AN IDP PROJECT REPORT

on

**“****Speech Emotion Recognition Using**

**Machine Learning Techniques”**

**Submitted**

**By**

221FA04266

K. Bhavyaratna

221FA04622

T. Soma

221FA04301

P. Sairam

221FA04631

P.B.V.S.Navya

***Under the guidance of***

***Mr. Sourav Mondal***

*Associate Professor*



**SCHOOL OF COMPUTING & INFORMATICS**

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**VIGNAN'S FOUNDATION FOR SCIENCE, TECHNOLOGY AND RESEARCH Deemed to be UNIVERSITY**

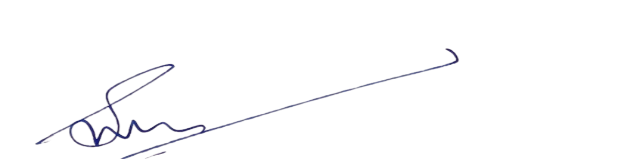
**Vadlamudi, Guntur.**

**ANDHRA PRADESH, INDIA, PIN-522213.**



**CERTIFICATE**

This is to certify that the Field Project entitled **“Speech Emotion Recognition Using Machine Learning Techniques”** that is being submitted by 221FA04266(K.Bhavyaratna), 221FA04301(P.Sairam), 221FA04622(T.Soma) and 221FA04631(P.B.V.S.Navya) for partial fulfilment of an IDP Project is a bonafide work carried out under the supervision of Mr.Sourav Mondal., Assistant Professor, Department of CSE.

Guide name& Signature

Designation HOD,CSE



**DECLARATION**

We hereby declare that the Field Project entitled “**Speech Emotion Recognition Using Machine Learning Techniques”** that is being submitted by 221FA04266(K.Bhavyaratna), 221FA04301(P.Sairam), 221FA04622(T.Soma) and 221FA04631(P.B.V.S.Navya) in partial fulfilment of an IDP Project course work. This is our original work, and this project has not formed the basis for the award of any degree. We have worked under the supervision of Mr. Sourav Mondal., Assistant Professor, Department of CSE.

**By**

**221FA04266(K.Bhavyaratna),**

**221FA04301(P.Sairam), 221FA04622(T.Soma),**

**221FA04631(P.B.V.S.Navya)**

## ABSTRACT

Speech Emotion Recognition (SER) is a rapidly advancing area in artificial intelligence and human-computer interaction, aimed at enabling machines to detect and interpret human emotions through vocal expressions. Emotions significantly influence communication, behavior, and decision-making, making their recognition crucial for developing empathetic AI systems. This project presents the development of a robust SER system utilizing the SAVEE dataset, which contains audio recordings labeled with seven emotional categories: anger, disgust, fear, happiness, sadness, surprise, and neutral. Audio preprocessing involves extracting Mel-Frequency Cepstral Coefficients (MFCCs), Spectral Contrast, Chroma to capture the acoustic features of speech, followed by systematic labeling and normalization to ensure consistent input for the models.

To improve classification accuracy, a stacked ensemble machine learning architecture is implemented, combining the strengths of Random Forest, Logistic Regression, and Support Vector Machine as base learners. The outputs of these models are further refined using meta-level classifiers—Support Vector Machine (SVM) and Random Forest (RF)—to make the final emotion predictions. This ensemble approach enhances the model's ability to generalize across speakers and emotion classes. The implementation is performed in Google Colab using Python, with evaluation based on epoch-wise accuracy tracking, loss visualization, confusion matrices, and classification reports. The proposed SER system demonstrates strong potential in creating emotionally responsive AI, with promising applications in virtual assistants, education technology, customer service, and mental health monitoring.

**TABLE OF CONTENTS**

1. [Introduction 1](#_bookmark0)
   1. [Background and Significance of Speech Emotion Recognition 2](#_bookmark1)
   2. Overview of Machine Learning in Speech Emotion Recognition 3
   3. [Research Objectives and Scope 4](#_bookmark3)
   4. Current Challenges in Speech Emotion Recognition 7
   5. Applications of ML to Speech Emotion Recognition 8
2. Literature Survey 11
   1. Literature review 12
   2. Motivation 14
3. Proposed System 16
   1. Input dataset 17

3.1.1 Detailed features of dataset 18

* 1. [Data Pre-processing 21](#_bookmark10)
  2. [Model Building 22](#_bookmark14)
  3. [Methodology of the system 24](#_bookmark16)
  4. Model Evaluation 26
  5. Constraints 32
  6. Cost and Sustainability Impact 48

1. [Implementation 51](#_bookmark19)
   1. Environment Setup 52
   2. Sample code for preprocessing and MLP operations 52
2. Experimentation and Result Analysis 56
3. Conclusion 60
4. References 62

**LIST OF FIGURES**

|  |  |
| --- | --- |
| Figure 1. Architecture of the proposed system | 24 |
| Figure 2. Various features in the dataset after Pre-Processing | 25 |
| Figure 3. Splitting as Training snd Testing | 26 |
| Figure 4. Confusion Matrix | 27 |
| Figure 5. Performance Outcomes | 28 |
| Figure 6.  Logistic Regression-Confusion Matrix | 29 |
| Figure 7. Random Forest-Confusion Matrix | 29 |
| Figure 8. Support Vector Machine (SVM) -Confusion Matrix | 30 |
| Figure 9. Confusion Matrix – Stack ensemble learning | 30 |
| Figure 10. ROC Curve | 31 |
| Figure 11. Confusion Matrix – Stack ensemble learning | 58 |
|  |  |

**LIST OF TABLES**

Table 1.Recorded Results for each Classifier 31

# CHAPTER-1 INTRODUCTION

### INTRODUCTION

**1.1 Background and Significance of Speech Emotion Recognition**

Speech Emotion Recognition (SER) is an interdisciplinary field that focuses on analyzing and understanding human emotions based on speech signals. Emotions are an essential component of communication, influencing not only the meaning of spoken words but also how messages are perceived and interpreted. Unlike text-based sentiment analysis, which relies solely on written content, SER incorporates vocal features such as tone, pitch, rhythm, and intensity to identify emotions more accurately.

The foundation of SER lies in speech signal processing and machine learning (ML) techniques. By extracting relevant acoustic and prosodic features—such as Mel-Frequency Cepstral Coefficients (MFCCs), spectral features, and pitch contours—machine learning models can classify speech into different emotional states, such as happiness, sadness, anger, fear, and neutrality. Over the years, advancements in deep learning and neural networks have further enhanced the accuracy and robustness of SER systems.

**Significance of Speech Emotion Recognition**

The ability to automatically recognize emotions from speech has profound implications across multiple industries and applications:

Enhancing Human-Computer Interaction (HCI): Emotion-aware virtual assistants (e.g., Siri, Alexa, Google Assistant) can provide more natural and personalized interactions.

Chatbots and AI-driven customer service agents can detect user frustration or satisfaction, improving user experience.

Mental Health and Well-being: SER can assist in detecting emotional distress, anxiety, and depression by analyzing speech patterns.

It enables early intervention in mental health care by identifying signs of emotional imbalance in patients.

Customer Service and Business Analytics: Call centers and customer service departments can use SER to analyze customer emotions during interactions, helping agents respond more empathetically.

Businesses can assess customer sentiment to improve service quality and customer satisfaction.

Education and E-Learning: Adaptive learning platforms can adjust teaching methods based on students' emotions (e.g., detecting frustration or engagement).

It helps in evaluating students' confidence levels in spoken assessments.

Security and Surveillance: SER can be integrated into security systems to detect stress or fear in individuals, aiding law enforcement.

It can enhance lie detection systems by analyzing emotional cues in speech.

**1.2 Overview of Machine Learning in Speech Emotion Recognition**

Machine Learning (ML) has revolutionized **Speech Emotion Recognition (SER)** by enabling computers to analyze speech signals, extract meaningful patterns, and classify emotions with high accuracy. ML-based SER systems enhance human-computer interaction by making machines more responsive to human emotions, contributing to fields such as healthcare, customer service, security, and entertainment.

**Machine Learning Applications in Speech Emotion Recognition:**

**Speech Signal Processing:**

Feature Extraction: ML models analyze speech features such as pitch, intensity, formants, Mel-Frequency Cepstral Coefficients (MFCCs), Spectral Contrast, Chroma features to capture emotion-related variations in speech signals.

**Emotion Classification Models:**

In this project, Mel-Frequency Cepstral Coefficients (MFCCs) are used as the primary feature extraction technique to capture the essential acoustic properties of speech signals relevant to emotion recognition. Once extracted, these features are used to train a stacked ensemble model comprising multiple machine learning algorithms. At the base level, Random Forest, Logistic Regression, and Support Vector Machine (SVM) are employed to independently learn from the data and make initial predictions. These predictions are then passed to meta-level classifiers - again including SVM and Random Forest—which combine the outputs to produce the final emotion classification. This layered approach enhances model accuracy, robustness, and generalization across different emotional categories and speaker variations.

**Healthcare and Mental Health Applications:**

Mental Health Monitoring: SER can analyze emotional distress in speech patterns to detect anxiety, depression, and mood disorders in patients.

Therapeutic Support: AI-driven virtual assistants can provide emotional support and track patients’ emotional states over time.

Speech Therapy: Emotion-aware speech therapy applications assist individuals with speech impairments by recognizing emotional cues and adjusting feedback accordingly.

**Human-Computer Interaction (HCI):**

Virtual Assistants and Chatbots: ML-powered assistants (e.g., Alexa, Google Assistant, Siri) can recognize emotions in speech and respond accordingly, making interactions more natural.

Customer Service Enhancement: Call center analytics leverage SER to detect frustration, dissatisfaction, or satisfaction in customer interactions, helping companies improve service quality.

**1.3 Research Objectives and Scope**

The objectives focus on enhancing emotional intelligence in human-computer interactions, improving mental health monitoring, and expanding the real-world applications of SER in various industries.

Enhance Emotion Detection Accuracy:

Develop machine learning models that improve the precision of emotion classification from speech by leveraging advanced feature extraction techniques, machine learning architectures, and multimodal approaches.

Develop Real-Time Emotion Recognition Systems:

Investigate how real-time speech emotion recognition can be achieved using low-latency processing techniques, making it viable for live applications such as virtual assistants and call center monitoring.

Improve Robustness Across Speaker Variability and Noisy Environments:

Address challenges related to variations in accent, language, speaking style, and background noise to develop generalized SER models that work effectively across different users and environments.

Apply SER to Mental Health Monitoring:

Explore how machine learning can detect emotional distress, depression, and anxiety from speech patterns, enabling early intervention in healthcare and psychological therapy applications.

Optimize SER for Personalized User Experiences:

Investigate methods for customizing emotion recognition models to individual speakers, allowing AI-driven applications to better understand and respond to user-specific emotional cues.

Expand SER Accessibility in Low-Resource Settings:

Develop SER models that can work on low-computational devices or offline systems, making them accessible in remote areas, mobile applications, and low-resource settings.

Reduce Bias and Improve Model Generalization:

Identify and mitigate biases in emotion recognition by training models on diverse datasets covering different demographics, languages, and emotional expressions.

**Research Scope**

1. Machine Learning Algorithms for SER

Exploring various machine learning including:

* 1. Traditional ML models: Support Vector Machines (SVM), Random Forest, Logistic Regression (LR).
  2. Stack Ensemble: Combines Random Forest, Logistic Regression, and Support Vector Machine as base learners. Their predictions are integrated by meta-level classifiers.

2. Applications of Speech Emotion Recognition

1. Healthcare & Mental Health: Detecting depression, anxiety, and stress through voice analysis.
2. Human-Computer Interaction: Enhancing voice assistants, chatbots, and robotics with emotional intelligence.
3. Customer Service & Business Analytics: Monitoring customer emotions in call centers to improve service.
4. Security & Forensic Analysis: Identifying emotions in criminal investigations, emergency response, and surveillance.
5. Education & E-Learning: Adapting educational content based on students' emotional states.
6. Entertainment & Gaming: Emotion-driven gaming experiences and personalized content recommendations.

3. Sources of Data

Utilizing diverse datasets for training and evaluation, including:

1. Speech emotion datasets (e.g., RAVDESS, EMO-DB, IEMOCAP, CREMA-D).
2. Real-world audio recordings from phone calls, interviews, and emotional speech databases.
3. Multimodal datasets integrating speech with facial expressions and physiological signals.

4. Ethical and Legal Considerations

Addressing issues such as:

1. User privacy and consent in collecting and analyzing speech data.
2. Bias and fairness in emotion recognition models across different demographic groups.
3. Compliance with data protection regulations like GDPR for handling voice recordings.

5. Challenges and Limitations

Addressing key challenges in SER, including:

1. Emotion ambiguity: Some emotions are subjective and vary across individuals.
2. Cross-linguistic variations: Emotional expressions differ across languages and cultures.
3. Background noise interference: Improving model performance in real-world noisy environments.

6. Model Evaluation Metrics

Assessing the performance and reliability of SER models using:

1. Accuracy, Precision, Recall, F1-score, ROC-AUC.
2. Confusion matrices to analyze misclassification patterns.
3. Real-time efficiency for deployment in interactive applications.

7.Impact on Technology and Society

Evaluating the broader impact of SER, including:

* + 1. Enhancing AI-driven interactions by making machines more empathetic.
    2. Improving mental health diagnostics through non-invasive voice analysis.
    3. Reducing stress in workplaces by monitoring employee emotional states.

8.Integration with Existing Technology

Investigating the incorporation of SER into:

1. Smart assistants (e.g., Siri, Alexa, Google Assistant).
2. AI-powered customer support for sentiment-aware conversations.
3. Healthcare applications for tracking emotional well-being over time.
4. Smart devices and IoT for emotion-aware user interactions.

**1.4 Current Challenges in Speech Emotion Recognition**

Despite the significant advancements in machine learning and artificial intelligence, Speech Emotion Recognition (SER) faces numerous challenges that affect its accuracy, generalizability, and real-world applicability. These challenges stem from the complexity of human emotions, variations in speech patterns, environmental factors, and technological limitations.

1. Variability in Emotional Expression

Subjectivity of Emotions: Emotions vary across individuals and cultures, making it difficult to build a universal SER model.

2. Impact of Linguistic and Cultural Differences

Cross-Language Generalization: SER models trained on a single language struggle to generalize across different languages and cultural contexts.

3. Challenges in Data Collection and Annotation

Lack of Large, High-Quality Datasets: Emotion datasets are limited and often biased, affecting model performance on diverse speakers.

4. Background Noise and Real-World Conditions

Noise and Environmental Factors: Background noise in real-world settings reduces the accuracy of SER models, making emotion detection unreliable.

5. Generalization and Model Adaptability

Overfitting to Specific Datasets: Models perform well on training data but fail to generalize to real-world or unseen speech data.

6. Ethical and Privacy Concerns

Privacy Issues in Speech Data Collection: Voice recordings contain sensitive personal data, raising privacy and security concerns.

7. Lack of Standardized Evaluation Metrics

Inconsistencies in Performance Metrics: Different studies use varied evaluation metrics, making model comparison difficult.

8. Limited Deployment of AI in Commercial Applications

Integration with Existing Systems: SER models lack seamless integration with virtual assistants, chatbots, and customer service platforms.

9. User Acceptance and Trust Issues

Resistance to Emotion Monitoring: Many users distrust AI emotion analysis, especially in workplace or surveillance applications.

10. High Computational Requirements

Real-Time Processing Challenges: SER models require high computational power, limiting deployment on mobile and embedded systems.

**1.5 Applications of Machine Learning in Speech Emotion Recognition (SER)**

Machine Learning (ML) has significantly transformed Speech Emotion Recognition (SER) by improving accuracy, adaptability, and real-time emotion detection. Through the analysis of acoustic features, linguistic patterns, and contextual cues, ML models have enhanced the ability to recognize human emotions. This has contributed to advancements in multiple sectors, including healthcare, customer service, education, and human-computer interaction.

One of the primary applications of ML in SER is the extraction and classification of audio features. Spectral and prosodic attributes such as pitch, tone, intensity, and rhythm are analyzed by ML algorithms to detect emotional variations in speech. Models such as Random Forest, Logistic Regression, and Support Vector Machines (SVM) are commonly used to classify these features into distinct emotional categories. Additionally, the integration of multiple machine learning models through ensemble techniques—such as stacking—has further improved classification performance by leveraging the strengths of different algorithms. Multimodal emotion detection, which integrates speech with facial expression and text analysis, also contributes to a more holistic understanding of emotions in human interactions.

Real-time emotion detection plays a critical role in human-computer interaction. AI-powered voice assistants, such as Alexa, Siri, and Google Assistant, utilize SER to respond empathetically based on user emotions. Similarly, human-robot interaction (HRI) in customer service, healthcare, and entertainment sectors incorporates SER to enhance engagement and response effectiveness, making interactions more natural and intuitive.

The integration of ML in SER has also contributed significantly to mental health and well-being monitoring. Emotion recognition algorithms are used to detect early signs of depression and anxiety by analyzing vocal changes such as slower speech rates and monotonous voice tones. SER is also employed in suicide prevention helplines, where distress detection helps in early intervention. In therapeutic applications, AI-driven emotional feedback supports mental health professionals by tracking mood variations over time, enabling more personalized treatment approaches.

In customer service, SER enhances sentiment analysis in call centers, helping businesses assess customer emotions during interactions. By identifying frustration or dissatisfaction in a caller’s tone, AI-driven systems can trigger real-time adjustments in responses to improve customer satisfaction. Automated complaint resolution benefits from SER by prioritizing urgent cases and ensuring that dissatisfied customers receive immediate attention. Furthermore, call center agents' performance is monitored using ML, providing insights into their tone, empathy, and communication effectiveness.

Education has also benefited from ML-driven SER, particularly in adaptive learning environments. Online education platforms leverage SER to adjust teaching methods based on students’ emotions, making learning experiences more engaging and effective. By detecting frustration or confusion in students’ speech, teachers in virtual classrooms can offer additional support when necessary. AI-powered tutors further personalize learning by modifying explanations and pacing according to the student’s emotional state, thereby improving comprehension and knowledge retention.

SER has critical applications in security and forensic investigations. In law enforcement, lie detection techniques analyze speech patterns to identify stress, anxiety, or deception during interrogations. In emergency response systems, AI-driven SER detects distress signals in 911 calls, ensuring prompt and appropriate intervention. Additionally, the banking sector utilizes SER for fraud detection by identifying nervousness or hesitation in speech patterns, helping financial institutions mitigate risks associated with fraudulent transactions.

The entertainment and gaming industries have integrated SER for personalized experiences. Streaming services analyze user emotions to recommend movies, music, or shows that align with their mood. In gaming, AI adapts the level of difficulty and character interactions based on player emotions, creating more immersive and interactive gameplay experiences.

The development of language-agnostic models has improved the application of SER across multiple languages and cultural contexts. Machine learning techniques are designed to generalize emotion recognition beyond a single language, making SER systems more effective for diverse populations. Additionally, cultural adaptation of SER models ensures that regional variations in speech and emotional expressions are considered, leading to more accurate recognition outcomes.

The benefits of ML-driven SER are evident in its improved emotion detection accuracy, real-time processing capabilities, and personalized user experiences. Ensemble methods and feature-based classification models improve recognition by capturing a variety of emotional cues. Real-time SER systems provide instant emotional feedback, making them highly effective in applications requiring immediate response. Furthermore, the adaptability of ML models ensures that SER can be tailored to individual speech styles, resulting in more precise and context-aware emotion recognition.

Despite these advancements, several challenges remain in implementing ML for SER. The availability and quality of labeled emotional speech datasets continue to pose difficulties, as biased or limited datasets can affect model performance. Background noise and variability in real-world environments present additional challenges, making accurate emotion detection more complex. Ethical concerns regarding user privacy and data security must also be addressed, particularly in applications that involve sensitive personal data. Moreover, the generalization of SER models across different contexts is difficult, as models trained on controlled datasets may not perform well in diverse real-world scenarios. Another key challenge is the computational cost associated with real-time ML model deployment, which requires efficient resource management.

Machine Learning is revolutionizing Speech Emotion Recognition by improving human-computer interaction, mental health monitoring, customer service, and security applications. While challenges such as data quality, privacy concerns, and computational demands remain, advancements in feature extraction techniques, ensemble learning, and real-time processing will continue to drive innovation in emotion-aware AI systems. As ML models become more refined and adaptive, their integration into various industries will enhance emotional intelligence in technology, making human-machine interactions more seamless and effective.

# CHAPTER-2 LITERATURE SURVEY

## 2. LITERATURE SURVEY

#### Literature review

Speech Emotion Recognition (SER) enables machines to identify human emotions from speech signals. This project develops an SER system using machine learning to classify emotions like happiness, sadness, anger, fear and neutrality. Key features including pitch and energy of speech are extracted. The system is trained on datasets and evaluated using accuracy, precision, recall and F1 score. Potential applications include sentiment analysis, virtual assistants and mental health with an emphasis on data processing, feature extraction, and classification, the paper examines many Machine Learning (ML) approaches in Speech Emotion Recognition (SER). IEMOCAP, SAVEE, RAVDESS, Emo-DB, and CASIA are examples of frequently used datasets. Prosodic, spectral, and deep characteristics from speech are used by machine learning algorithms like SVM, k-NN, GMM, RNN, and neural networks to categorize emotions. Speaker dependency, emotion overlap, and reduced accuracy compared to speaker independent models are some of the issues that SER models must deal with. While performance varies by dataset and algorithm, RNNs and Deep Learning methods typically outperform more conventional machine learning models such as SVM and k-NN.[1]

Classifiers and feature selection techniques are compared in a number of SER research. For instance, SVM, MLR, RNN, MFCC, and Recursive Feature Elimination (RFE) were used for feature selection in a study on the Spanish Database (6041 utterances, 2 speakers, 7 emotions) and the Berlin Database (535 utterances, 10 speakers, 7 emotions). The results show that feature selection affects accuracy; the Spanish dataset achieved 94% accuracy with RNN and feature selection, while the Berlin dataset achieved 83% accuracy with feature selection and speaker normalization. [2]

The Berlin Emotional Speech Database (493 utterances, 10 speakers, 7 emotions) was used in another work to assess kernel functions and develop SER using SVM. With MFCC and MEDC for feature extraction, the RBF and Polynomial kernels in the SVM model demonstrated varying accuracy: the RBF Kernel reached up to 100% accuracy for female speakers and 93.75% for gender-independent classification, while the Polynomial Kernel reached 100% for female speakers and 96.25% for gender-independent classification.[3]

Several datasets, such as IEMOCAP, Emo-DB, SAVEE, and eNTERFACE05, have been used to compare deep learning approaches—such as CNN, RNN, LSTM, DBN, and DBM—with conventional machine learning techniques. These models have drawbacks such high computational cost, overfitting, and the requirement for huge datasets, despite the fact that they provide automatic feature extraction and better emotion classification. Although accuracy varies depending on the dataset, it has been demonstrated that CNN in conjunction with LSTM performs better than typical ML models.[4]

Another study that used Deep Learning to analyze SER on datasets like EMO-DB, SAVEE, and RAVDESS discovered that DNN, CNN, RNN, and DBN models outperformed conventional ML techniques, achieving up to 97% accuracy on EMO-DB. [5]

Additionally, hybrid models have been investigated, such as the use of DBN and HMM to enhance the recognition of emotions in unscripted speech. A hybrid DBN-HMM classifier was used for temporal emotion modeling utilizing 39-dimensional MFCC features in a study on the FAU Aibo Dataset (51 children, 9,959 training utterances, 8,257 test utterances, and 5 emotions). This method outperformed earlier models (44.8%) with a best-case accuracy of 46.36% with speaker normalization. [6]

Using ML approaches on the TESS (2800 samples) and KEEL (593 samples) datasets, another study created a SER system. MFCC, Chroma, and Spectral Features were combined with SVM, Decision Tree, KNN, Logistic Regression, RF, and Gradient Boosting. With 92% accuracy on TESS, the SVM model outperformed the others; but, because of the reduced sample size on KEEL, accuracy decreased to 67%.[7]

SER has also investigated hybrid feature extraction in conjunction with CNN utilizing Berlin Emo-DB (535 data), SAVEE (480 samples), and RAVDESS (1440 samples). A hybrid method that included Time-Domain Features (MFCCT) and MFCC was evaluated against RF, SVM, and k-NN classifiers. CNN achieved 97% accuracy on EMO-DB, 93% accuracy on SAVEE, and 92% accuracy on RAVDESS, according to the results. [8]

Surpassing machine learning models. A different SVM-based SER system with an RBF Kernel used Energy, Pitch, LPCC, MFCC, and LPCMCC characteristics with Berlin EMO-DB (500 samples, 7 emotions). This model achieved a maximum accuracy of 82.5%, with Energy Pitch obtaining 66.02% and LPCMCC reaching 70.7%, despite issues with speaker reliance and overfitting.[9]

A different method for SER used a bespoke dataset of 60 speakers and three emotions (happy, angry, and sad) in conjunction with MFCC feature extraction and standard deviation-based classification. The mean and standard deviation of MFCCs were used for classification, together with FFT, MFCC, and a Hamming Window for feature extraction. Even in noisy surroundings, the model reached 80% efficiency despite having few classes and overlapping emotions.[10]

#### 2.2 Motivation

The growing significance of human-computer interaction (HCI) and affective computing underscores the need for Speech Emotion Recognition (SER). Emotions play a crucial role in communication, influencing both verbal and non-verbal exchanges. Enabling machines to accurately interpret human emotions can significantly enhance artificial intelligence applications, particularly in virtual assistants, mental health monitoring, and sentiment analysis. Despite advancements in speech processing, traditional emotion detection techniques face challenges such as low accuracy, speaker dependency, and difficulty in handling overlapping emotions. These limitations highlight the necessity of employing machine learning (ML) techniques to improve the efficiency and accuracy of emotion classification from speech signals.

Machine learning algorithms offer powerful solutions for SER by analyzing patterns in speech and categorizing emotions based on extracted features. Models such as Support Vector Machines (SVM), Random Forest(RF), Logistic Regression(LR) have been widely applied in this domain. These models rely on different speech characteristics, including prosodic features like pitch, energy, and duration, as well as spectral features such as Mel-Frequency Cepstral Coefficients (MFCCs) , Chroma and Spectral-contrast .

Preprocessing and feature extraction play a critical role in enhancing SER performance. Noise reduction techniques, such as Gaussian and median filtering, help refine raw speech signals by eliminating distortions that may affect model accuracy. Segmentation methods, including voice activity detection (VAD), facilitate the isolation of speech components relevant to emotion classification. These preprocessing steps contribute to improving model robustness and reducing misclassification errors.

The increasing demand for real-time emotion recognition and context-aware AI systems has motivated ongoing research in SER. Future advancements in transformer-based architectures, multi-modal learning that integrates speech with facial expressions, and transfer learning approaches are expected to enhance the generalization ability of emotion recognition models. By combining machine learning techniques with improvements in speech processing, SER research aims to develop more reliable and intelligent systems. These innovations will not only enhance user experience in interactive AI applications but also contribute to critical fields such as mental health support and human-robot interaction, paving the way for emotionally aware AI-driven technologies.

# CHAPTER-3

# PROPOSED SYSTEM

### 3. PROPOSED SYSTEM

**3.1 Input dataset**

**A.** Four native English speakers (DC, JE, JK, and KL) who were postgraduate students and researchers at the University of Surrey and ranged in age from 27 to 31 years provided information for the SAVEE database. Anger, disgust, fear, happiness, sadness, and surprise are the distinct categories into which psychology has classified emotion. In order to provide recordings of seven emotion categories, a neutral category has also been added. There were 15 TIMIT sentences per emotion in the text: 3 common, 2 emotion-specific, and 10 generic sentences that were phonetically balanced and unique for each emotion. Thirty neutral sentences were created by recording the three common and two × six = twelve emotion-specific sentences. As a result, each speaker produced 120 utterances in total.

**B.** Data Preprocessing: Four native English speakers (DC, JE, JK, and KL) who were postgraduate students and researchers at the University of Surrey and ranged in age from 27 to 31 years provided information for the SAVEE database. Anger, disgust, fear, happiness, sadness, and surprise are the distinct categories into which psychology has classified emotion. In order to provide recordings of seven emotion categories, a neutral category has also been added. There were 15 TIMIT sentences per emotion in the text: 3 common, 2 emotion-specific, and 10 generic sentences that were phonetically balanced and unique for each emotion. Thirty neutral sentences were created by recording the three common and two × six = twelve emotion-specific sentences. As a result, each speaker produced 120 utterances in total.

**C.** Feature Extraction: identifies speech emotions by extracting features from preprocessed audio files. Following the retrieval of speaker and emotion labels, three different feature types are extracted: CNN-based spectrogram features, handcrafted features (MFCCs, spectral contrast, and chroma). After being combined, the extracted features are saved as a CSV file for additional examination.

**D.**Apply ML Models: Different machine learning models can be applied to classification, identification, and clustering problems after features are collected and dimensionality is decreased. An outline of the models you described is provided below:

1. Load the necessary libraries Import the Python libraries required for data processing, visualization, and machine learning.

2. Dataset Split Divide the data into two sets: 20% for testing and 80% for training.

3. Preprocess and load the dataset Open a CSV . Use resampling, apply filters, noise removal for preprocessing .

4. Apply feature extraction techniques like MFCC, spectral, Chroma and CNN.

5. Specify the models and hyperparameters Explain SVM, Random Forest, and Logistic Regression. Use GridSearchCV to adjust models by setting hyperparameter grids.

6. Develop and Assess Every Model Train every model (some may require hyperparameter tuning). Forecast the test set’s emotions. Determine the accuracy and show the classification report. Plot a matrix of confusion to visualize performance.

7. Develop a Classifier for Voting Combine Naive Bayes, SVM, Input dataset

#### Detailed Features of the Dataset

The **SAVEE (Surrey Audio-Visual Expressed Emotion) dataset** was developed to support research in emotion recognition using speech and facial expressions. It contains audio-visual recordings of emotional speech performed by four male speakers who were postgraduate students and researchers at the University of Surrey. Below are the detailed features of the dataset:

1. General Information

Dataset Name: SAVEE (Surrey Audio-Visual Expressed Emotion)

Institution: University of Surrey

Speakers: 4 native English speakers (DC,JE, JK, KL)

Speaker Details: Postgraduate students and researchers

Age Range: 27 to 31 years

Total Number of Utterances per Speaker: 120

Total Number of Utterances in the Dataset: 120×4=480120 \times 4 = 480120×4=480

2. Emotion Categories

SAVEE includes seven distinct emotions, plus a neutral category:

|  |  |
| --- | --- |
| Anger (A) | Expressing strong displeasure or hostility |

|  |  |
| --- | --- |
| Disgust (D) | Feeling of revulsion or distaste |

|  |  |
| --- | --- |
| Fear (F) | Expressing anxiety or threat perception |

|  |  |
| --- | --- |
| Happiness (H) | State of joy or pleasure |

|  |  |
| --- | --- |
| Sadness (S) | State of sorrow or unhappiness |

|  |  |
| --- | --- |
| Surprise (Su) | Expressing astonishment or unexpectedness |

|  |
| --- |
| Neutral (N) Emotionless |
| Each emotion is represented by 15 sentences per speaker, leading to a total of 15×7=10515 \times 7 = 10515×7=105 sentences per speaker for emotions. Additionally, 30 neutral sentences are recorded per speaker.  3. Sentence Structure  Total Sentences per Emotion: 15  3 Common Sentences (same for all emotions)  2 Emotion-Specific Sentences (tailored to each emotion)  10 Unique, Phonetically Balanced Sentences (different for each emotion)  Total Neutral Sentences: 30  3 Common Sentences  12 Emotion-Specific Sentences (2 per emotion × 6 emotions)  Each speaker recorded 120 utterances:  105 emotional sentences  30 neutral sentences  Total dataset size: 480 utterances (4 speakers × 120 sentences each)  4. Modalities in the Dataset  The dataset consists of both audio and visual components:  Audio Recordings (speech data)  Visual Recordings (facial expressions)  The video files include head-mounted microphone audio and frame-synchronized facial expressions, which can be used for multimodal emotion recognition.  5. File Format and Annotations  Audio Format: WAV files (high-quality recordings)  Video Format: AVI files (RGB facial expressions)  Annotations:  Each recording is labelled with the corresponding emotion category.  Phoneme transcriptions and timestamps are available.  Metadata includes speaker ID and sentence type.  6. Applications of the SAVEE Dataset  The SAVEE dataset is widely used for:  Speech Emotion Recognition (SER) models.  Multimodal Emotion Recognition using facial expressions and speech.  Deep Learning and Machine Learning applications in affective computing.  Human-Computer Interaction (HCI) research.  Psychological and Linguistic Studies on emotional speech. |

#### 3.2Data Pre-processing

Data pre-processing is the essential process of preparing raw data for analysis and modelling by cleaning, transforming, and structuring it to enhance data quality and utility. It involves tasks like handling missing values, correcting errors, encoding features, and scaling data to ensure it's in an optimal form for further analysis. It encompasses a range of operations and transformations designed to refine raw data, ensuring that it is clean, structured, and amenity subsequent analysis. This process is driven by its manifold significance in data science and analysis.

Through meticulous data cleaning, transformation, feature engineering, dimensionality reduction, outlier handling, scaling, and data splitting, it prepares raw data for more accurate and reliable analysis and modelling. Ultimately, the goal is to obtain more meaningful insights, make informed decisions, and optimize predictive models for a wide range of applications in data science and analysis.

The system processes audio files from the SAVEE dataset to prepare them for emotion recognition using speech features. It begins by importing necessary libraries, including librosa for audio processing, scipy.signal for noise reduction, and transformers to load the Wav2Vec2 model, which is useful for speech feature extraction. The script defines input and output directories, ensuring that the processed audio files are stored in a designated folder, and also maintains a log file (preprocessed\_files.txt) to record the details of the processed files.

Next, it maps emotion codes from filenames to their corresponding emotions using a predefined dictionary. Each filename in the dataset follows the format "Speaker\_EmotionCodeXX.wav", where the first part represents the speaker ID (e.g., DC, JE, JK, KL) and the second part encodes the emotion. A function is implemented to extract these details, ensuring correct classification of each audio file into one of the seven emotions (anger, disgust, fear, happiness, neutral, sadness, and surprise).

To enhance the quality of the audio data, the script applies three preprocessing steps: (1) Wiener filtering for noise reduction, (2) normalization to maintain a consistent amplitude level across recordings, and (3) resampling all audio files to 16 kHz, the optimal sampling rate for Wav2Vec2-based models. These steps ensure that the audio data is clean and standardized for further analysis.

After preprocessing, the script iterates through all .wav files in the dataset, processes each file, and saves the cleaned version in the output directory. It logs each file’s speaker, emotion, and saved path into a text file for future reference. Finally, it prints a completion message, confirming that all audio files have been successfully processed and saved.

#### Model Building

The next step is to build an emotion recognition model. The process involves preprocessing, feature extraction, model selection, training, evaluation, and deployment. Below is a structured workflow for building the model.

we need to prepare the dataset along with pre-processing by storing the features along with their corresponding emotion labels. This structured dataset is usually saved in a CSV file, making it easier to train and evaluate models. The dataset is then split into training and testing sets,and apply pre-processing typically in an 80-20 ratio, ensuring that the model can generalize well to unseen data. Labels are also encoded into numerical values using LabelEncoder for compatibility with machine learning models.

The next step is feature extraction, where we convert raw audio signals into meaningful numerical representations. One common approach is to use Mel-Frequency Cepstral Coefficients (MFCCs), Chroma Features, Spectral Contrast which capture important frequency characteristics of human speechThese extracted features are compiled into a feature matrix and stored in a CSV file. This matrix is then used for training and testing various machine learning classifiers. By combining spectral, prosodic, and temporal features, the system achieves a rich representation of the emotional content in speech.

For model training, we can use either machine learning. A stacked ensemble learning model is used for classification. At the base level, Random Forest, Logistic Regression, and Support Vector Machine (SVM) are trained independently to make initial predictions. These base models are chosen for their complementary strengths in handling non-linearity, high dimensionality, and linearly separable data. The outputs (predicted probabilities) of these base classifiers are then fed into meta-level learners—SVM and Random Forest—which combine these inputs to make the final classification decision. This two-tiered structure enhances generalization and reduces overfitting.

After training, the model is evaluated using metrics such as accuracy, classification report, and confusion matrix. This helps us understand the model’s performance in distinguishing between different emotions.

The model was used to forecast the test set's cancer severity after it had been trained. The model's fit to the data was evaluated by calculating both training and testing accuracies. While the training accuracy gauges how well the model learned from the training data, the testing accuracy offers information about how well the model performs on fresh, unseen data.

Important metrics including accuracy, precision, recall, and F1-score were calculated in order to assess the model further. A thorough understanding of the model's performance is offered by these metrics:

Accuracy gauges how accurate the model is overall.

The number of projected positive cases (such as high severity) that were actually true is known as precision.

The model's recall indicates how effectively it represented every real positive instance.

The F1-score is helpful when the dataset is unbalanced since it offers a balance between precision and recall.

#### 3.4 Methodology of the system

The proposed system architecture for recognizing emotions from speech involves several interconnected steps, including data collection, preprocessing, feature extraction, model training, and classification. The architecture consists of the following layers:

Input Layer: The system collects speech recordings from the dataset, such as the SAVEE dataset, containing utterances labeled with emotions.

Preprocessing Layer: This stage involves noise reduction, normalization, and resampling to enhance the quality of speech signals before feature extraction.

Feature Extraction Layer: Extracts meaningful representations such as Mel-Frequency Cepstral Coefficients (MFCCs), spectrogram features.

Classifier: A machine learning (SVM, Naïve Bayes, Random Forest) predicts the emotion based on extracted features.

Output Layer: Displays the classified emotion (e.g., Anger, Happiness, Sadness, etc.) based on the processed input speech signal.

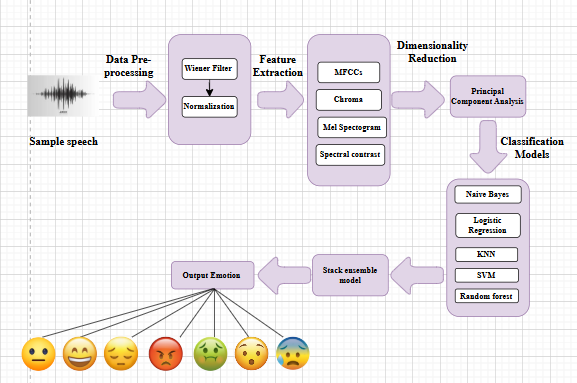


Figure 1. Architecture of the proposed system

B. Training and Preprocessing of Data

Proper data preprocessing ensures the dataset is suitable for machine learning models. The preprocessing steps include:

Data Cleaning: Unnecessary metadata, low-quality recordings, and silent sections are removed.

Label Encoding: The emotion labels are converted into numerical values (e.g., Anger = 0, Happiness = 1, etc.) for compatibility with machine learning models.

Feature Scaling: Normalization ensures that all extracted features have a similar range, preventing certain features from dominating the learning process.

Data Splitting: The dataset is divided into training (80%) and testing (20%) sets to ensure generalization to unseen speech samples.

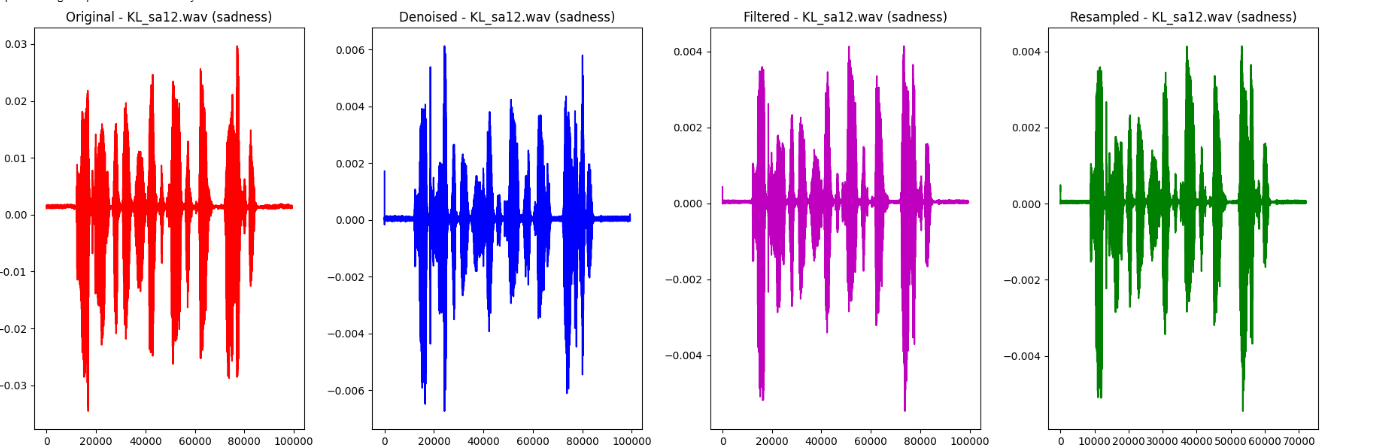


Figure 2. Sample after Pre-Processing

C. Extraction of Features

Feature extraction transforms raw speech signals into relevant numerical representations that can be used for classification. Key extracted features include:

MFCCs: Capture speech characteristics based on human auditory perception.

Chroma Features: Represent tonal properties of the speech signal.

Spectral Features: Include spectral centroid, roll-off, and bandwidth, providing insights into frequency distribution.

D. Classification Model

A classifier is trained to predict the emotion category from extracted features.

E. Classification

The trained model classifies speech into one of the emotion categories (Anger, Disgust, Fear, Happiness, Neutral, Sadness, Surprise). Evaluation is conducted using:

Accuracy: Measures the proportion of correctly classified emotions.

Precision, Recall, and F1-Score: Provide insights into the model's classification effectiveness.

Confusion Matrix: Visualizes misclassifications between different emotions.

F. Results

The system outputs the detected emotion for each input speech sample. Once trained, the model can predict the emotional state of new speech recordings. Healthcare, customer service, and human-computer interaction applications can benefit from this system by detecting emotions in real-time. The system’s performance is measured by classification accuracy, showing its potential for practical deployment in sentiment analysis, virtual assistants, and mental health monitoring.

#### Model Evaluation

A. Accuracy of Training and Testing

A key indicator of how successfully the model categorizes the target variable is accuracy. To determine how well the model fit the training data and how well it generalized to new data, both training and testing accuracy were computed.

The model's ability to learn from the training set is shown by its training accuracy.

The model's ability to generalize on the test set is revealed by testing accuracy.

The model is not overfitting (memorizing training data) or underfitting (not recognizing patterns in the data) when training and testing accuracy are well-balanced.

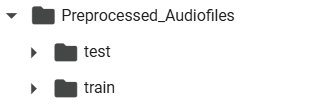


Figure 3. Splitting as Training and Testing

B. Confusion Matrix  
The model's classification performance was assessed using the confusion matrix, which offers a thorough analysis of true positives, false positives, true negatives, and false negatives. The matrix assisted in figuring out:

How often the model successfully classified each severity level.

Locations where the model misclassified a class (for example, Happy as Sad).

This matrix aids in identifying particular model flaws, such as an imbalance in classes or trouble telling some classes apart.

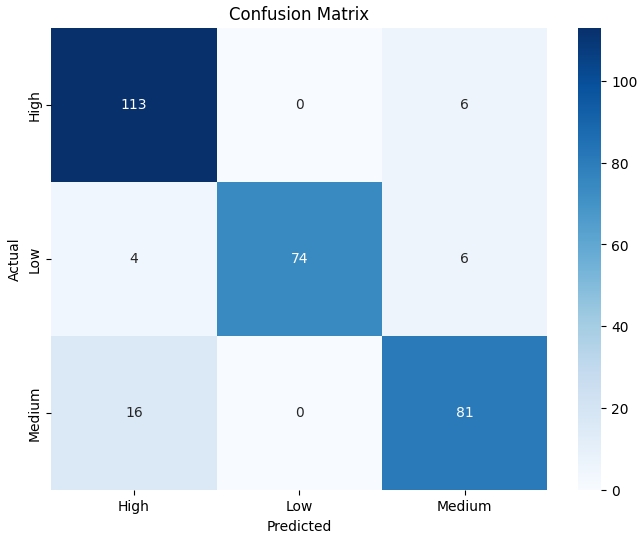


Figure 4. Confusion Matrix

C. Accuracy

Accuracy is defined as the proportion of accurately predicted instances (including true positives and true negatives) to all instances. Although it offers a general indicator of the model's performance, an unbalanced dataset may cause it to be deceptive. Here, accuracy is used as a starting point.

D. Precision

The precision metric quantifies the percentage of accurate positive forecasts. In this study, it shows the proportion of instances that fell into the emotion that was predicted. Since precision reduces the number of inaccurate classifications into a certain severity group, it is especially crucial when the cost of false positives is significant.

E. Recall

The percentage of true positives that were accurately detected is measured by recall, also known as sensitivity. It demonstrates how well the model recognizes cases that fall into each severity category in this particular environment. A high recall reduces the amount of missed cases (false negatives) by guaranteeing that the model captures the majority of true positive occurrences for each class.

F. F1-Score

The harmonic mean of recall and precision is the F1-score. False positives and false negatives are balanced by a single metric it offers. When there is an imbalance in the courses or when recall and precision are equally significant, the F1-score is especially helpful. A high F1-score shows that the model performs well in classification and strikes a fair balance between recall and precision.

G. Outcomes of Performance

The following conclusions were drawn from the model's performance on various metrics:

Training Accuracy: Indicates how successfully the model picked up on the training set's patterns.

Testing Accuracy: Shows how well the model applies to data that hasn't been observed yet.

Precision and Recall: Aided in evaluating the model's ability to correctly classify particular emotion and steer clear of incorrect classifications.

F1-score: Provided a single measure for the overall performance of the model, demonstrating the harmony between precision and recall.

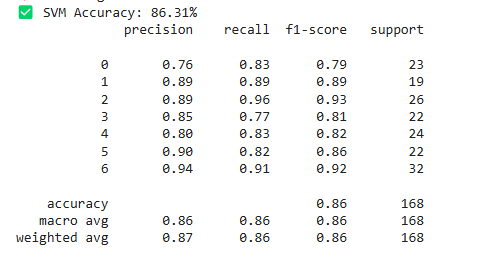


Figure 5. Performance Outcomes

According to the evaluation results, the SVM classifier is a good model for this dataset because it performs well across all severity levels and has a respectable accuracy. Nevertheless, more optimization (such as feature selection and tuning) might improve the model's capacity to distinguish across severity levels.

To see each classifier's performance, confusion matrices were plotted. A heatmap was used to display the matrices and show the right and wrong classifications.

**Logistic Regression**

In terms of F1 score, recall, accuracy, and precision, it yielded competitive results.

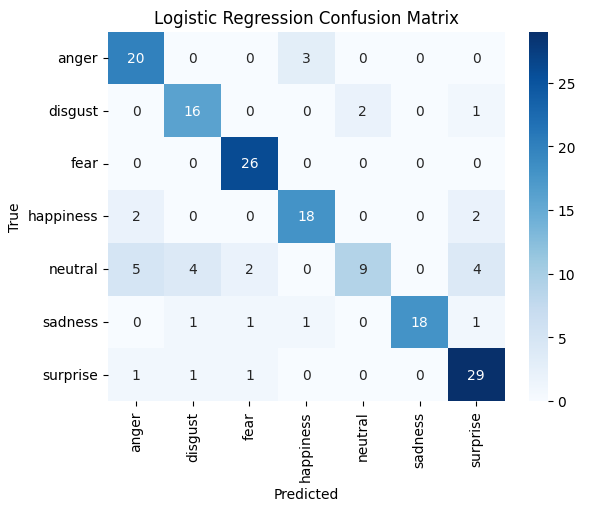


Figure 6. Logistic Regression – Confusion Matrix

**Random Forest**

Random Forest demonstrated solid performance after being trained with 100 trees (estimators=100). Because Random Forest is an ensemble approach, it is resistant to overfitting and typically produces good accuracy.

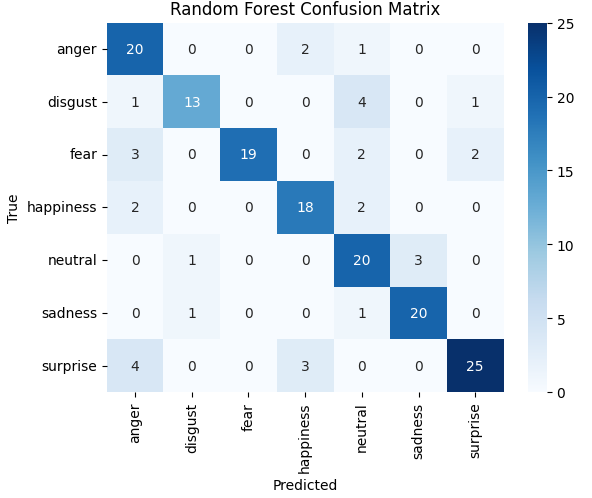


Figure 7. Random Forest – Confusion Matrix

**Support Vector Machine (SVM)**

Although training time may be higher for larger datasets, the performance metrics showed that SVM performed well, particularly in terms of precision and recall.

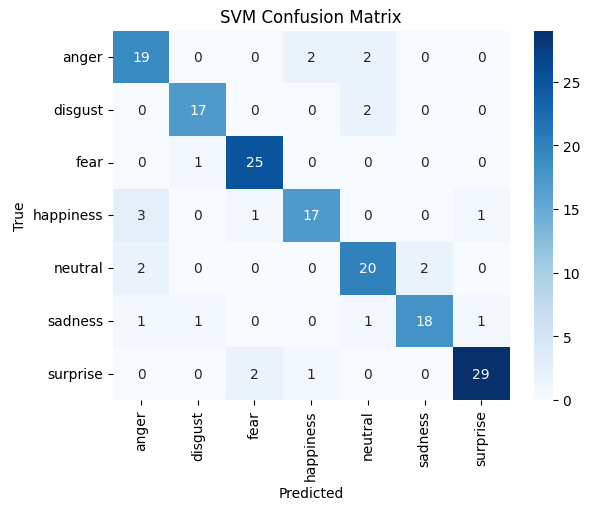


Figure 8. Support Vector Machine (SVM) - Confusion Matrix

**Stacked Model**

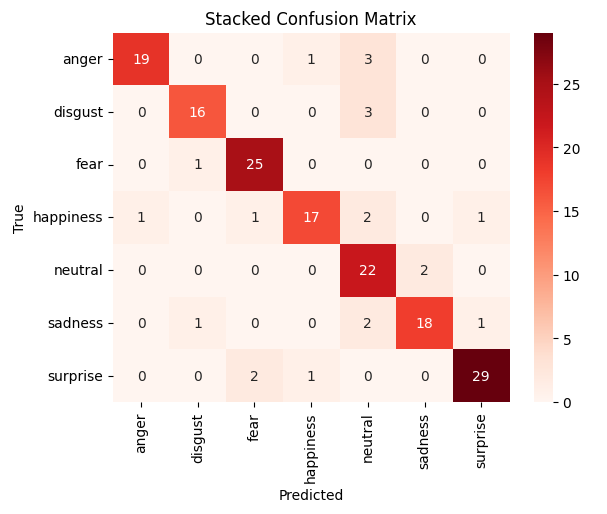


Figure 9. Stacked Model – Confusion Matrix

|  |  |
| --- | --- |
| ML MODELS | ACCURACY |
| Support Vector Machine | 86.31 |
| Random Forest | 80.36 |
| Logistic Regression | 80.95 |
| Stack Ensemble | 86.90 |

Table 1: Recorded Results for each Classifier:

The Receiver Operating Characteristic (ROC) curve:

The Receiver Operating Characteristic (ROC) curve is a graphical representation used to evaluate the performance of a classification model. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at different classification thresholds. The ROC curve helps determine how well a model distinguishes between different emotion classes in speech emotion recognition (SER).

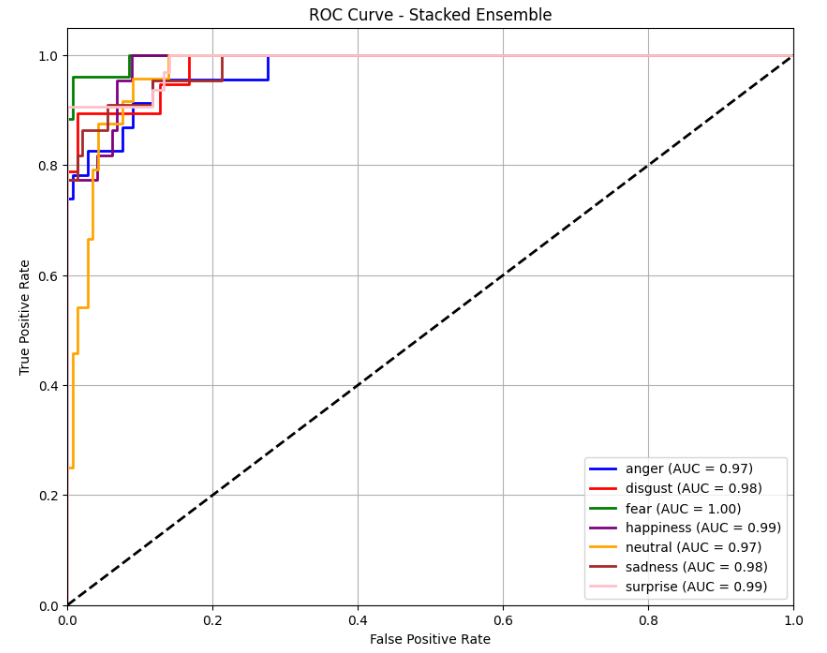


Figure 10. ROC curve

1. **Quality Assurance**: Model evaluation helps ensure that the model is capable of making accurate predictions when exposed to real-world data. It acts as a quality control mechanism to validate the model's generalization ability.
2. **Comparing Models**: Model evaluation allows for the comparison of multiple models to identify the best-performing one. It helps data scientists and stakeholders make informed decisions about which model to deploy.
3. **Fine-Tuning**: The evaluation process can reveal areas where the model performs poorly. This information is valuable for refining the model, making it more robust, and addressing its limitations.
4. **Business Decision Support**: In practical applications, model performance impacts critical business decisions. A well-evaluated model provides confidence to stakeholders, leading to better decision-making.
5. **Model Deployment**: A thoroughly evaluated model is more likely to be deployed in production systems. It instils trust in the model's predictions, which is essential in real- world applications.

### 3.6 Constraints

In the development of our Speech Emotion Recognition (SER) system, we operate within specific limitations that influence our approach to design and implementation. These constraints ensure that the system adheres to essential factors related to data, security, and performance in real-world applications:

i. **Authenticity**: We recognize that speech data may be imperfect due to environmental noise, speaker variability, and emotional context that is not always captured accurately. Some audio samples may be distorted or unclear, which could affect the performance of the model. To mitigate this risk, we implement noise reduction techniques, data augmentation, and verification steps to improve the robustness of the model and handle incomplete or ambiguous data.

ii. **Privacy**: As with any technology that processes sensitive information, privacy is a critical concern. Since speech data can sometimes contain personal or sensitive details, we adhere to strict privacy and data protection standards, ensuring compliance with regulations such as GDPR and HIPAA. Data anonymization techniques are applied to prevent the identification of individuals, and we take care to limit access to data based on need-to-know protocols to safeguard privacy.

iii. **Cost**: Although publicly available datasets may help reduce costs during the development phase, obtaining high-quality labelled speech datasets for emotion recognition can be expensive. In some cases, the costs for gathering specific emotion-labelled data or the hardware needed for recording high-quality audio can add to the budget. We seek to optimize resource usage while balancing the quality of data acquisition with project goals to ensure cost-effectiveness in model development.

iv. **Data Quality**: The accuracy and efficiency of our Speech Emotion Recognition model heavily depend on the quality of the audio data. Datasets may contain noise, non-standardized speech patterns, or imbalances in emotion representation. Ensuring high data quality requires constant attention to data cleansing, verification, and enhancement, such as improving signal-to-noise ratio and maintaining a balanced dataset for training the model.

v. **Resource Availability**: The limitations of available computational resources, including processing power and storage capacity, can affect the scale and speed of model training. Additionally, the expertise required for speech signal processing and machine learning may be limited by human resource constraints. To maximize available resources, we select efficient algorithms and methods that can provide good performance while balancing the computational load.

### 3.7 Cost and Sustainability Impact

The creation and implementation of the Speech Emotion Recognition (SER) system also involve considering the financial and long-term sustainability impacts. Below is an analysis of the project's costs and potential sustainability effects in the context of healthcare and other applications.

#### A. ****Cost Consequences****

1. **Infrastructure and Equipment**: The implementation of the SER system requires investment in hardware and software infrastructure. This includes servers, storage, and computational power, especially when training large models on extensive audio datasets. Additionally, specialized equipment for high-quality audio recording may also contribute to the initial setup costs.
2. **Costs of Operations**: Ongoing operational costs include system maintenance, software upgrades, and continuous monitoring to ensure the system’s accuracy and reliability. Moreover, hiring and training skilled personnel, including speech data annotators and machine learning engineers, can be a significant operational expense.
3. **Costs of Data Acquisition**: While some publicly available speech emotion datasets are accessible for research purposes, obtaining more specialized or proprietary datasets can be costly. These costs might involve acquiring permission from research institutions, purchasing access to private databases, or paying for crowd-sourced data collection efforts.

#### ****Benefit-Cost Analysis****

A benefit-cost analysis is essential to gauge the potential financial returns of implementing the Speech Emotion Recognition system. Some of the benefits that can offset the initial costs include:

1. **Enhanced User Experience**: For applications in customer service, mental health support, and human-computer interaction, the ability to detect emotions can lead to more personalized, empathetic responses and improved user satisfaction.
2. **Reduced Miscommunication**: In contexts like virtual meetings or telehealth, accurately detecting emotional cues can improve communication, reducing misunderstandings that may arise from tone and intent.
3. **Cost Savings in Healthcare**: Emotion recognition can play a role in mental health diagnostics and remote patient monitoring. Early identification of emotional distress or disorders could lead to more timely interventions, potentially reducing healthcare costs associated with untreated conditions.

#### ****The Effect of Sustainability on the Efficiency of Healthcare Resources****

The application of speech emotion recognition in healthcare and support services could enhance the allocation and use of healthcare resources:

1. **Mental Health Support**: SER could enable early detection of emotional distress or mental health conditions, helping healthcare providers prioritize patients who may need immediate attention or therapy. This could lead to better resource utilization in mental health services, especially in regions with limited access to professionals.
2. **Resource Allocation in Customer Support**: For organizations providing remote support (e.g., helplines), SER systems could help allocate resources by detecting emotional urgency and directing customers to the appropriate services more efficiently.

#### ****Sustainability of the Environment****

The environmental impact of implementing speech emotion recognition systems can be minimized by:

1. **Reducing the Need for Physical Interaction**: In applications like telehealth and virtual customer service, SER reduces the need for in-person interactions, thus decreasing the carbon footprint associated with travel and physical infrastructure.
2. **Energy Efficiency in Cloud-based Systems**: Leveraging cloud computing platforms for the processing and storage of audio data can improve the overall energy efficiency compared to traditional on-premise solutions. Efficient server use and resource sharing can also lower environmental impact.

#### ****Long-Term Health Outcomes****

The integration of SER into healthcare and mental health services could lead to:

1. **Improved Mental Health Outcomes**: By providing real-time emotional insights, SER systems can help healthcare providers offer more timely and personalized care, potentially improving long-term patient outcomes in mental health.
2. **Lower Healthcare Costs**: Early identification of emotional issues and more targeted interventions can lead to long-term savings by reducing the need for more intensive, expensive treatments in the future.

#### ****Community Involvement and Awareness****

By using SER systems in educational and healthcare settings, we can:

1. **Raise Awareness**: Public awareness campaigns about emotional well-being and the role of technology in mental health could encourage individuals to seek help earlier, leading to improved community health outcomes.
2. **Encourage Preventive Care**: The detection of emotional states in patient interactions may foster more preventive health behaviors and early intervention in cases of mental health distress.

#### ****Scalability and Accessibility****

A key benefit of SER technology is its ability to scale, especially in underserved regions:

1. **Remote Access to Support**: SER systems can expand access to mental health support and other emotional intelligence applications in rural or remote areas where human resources are limited.
2. **Affordable Alternatives**: By utilizing cost-effective machine learning models and cloud-based services, we can make emotion recognition technology accessible to a wider audience, promoting equity in healthcare and customer service applications.

# CHAPTER-4 IMPLEMENTATION

**4.Implementation**

# 4.1 Environment Setup

To ensure the efficient operation of our Speech Emotion Recognition (SER) models, we utilized a well-structured environment designed for audio processing, feature extraction, and machine learning tasks. The project was developed using Python, supported by key libraries for speech processing, data handling, and model training.

We employed Librosa for audio processing and feature extraction, NumPy for numerical computations, pandas for data management, and Matplotlib/Seaborn for visualization. Machine learning models were implemented using scikit-learn (for SVMs, Random Forest, and Logistic Regression).

The development environment was set up using Anaconda, facilitating seamless package management. The dataset was loaded from local storage and preprocessed by applying noise reduction, normalization, resampling, and feature extraction (MFCCs, spectrograms). The models were trained on a desktop with at least 16GB RAM and an Intel i7 processor, ensuring efficient computation and real-time inference capabilities.

# 4.2 Sample Code for Pre-processing

import os

import librosa

import numpy as np

import soundfile as sf

import matplotlib.pyplot as plt

from tqdm import tqdm

from sklearn.model\_selection import train\_test\_split

# Correct Emotion Mapping

emotion\_mapping = {

    'a': 'anger',

    'd': 'disgust',

    'f': 'fear',

    'h': 'happiness',

    'n': 'neutral',

    'sa': 'sadness',

    'su': 'surprise'

}

def remove\_noise(y, sr):

    return librosa.effects.preemphasis(y)

def apply\_filtering(y):

    kernel\_size = 5

    kernel = np.ones(kernel\_size) / kernel\_size

    return np.convolve(y, kernel, mode='same')

def resample\_audio(y, sr, target\_sr=16000):

    return librosa.resample(y, orig\_sr=sr, target\_sr=target\_sr), target\_sr

# Paths

dataset\_dir = "/content/extracted"

audio\_output\_dir = "/content/Preprocessed\_Audiofiles"

text\_output\_dir = "/content/pre-Processed\_Text"

train\_dir = os.path.join(audio\_output\_dir, "train")

test\_dir = os.path.join(audio\_output\_dir, "test")

# Ensure directories exist

os.makedirs(train\_dir, exist\_ok=True)

os.makedirs(test\_dir, exist\_ok=True)

os.makedirs(text\_output\_dir, exist\_ok=True)

processed\_files = []

# Process dataset

for root, \_, files in os.walk(dataset\_dir):

    for file in tqdm(files, desc="Processing Audio Files"):

        if file.endswith('.wav'):

            file\_path = os.path.join(root, file)

            # Extract emotion from filename (corrected for SAVEE dataset format)

            filename\_parts = file.split("\_")

            emotion\_code = filename\_parts[-1].split(".")[0]  # Extract last part before .wav

            emotion\_code = ''.join([i for i in emotion\_code if not i.isdigit()])  # Remove digits

            emotion = emotion\_mapping.get(emotion\_code, "unknown")

            if emotion == "unknown":

                print(f"Warning: Emotion not identified correctly for file {file}. Expected one of {list(emotion\_mapping.keys())}, but got '{emotion\_code}'.")

                continue

            y, sr = librosa.load(file\_path, sr=22050)

            y\_denoised = remove\_noise(y, sr)

            y\_filtered = apply\_filtering(y\_denoised)

            y\_resampled, new\_sr = resample\_audio(y\_filtered, sr)

            processed\_files.append((file, emotion, y, y\_denoised, y\_filtered, y\_resampled, new\_sr))

# Split data

if len(processed\_files) > 0:

    train\_files, test\_files = train\_test\_split(processed\_files, test\_size=0.2, random\_state=42)

    # Save processed files

    for file, emotion, \_, \_, \_, y\_resampled, new\_sr in train\_files:

        sf.write(os.path.join(train\_dir, f"preprocessed\_{file}"), y\_resampled, new\_sr)

    for file, emotion, \_, \_, \_, y\_resampled, new\_sr in test\_files:

        sf.write(os.path.join(test\_dir, f"preprocessed\_{file}"), y\_resampled, new\_sr)

    # Save file names

    with open(os.path.join(text\_output\_dir, "train\_files.txt"), "w") as f:

        f.writelines("\n".join([f"{file},{emotion}" for file, emotion, \_, \_, \_, \_, \_ in train\_files]))

    with open(os.path.join(text\_output\_dir, "test\_files.txt"), "w") as f:

        f.writelines("\n".join([f"{file},{emotion}" for file, emotion, \_, \_, \_, \_, \_ in test\_files]))

    print("Preprocessing completed successfully.")

else:

    print("Error: No valid audio files processed. Check file naming conventions.")

# Visualization

if len(processed\_files) > 0:

    fig, axes = plt.subplots(min(3, len(train\_files)), 4, figsize=(18, 6 \* min(3, len(train\_files))))

    for i, (filename, emotion, original, denoised, filtered, resampled, \_) in enumerate(train\_files[:3]):

        axes[i, 0].plot(original, color='r')

        axes[i, 0].set\_title(f'Original - {filename} ({emotion})')

        axes[i, 1].plot(denoised, color='b')

        axes[i, 1].set\_title(f'Denoised - {filename} ({emotion})')

        axes[i, 2].plot(filtered, color='m')

        axes[i, 2].set\_title(f'Filtered - {filename} ({emotion})')

        axes[i, 3].plot(resampled, color='g')

        axes[i, 3].set\_title(f'Resampled - {filename} ({emotion})')

    plt.tight\_layout()

    plt.show()

# CHAPTER-5

**Experimentation and Result Analysis**

**5. Experimentation and Result Analysis**

Using the Speech emotion dataset, several machine learning models were trained during the experimentation phase, and their performance was assessed using a range of metrics. To determine how well each model predicted the emotion of the audio, we methodically evaluated its accuracy, precision, recall, and F1 score.

The findings showed that ensemble approaches performed better than more conventional models like support vector machines, Random forest and logistic regression. The model performed better because it was resilient against overfitting and could accommodate missing values.

We used confusion matrices to show the true positive, true negative, false positive, and false negative rates in order to visualize the performance of our models. This study shed light on the models' advantages and disadvantages by identifying instances of incorrect classification.

A **stacking ensemble model** is implemented using multiple base models and a meta-model to improve prediction performance. three models are used: Support Vector Machine (SVM), Random Forest (RF), Logistic Regression(LR).These models are selected from the dictionary best\_models, which presumably stores the best-performing versions of these algorithms.

### Key Points:

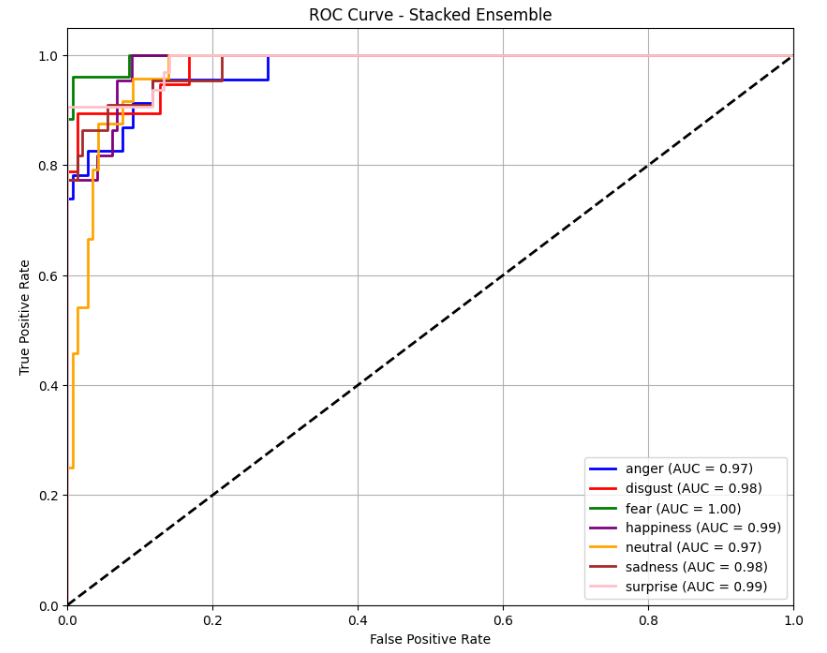
**Stacking Ensemble**: Combines multiple models to make better predictions.

**Base Models:** SVM, Random Forest.

**Meta-Model:** Logistic Regression used to combine base model predictions.

**Cross-Validation**: 5-fold cross-validation is used to assess the model’s robustness.

**Evaluation Metrics**: Accuracy, classification report, and confusion matrix provide a comprehensive evaluation of the model's performance.



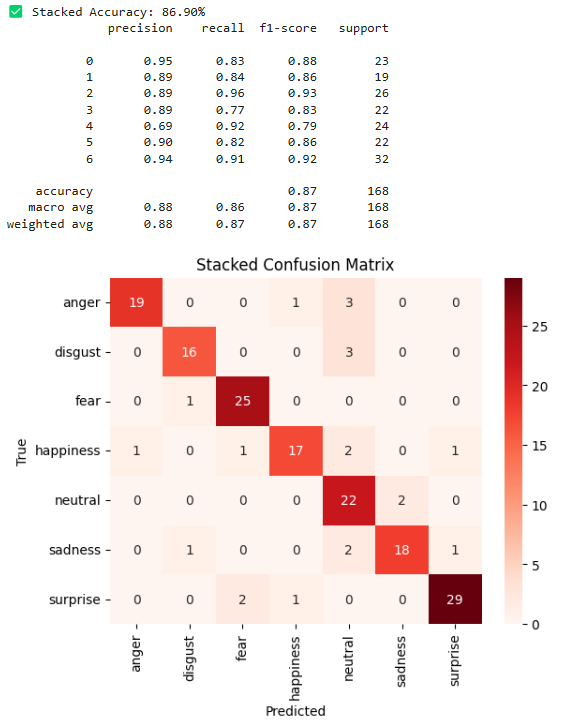


Figure 11. Confusion Matrix for Stacked Model

# CHAPTER-6

**CONCLUSION**

**6.Conclusion**

In this Speech Emotion Recognition (SER) project, the aim was to create a system able to precisely detect and classify emotions from speech data. It is a crucial task for use in areas like human-computer interaction, customer service, and mental illness monitoring, where emotional context comprehension may promote communication and decision-making. Throughout the project, machine learning models were used to identify patterns in the audio features and map them to different emotional states, leading to improved user experiences in systems that require emotional intelligence.

One of the key techniques employed in this project was the use of a stacking ensemble model, which combines multiple base models to improve prediction accuracy. The stacking model takes advantage of the individual strengths of different classifiers, such as Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression(LR). Each of these models interprets the data differently, capturing different features of the speech features that are responsible for emotion detection. For example, SVM is effective when dealing with high-dimensional data, Random Forest handles complex data relationships well without overfitting, and Logistic Regression performs linear model used for multi-class classification based on probability estimation. Mixing these heterogeneous models, stacking avoids the drawbacks of single models and decreases overall bias and variance to produce a better and more reliable system.

Beyond the base models, the meta-model—in this particular instance, Logistic Regression—is vital to the process of stacking. The meta-model learns to effectively choose how best to combine predictions from the base models, ultimately determining which models to believe in which situations. This extra layer of learning guarantees that the ensemble is not just an aggregation of models but a unified system that enhances performance by giving weight to the contribution of each base model depending on its capabilities. This meta-model layer hence introduces another layer of optimization, further boosting the overall predictive ability of the system.

The stacking model's flexibility is also well-suited to dynamic settings. With the introduction of new models or features, the stacking framework can readily add them to the current system without breaking the whole architecture. This flexibility is what ensures that the emotion recognition system continues to adapt with the data in pursuit of higher accuracy and reliability.

Overall, the application of a stacking ensemble model in this Speech Emotion Recognition task was critical to enhancing classification accuracy through the aggregation of multiple models with differing strengths. The capability of the meta-model to refine the predictions of base models increases system robustness and reliability and suits it for high-complexity tasks such as detecting emotion from speech. The results prove the strength of ensemble learning and provide a versatile tool for the development of emotion-aware systems with the ability to better sense and react to human emotions, setting the stage for more responsive and intuitive user interfaces in many types of applications.

### REFERENCES

[1]. Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub, et al.. Automatic Speech Emotion Recognition Using Machine Learning. Social Media and Machine Learning [Working Title], IntechOpen, 2019, ff10.5772/intechopen.84856ff. ffhal-02432557f. Submitted on 4 Jun 2022

[2]. Samaneh Madanian, Talen Chen a, Olayinka Adeleye,Department of Computer Science and Software Engineering, Auckland University of Technology (AUT), Auckland, New Zealand, John Michael Templeton b, b University of South Florida, Department of Computer Science and Engineering, Tampa, FL, USA

[3]. Yashpalsing Chavhan Student VIT, Pune India, M. L. Dhore Professor VIT, Pune India, Pallavi Yesaware Student VIT, Pune India(©2010 International Journal of Computer Applications (0975 - 8887))

[4]. Smith K. Khare a,∗ , Victoria Blanes-Vidal a , Esmaeil S. Nadimi a , U. Rajendra Acharya b a Applied AI and Data Science Unit, Mærsk Mc-Kinney Møller Institute, Faculty of Engineering, University of Southern Denmark, Denmark b School of Mathematics, Physics and Computing, University of Southern Queensland, Springfield, Australia{Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations}

[5]. RUHUL AMIN KHALIL1, 1Department of Electrical Engineering, Faculty of Electrical and Computer Engineering, University of Engineering and Technology, Peshawar, Pakistan .This article was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah. The authors therefore, acknowledge with thanks DSR for technical and financial support.

[6]. Duc Le and Emily Mower Provost University of Michigan Computer Science and Engineering, Ann Arbor, MI 48109 {ducle, [emilykmp}@umich.edu](mailto:emilykmp%7d@umich.edu)

[7]. Sundarprasad, Neethu, "SPEECH EMOTION DETECTION USING MACHINE LEARNING TECHNIQUES" (2018). Master's Projects. 628. DOI: https://doi.org/10.31979/etd.a5c2-v7e2 <https://scholarworks.sjsu.edu/etd_projects/628>

[8]. Ala Saleh Alluhaidan 1 , Oumaima Saidani, Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia(Appl. Sci. 2023, 13, 4750)

[9]. Ashwini S.Shinde Research Scholar, E&Tc Dept AISSMS Institute of Information Technology Asst.Prof. Pimpri Chichwad College of Engineering Pune, India ashwinik09@gmail.com(2022 6th International Conference On Computing, Communication, Control And Automation (ICCUBEA) Pimpri Chinchwad College of Engineering (PCCOE), Pune, India. Aug 26-27, 2022