. The research findings reveal that the proposed model effectively identifies emotions, including happiness, sadness, anger, neutrality, and fear, showcasing its potential in applications like AI-driven virtual assistants and mental health assessment tools. The core functionality involves real-time voice recording, where audio is captured and processed for feature extraction.

Comparative analysis highlights the advantages and limitations of each model, showcasing their performance in terms of accuracy. Furthermore, the system is deployed using a Flask-based web application for real-time emotion detection. The system is implemented in a Flask-based web environment, allowing real-time emotion prediction from voice inputs. The system is designed with a dark-themed user interface, featuring navigation options like Home, About, Analyze, and Realtime Prediction to enhance usability. Once the voice is recorded, the extracted features are passed through a pre-trained emotion classification model, deployed in the Flask environment. The backend processes the audio data, applies the trained model, and returns the detected emotion to the web interface in real time. The paper aims to provide a robust and scalable solution for emotion detection with applications in mental health monitoring, customer service, and AI-driven personal assistants. By leveraging real-time speech processing, efficient machine learning algorithms, and a user-friendly web interface, this system contributes to advancements in speech-based affective computing and human-computer interaction.

Keywords: Speech Emotion Recognition, Machine Learning, SVM, RF, MLP, KNN, NB, MFCC, Mel Spectrogram, Flask, Human-Computer Interaction.

1. Introduction

Emotions greatly impact human communication, influencing speech patterns, decision-making, and social interactions. SER is a growing domain within artificial intelligence that aims to extract and analyze emotional cues from speech signals, facilitating more intuitive and responsive machine interactions. Earlier methods focused on facial cues and biological indicators, whereas speech-based emotion detection offers a non-intrusive and effective alternative.

With advancements in machine learning, the accuracy and reliability of SER systems have greatly improved. This study employs feature extraction techniques such as MFCC, Mel Spectrogram, and Chroma Features, in conjunction with classifiers like SVM, RF, MLP, KNN, and NB, to develop an optimized SER model. The effectiveness of these models is evaluated using standard dataset, including TESS, which provide a diverse range of emotional speech samples.

SER has a wide range of applications, they are:

- **Virtual Assistants & Chatbots:** SER enables AI-powered virtual assistants (e.g., Alexa, Siri, Google Assistant) to detect user emotions and respond more naturally. By recognizing stress, frustration, or happiness, these systems can adapt their tone and responses to provide a more engaging and personalized interaction.
- **Healthcare & Mental Health Monitoring:** Emotion recognition plays a vital role in detecting early signs of stress, anxiety, and depression. SER can assist in mental health applications by monitoring patient emotions during telemedicine sessions or therapy sessions, providing clinicians with valuable insights into emotional well-being.
- **Call Centers & Customer Support**: Call centers utilize SER to analyze customer emotions in real-time. If a customer expresses dissatisfaction or anger, the system can redirect the call to a senior representative ensuring better customer service. This helps improve client satisfaction and optimizes customer-agent interactions.

***Corresponding Author:** G Anitha

- **Entertainment & Gaming:** SER enhances immersive experiences in gaming and entertainment by adjusting the storyline, background music, or difficulty level based on the player's emotional state. Emotion-aware AI in gaming adapts dynamically to player responses, making interactions more engaging.
- **Human-Robot Interaction:** Robots in various industries, including education, healthcare, and customer service, can benefit from SER to understand human emotions and respond appropriately. This is crucial for social robots interacting with humans in real-world settings, such as elderly care or assistive technology**.**

Integrating SER into real-world applications enhances user experiences by allowing machine learning-powered system to detect emotions and adapt responses accordingly. The primary goal of this research is to compare various machine learning methods for speech emotion recognition and deploy the most effective model using a Flask-based web interface for real-time emotion detection.

2. Literature Review

Zhang *et al*. (2021) [1] presented a multi-task deep learning model for speech emotion recognition, employing shared feature learning to enhance classification accuracy. The results demonstrated improved performance compared to single-task models. Wang *et al*. (2021) [2] introduced an adaptive feature fusion approach using deep neural networks for speech emotion recognition. The study highlighted the benefits of integrating multiple feature extraction techniques for enhanced classification accuracy. Jaiswal and Sahu (2021) [3] explored the use of transfer learning and convolutional neural networks (CNNs) for speech emotion recognition, evaluating pre-trained CNN architectures on emotional speech datasets and reporting high accuracy levels.

Li *et al*. (2021) [5] investigated a multi-task learning model with a self-attention mechanism for speech emotion recognition. The model outperformed conventional deep learning approaches by dynamically adjusting feature importance during training.

El-Hajj *et al*. (2021) [6] discussed the use of deep learning algorithms with an attention mechanism to enhance speech emotion recognition. The results indicated significant improvements in classification accuracy. Schuller *et al*. (2018) [7] presented the INTERSPEECH 2018 Computational Paralinguistics Challenge, focusing on atypical speech and self-assessed affect recognition. The study provided benchmark datasets and methodologies for emotion detection.

Fayek *et al*. (2017) [8] evaluated different deep learning architectures for speech emotion recognition, including fully connected, convolutional, and recurrent networks, comparing their performance across various datasets. Eyben *et al*. (2010) [9] introduced OpenSMILE, a widely used open-source feature extraction toolkit for speech emotion recognition, detailing its implementation and effectiveness in emotion classification tasks. Cowie *et al*. (2000) [10] described the development of FEELTRACE, an instrument designed for real-time recording of perceived emotions in speech, contributing to the field of affective computing by providing a reliable annotation tool.

Kim and Provost (2013) [11] proposed an emotion recognition framework based on hidden Markov models (HMMs). The study highlighted the effectiveness of sequence modeling techniques for speech-based emotion detection.

3. Related Work

Numerous studies have investigated Speech Emotion Recognition (SER) using diverse machine learning and deep learning methodologies. Traditional approaches primarily relied on handcrafted feature extraction techniques combined with classical classifiers such as Support Vector Machines (SVM) and Random Forest (RF). Researchers have utilized feature sets like Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), and Chroma features to train models for emotion classification. While these methods achieved moderate success, their performance was often constrained by the inability to capture intricate hierarchical structures in speech data.

With the rise of deep learning, architectures such as Multi-Layer Perceptron (MLP), Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks have significantly enhanced SER accuracy. CNNs are particularly effective in deriving spatial characteristics from spectrogram representations, while LSTMs excel in handling the temporal dependencies of speech signals. Hybrid models that integrate CNN and LSTM have demonstrated superior performance by leveraging both spatial and sequential information.

Several comparative analyses have also examined the effectiveness of k-Nearest Neighbors (KNN) and Naïve Bayes (NB) classifiers. While KNN performs well for emotion classification based on similarity metrics, its computational inefficiency in large datasets poses a limitation. Conversely, Naïve Bayes employs probabilistic techniques, offering efficient classification but struggling with the interdependence of speech features.

The role of training datasets is crucial in improving SER models. Publicly available datasets such as TESS, RAVDESS, Emo-DB, and CREMA-D provide extensive emotional speech samples that facilitate robust model evaluation. Researchers have also explored data augmentation techniques, including pitch shifting, noise addition, and time-stretching, to improve the generalizability of SER models across varied acoustic environments.

Recent advancements have focused on the deployment of SER models in real-world applications. Integration into web frameworks like Flask has enabled the creation of interactive SER systems capable of real-time emotion prediction. Such models are being incorporated into AI-driven virtual assistants, mental health monitoring tools, and customer service analytics, demonstrating the growing relevance of SER in enhancing human-computer interaction. This study builds upon these existing works by implementing a comparative analysis of multiple machine learning models and deploying the most effective model for real-time emotion recognition.

4. Methodology

This study's methodology consists of several essential steps, including data collection, feature extraction, model training, evaluation, and deployment. The workflow of the envisioned system is illustrated in the diagram below.

Data Collection: This research leverages the Toronto Emotional Speech Set (TESS), a dataset containing speech recordings annotated with various emotions, including happiness, sadness, anger, neutrality, and fear, ensuring diversity in emotional representation. This dataset provide a balanced representation of emotions, ensuring reliable model training and evaluation.

Feature Extraction: To convert raw audio signals into meaningful data, three primary feature extraction methods are used:

- **Mel Frequency Cepstral Coefficients (MFCC):** Captures essential spectral characteristics of speech.
- **Mel Spectrogram:** Provides a visual representation of frequency changes over time.
- **Chroma Features:** Extracts tonal and pitch-related information critical for emotion detection.

Model Training

Five classification models are trained to categorize emotions:

- **Multi-Layer Perceptron (MLP):** A deep learning model that captures complex speech variations.
- **Support Vector Machines (SVM):** A robust classifier that identifies optimal decision boundaries.
- **Random Forest (RF):** An aggregated learning approach that enhances classification precision by building multiple decision trees.
- **k-Nearest Neighbors (KNN):** A proximity-based classification technique that labels data based on nearby points.
- **Naïve Bayes (NB):** A probability-based classifier that presumes feature independence for fast categorization.

Model Evaluation

The trained models are evaluated using accuracy, precision, recall, and F1-score to determine their effectiveness in classifying speech emotions. A comparative analysis is performed to identify the top-performing model.

Deployment

The most effective model is integrated into a Flask-based web application, allowing real-time emotion recognition from speech inputs. Users can upload an audio file or provide live speech input for analysis.

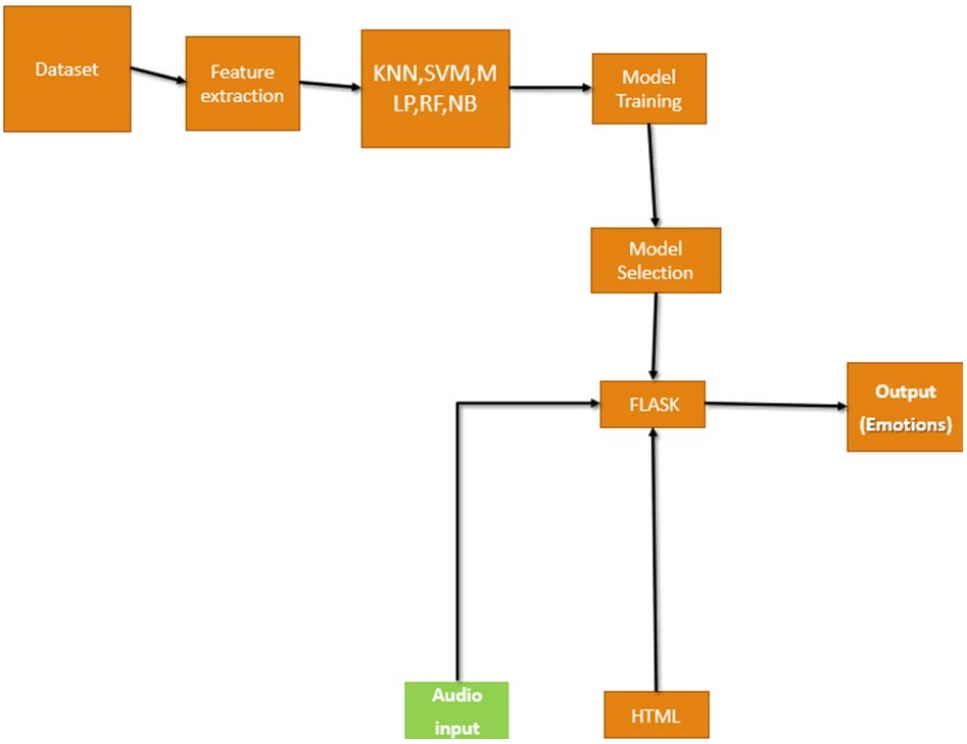

Fig 1: Process flow diagram

Fig 1 illustrates the process flow of a Speech Emotion Recognition system using different models. The process begins with the TESS dataset, from which relevant features are extracted. These features are then used to train multiple classifiers, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Multilayer Perceptron (MLP), Random Forest (RF), and Naïve Bayes (NB). The trained models undergo model selection to determine the best-performing classifier. A Flask-based web application is used for real-time emotion prediction. The system accepts audio input from users and processes it through the selected model via the Flask framework. The output, representing the detected emotion, is displayed through an HTML interface. This architecture enables efficient speech emotion recognition and real-time interaction for various applications.

4.1 Feature Extraction

Preparing audio data for Speech Emotion Recognition (SER) begins by loading audio files and extracting key features such as Chroma, Mel-Frequency Cepstral Coefficients (MFCCs), and Mel Spectrograms using the Librosa library. These features capture crucial aspects of speech, including tonal characteristics, energy distribution, and frequency variations, which help differentiate emotional expressions. Once extracted, the features are linked to their corresponding emotion labels, forming a structured dataframe that ensures a clear association between data and emotion categories. To minimize any biases caused by ordering, the dataset is randomized to enhance model generalization. Additionally, unnecessary indexing columns are removed to avoid influencing the training process. After preprocessing, the dataset is divided into input features and corresponding labels, preparing it for model training.

Mel Frequency Cepstral Coefficients (MFCC): MFCCs are a collection of attributes that offer a concise representation of an audio signal's short-term power spectrum. They are designed to mimic human auditory perception by

incorporating the Mel scale, which prioritizes lower frequencies, as the human ear is more sensitive to them.
The process of computing MFCCs consists of the following steps:

- The speech signal is divided into small overlapping segments.
- A Fourier Transform is used to convert the signal from the time domain to the frequency domain.
- The obtained power spectrum is projected onto the Mel scale using a set of triangular filters.
- A Discrete Cosine Transform (DCT) is then applied to minimize correlation among the Mel spectrum values.
- Generally, the first 13 to 20 MFCC coefficients are chosen as essential attributes for speech-related applications.

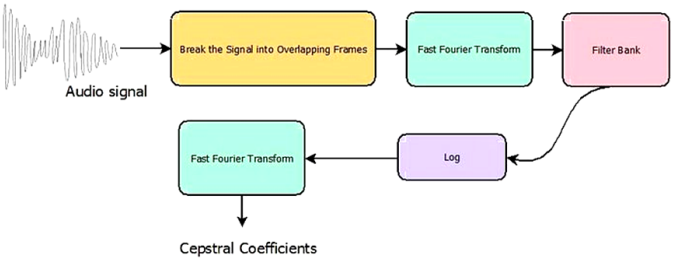

Fig 2: MFCC Process

Mel Scale
The Mel scale is a sensory-based scale of tones that closely follows how humans perceive sound frequencies. It maps actual frequency values (in Hertz) to a scale where equal distances correspond to perceived equal pitch differences.

Use of Mel Scale

- The human ear is more sensitive to lower frequencies (e.g., 100 Hz to 1000 Hz) than to higher frequencies (e.g., 8000 Hz to 9000 Hz).
- The Mel scale compresses higher frequencies while keeping lower frequencies more detailed, making it more aligned with human perception.
- Used in speech and audio processing to improve accuracy in tasks like speech recognition and emotion detection.

Calculation of Mel Scale
The approximate conversion from Hertz (Hz) to Mel (m) is given by:

$$m = 2595 \times \log_{10}\left(1 + \frac{f}{700}\right)$$

Mel Spectrogram
A Mel spectrogram provides a time-frequency analysis of an audio stream, with the frequency axis adjusted to the Mel scale to match human hearing perception. The process involves:

- Splitting the audio signal into small overlapping frames.
- Performing a Short-Time Fourier Transform (STFT) to capture time-dependent frequency characteristics.
- Transforming the frequency axis to the Mel scale using a series of triangular filter banks, highlighting frequencies that are more perceptible to the human ear.

Chroma Short-Time Fourier Transform (STFT)
STFT is a method used to examine the spectral makeup of a signal that changes over time. It divides the signal into brief sections and applies the Fourier Transform to each portion, generating a time-frequency depiction. This technique enables the analysis of how various frequency elements shift over time, making it useful for speech and sound processing.

$$STFT\{x(t)\}(m,k) = \sum_{n=-\infty}^{\infty} x(n)w(n-m)e^{-j2\pi kn/N}$$

- x(n) is the sound wave.
- w(n−m) is a weighting function (such as Hamming or Hann function).
- m is the time segment index.
- k is the spectral bin index.
- N is the frame length.
- $e^{-j2\pi kn/N}$ represents the Fourier Transform.

Chroma Features
Chroma features (or chroma vectors) are illustrations of the tonal content of a sound wave. They capture the intensity of different pitch classes (C, C#, D, D#,..., B) within a given frame, making them useful in music and speech analysis.
In speech and emotion recognition, chroma features help detect intonation patterns and vocal expressions. They are particularly useful for recognizing speaker emotions based on intonation and harmonic structure.

- The sound wave is divided into frames.
- A harmonic Transform is implemented to obtain frequency components.
- The frequency content is mapped into 12 chroma bins, each representing a pitch class.

5. Models Training

The model training process begins with preprocessed audio features extracted from the TESS dataset, including MFCCs, Mel Spectrograms, and Chroma features. These features serve as the input for training five machine learning models: Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), K-Nearest Neighbors (KNN), Random Forest (RF), and Naïve Bayes (NB). Each model undergoes hyperparameter tuning to optimize classification accuracy, where techniques like grid search or random search are applied to determine the best parameters for each algorithm. Feature scaling (such as Min-Max normalization or Standardization) is applied to models that require it, particularly SVM, MLP, and KNN, to ensure consistent feature ranges for better performance.
Each algorithm learns patterns from the training data using its unique approach. SVM finds an optimal hyperplane that separates different emotions, making it effective for high-dimensional data. MLP, a deep learning-based neural network, learns complex emotional representations using multiple layers of neurons. KNN classifies emotions by comparing new instances with the closest training examples, while RF, an aggregated learning approach, builds multiple decision trees to improve accuracy and reduce overfitting. Finally, NB, based on probability distributions, assumes feature independence and performs well with smaller datasets.
K-Nearest Neighbors (KNN): K-Nearest Neighbors (KNN) is a straightforward yet powerful categorization method that operates on the concept of closeness. It assigns a label to a new data instance based on the predominant category of its closest data points. The gap between points is measured using techniques like Euclidean metric, and a preset number of nearest points (K) is selected. KNN is model-free and performs effectively with limited datasets but can be computationally demanding for large-scale datasets.

Multi-Layer Perceptron (MLP): A Multi-Layer Perceptron (MLP) is a form of artificial neural network that comprises multiple tiers, including an input tier, hidden tiers, and an output tier. Each node in a tier is linked to nodes in the following tier through weighted links. MLP employs error correction (backpropagation) and an activation mechanism to identify intricate patterns in information. It is particularly advantageous for deep learning applications such as speech emotion analysis due to its capability to represent non-linear dependencies.

Support Vector Machine (SVM): A Support Vector Machine (SVM) is a robust supervised learning technique that determines an optimal decision boundary to distinguish various categories in a high-dimensional domain. It utilizes kernel methods (such as linear, polynomial, and radial basis function (RBF)) to address complex, non-linear classification challenges. SVM is effective for high-dimensional datasets and is resistant to overfitting, making it a dependable option for emotion classification.

Random Forest (RF): A Random Forest (RF) is a collective learning strategy that constructs multiple decision trees and aggregates their outcomes to boost precision and mitigate overfitting. It randomly picks attributes and data subsets for each tree, enhancing its resilience against data noise. RF is highly interpretable and excels in structured datasets, making it ideal for attribute-rich classification problems.

Naïve Bayes (NB): A Naïve Bayes (NB) classifier is a probabilistic model founded on Bayes' Rule, assuming independence among attributes. It computes the likelihood of an instance belonging to a category and assigns it to the category with the highest probability score. Despite its simplicity, NB is efficient for text analysis, voice recognition, and other domains where the independence hypothesis is reasonably valid.

6. Results

Model assessment is an imperative phase in artificial intelligence that guarantees the dependability and efficiency of a trained algorithm. It comprises assessing the model's performance using various metrics that measure its ability to generalize to unseen data. The trained models are assessed using accuracy, precision, recall, and F1-score to determine their performance.

Table 1: Accuracy for each model

Algorithms	Accuracy
SVM (Support Vector Machine)	98.75
MLP (Multi-Layer Precepton)	99.82
KNN (K-Nearest Neighbors)	99.64
Random Forest(RF)	99.82
Naïve Bayes(NB)	80.89

The table presents the accuracy performance employing five diverse machine learning techniques for speech emotion recognition. Among them, Multi-Layer Perceptron (MLP) and Random Forest (RF) achieved the highest accuracy of 99.82%, making them the most effective models for this task. K-Nearest Neighbors (KNN) followed closely with 99.64% accuracy, demonstrating strong classification performance. Support Vector Machine (SVM) also performed well, achieving 98.75% accuracy, proving to be a reliable model for speech-based emotion detection. However, Naïve Bayes (NB) had the lowest accuracy at 80.89%, indicating that its assumption of feature independence negatively impacted its performance in this context. Since MLP was chosen for deployment, it was chosen for its capability to identify intricate trends in speech attributes, guaranteeing precise and resilient emotion detection.

The decision to select MLP over other models is based on its ability to generalize well, even when working with intricate features such as MFCCs, mel spectrograms, and chroma features extracted from audio data. Unlike SVM and KNN, which may struggle with high-dimensional feature spaces, MLP can capture hierarchical representations of data, improving classification accuracy. Additionally, while Random Forest is robust and interpretable, it may not perform as well on sequential data like speech features. Naïve Bayes, although computationally efficient, assumes feature independence, which is often unrealistic for correlated audio features. Given these considerations, MLP proves to be the best choice due to its ability to learn deep representations, adaptability to complex feature relationships, and superior classification performance in speech emotion recognition tasks.

6.1. Output

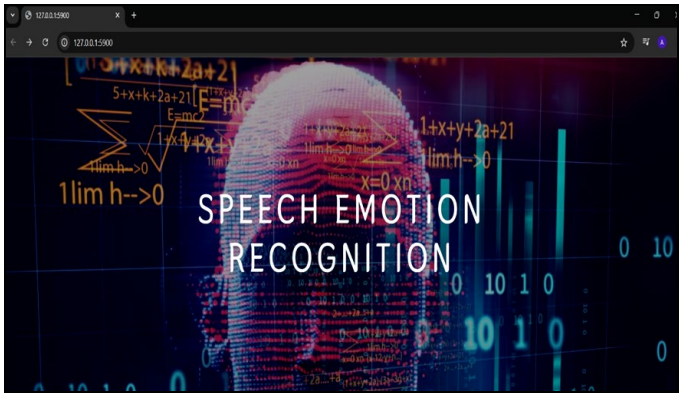

Fig 2: Flask Web Framework

Fig 2, displays the homepage with a visually appealing digital background, featuring mathematical notations and data visualizations, emphasizing the system's AI-driven analytical capabilities. The title "Speech Emotion Recognition" is prominently displayed, giving users an intuitive understanding of the application's purpose.

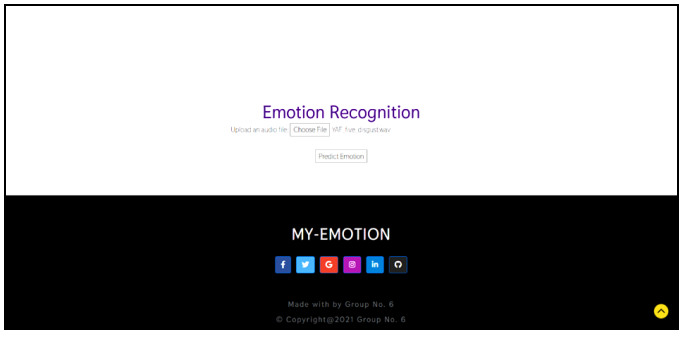

Fig 3: The Emotion Recognition Interface

In fig 3, users can upload an audio file and click the "Predict Emotion" button. This demonstrates the model's integration with Flask, allowing users to interact with the trained emotion recognition system in real-time. The clean and minimalistic design ensures ease of use.

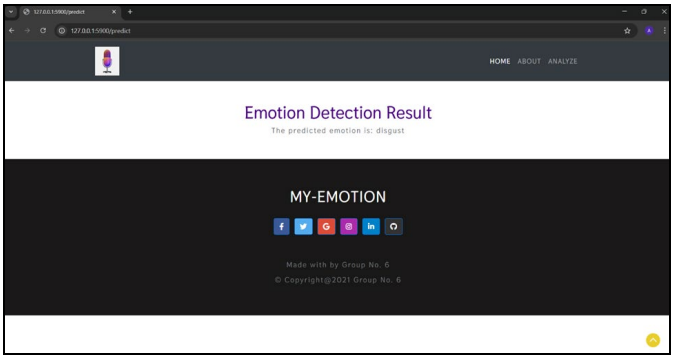

Fig 4: The Predicted Result Page

Fig 4, displays "The predicted emotion is: disgust." This confirms that the uploaded audio file was successfully processed by the model, which identified the emotional tone.

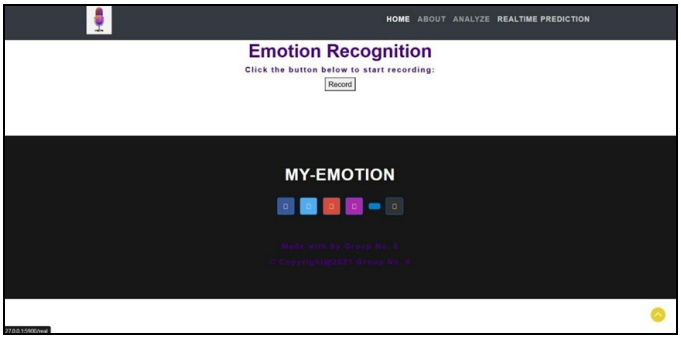

Fig 5: Voice Recording option

Figure 5 presents a web interface for an Emotion Recognition system, designed for voice-based emotion analysis. The interface features a dark-themed header with navigation options such as Home, About, Analyze, and Realtime Prediction, indicating multiple functionalities related to emotion detection. The main title, "Emotion Recognition," is highlighted in purple for visual emphasis. Below the title, an instruction prompts the initiation of recording by clicking the Record button, signaling that the system processes audio input to analyze emotions.

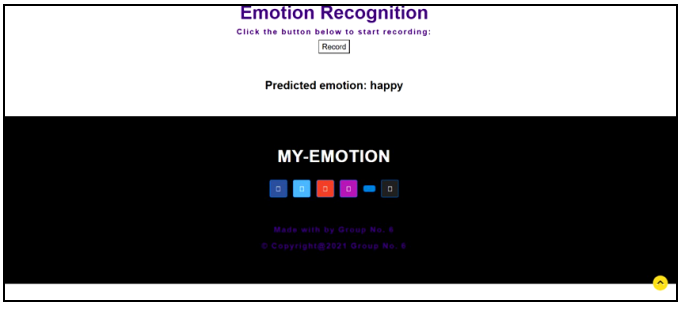

Fig 6: Prediction of the user voice

Fig 6, shows prediction "Predicted emotion: happy" text, which appears below the recording section. This indicates that the system has successfully processed the recorded audio and classified the emotion as "happy." This suggests the system utilizes machine learning or deep learning techniques to analyze vocal features and determine emotional states.

7. Conclusion

This study designs a sophisticated SER framework leveraging deep learning techniques. The Toronto Emotional Speech Set (TESS) dataset was utilized for training, and diverse speech attributes, including Mel Frequency Cepstral Coefficients (MFCCs), Chroma, and Mel Spectrograms, were derived using the Librosa toolkit. By implementing cutting-edge attribute extraction methods and robust machine learning classifiers, the suggested model attains high precision in emotion categorization. The dataset underwent preprocessing, and the extracted characteristics were employed to train multiple machine learning algorithms, ensuring a well-structured and effective strategy for emotion detection.

Five machine learning algorithms—Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), K-Nearest Neighbors (KNN), Random Forest (RF), and Naïve Bayes (NB)—were evaluated based on their accuracy in recognizing emotions. Among these, MLP and Random Forest achieved the highest accuracy of 99.82%, making them the most effective classifiers for this task. KNN and SVM also performed well, achieving 99.64% and 98.75% accuracy, respectively, whereas Naïve Bayes had the lowest accuracy (80.89%), indicating its limitations in handling complex speech features. The comparison of these models highlights the advantages of neural networks and ensemble learning methods in emotion recognition.

To ensure real-world usability, the Flask framework was utilized for deploying the trained model, creating a user-friendly interface that allows real-time speech emotion recognition. Users can input an audio file, which is then processed using the trained MLP model to classify the emotion. This web-based deployment makes the system accessible and interactive, demonstrating the feasibility of implementing Speech emotion detection in real-world applications like human-machine communication, digital assistants, and psychological health tracking.

Overall, this project demonstrates the potential of machine learning in speech emotion recognition, showcasing how different algorithms perform in classifying emotions from speech. The results indicate that MLP is the most effective model for this task, given its ability to learn complex patterns in speech data. Future improvements may include integrating convolutional and recurrent neural networks, increasing dataset diversity, and refining real-time processing capabilities for practical applications.

References

1. Zhang X, Li B, Wang J & Li L. Speech emotion recognition using a multi-task deep learning model. IEEE Access. 2021; 9:57521-57531.
2. Wang J, Hu X, Chen Y & Lai J. Speech emotion recognition using deep neural networks with adaptive feature fusion. IEEE Transactions on Affective Computing. 2021; 12(3):537-550.
3. Jaiswal S & Sahu K. Speech emotion recognition using transfer learning and convolutional neural network. *Journal of Ambient Intelligence and Humanized Computing*. 2021; 12(4):4167-4178.
4. Gandomi M, Ramezani R & Shojafar M. A novel ensemble model for speech emotion recognition using convolutional neural network and long short-term memory. Soft Computing. 2021; 25(8):5425-5439.
5. Li Y, Zhang X & Guo Q. Speech emotion recognition based on multi-task learning with self-attention mechanism. Multimedia Tools and Applications. 2021; 80(3):3353-3373.
6. El-Hajj M, El-Said M & Mokhtar H. Speech emotion recognition using deep learning algorithms with attention mechanism. Soft Computing. 2021; 25(7):4987-5001.

7. Schuller B, Steidl S & Batliner A. The INTERSPEECH 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats. Proceedings of Interspeech, 2018, 122-126.
8. Fayek HM, Lech M & Cavedon L. Evaluating deep learning architectures for Speech Emotion Recognition. Neural Networks. 2017; 92:60-68.
9. Eyben F, Wöllmer M & Schuller B. OpenSMILE – The Munich versatile and fast open-source audio feature extractor. Proceedings of ACM Multimedia, 2010, 1459-1462.
10. Cowie R, Douglas-Cowie E & Savvidou S. 'FEELTRACE': An instrument for recording perceived emotion in real time. Proceedings of Speech Emotion Recognition Workshop, 2000, 19-24.
11. Kim J & Provost EM. Emotion recognition from speech using hidden Markov models. IEEE Transactions on Affective Computing. 2013; 4(4):435-445.
12. Busso C, Bulut M, Narayanan SS. Toward effective automatic recognition systems of emotion in speech. Speech Communication. 2008; 50(5):455-470.
13. Tao J & Tan T. Affective computing: A review. Affective Computing and Intelligent Interaction, 2005, 981-995.
14. Picard RW. Affective Computing. MIT Press, 1997.
15. Poria S, Cambria E, Howard N, Huang G & Hussain A. Fusing audio, visual, and textual clues for sentiment analysis. IEEE Transactions on Affective Computing. 2016; 7(3):345-358.
16. Lee CM & Narayanan S. Toward detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing. 2005; 13(2):293-303.
17. Hinton GE, Osindero S & Teh YW. A fast learning algorithm for deep belief nets. Neural Computation. 2006; 18(7):1527-1554.
18. Krizhevsky A, Sutskever I & Hinton G. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems (NeurIPS):1097-1105.
19. Goodfellow I, Bengio Y & Courville A. Deep Learning. *MIT Press,* 2016.
20. Pan SJ & Yang Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering.* 2010; 22(10):1345-1359.