# SPEECH EMOTION RECOGNITION SYSTEM USING MACHINE LEARNING

Nithyarubini D [1], Manikandan S [2], Sharmekaa S V [3], Nithin P [4]

*Bachelor of Engineering, Computer Science, Bannari Amman Institute of Technology, Erode, India*

*Bachelor of Engineering, Computer Science and Business System, Bannari Amman Institute of Technology, Erode, India*

*Bachelor of Engineering, Computer Science, Bannari Amman Institute of Technology, Erode, India*

*Bachelor of Engineering, Artificial Intelligence and Machine Learning, Bannari Amman Institute of Technology, Erode, India*

## ABSTRACT

*This project explores Speech Emotional Recognition through Machine Learning, a captivating blend of artificial intelligence and human emotion analysis. In an era where technology increasingly interfaces with human experience, the ability to understand and respond to emotional cues in speech holds transformative potential for communication and interaction. This abstract provides a comprehensive overview of the project's objectives, methodologies, and potential implications. Understanding emotions in spoken language is crucial for effective human communication. Speech Emotional Recognition (SER) using Machine Learning (ML) presents an innovative approach to automate this nuanced process. The project explores diverse ML techniques to enhance the accuracy of recognizing emotions in speech signals. The primary goal is to develop robust ML models capable of accurately identifying and classifying a broad spectrum of emotional states within spoken language. The focus is on creating a system that not only recognizes basic emotions but also captures subtle variations and complex emotional nuances. The project employs a multifaceted methodology, starting with diverse speech dataset acquisition and curation. These datasets cover emotional contexts, cultural influences, and linguistic variations to ensure adaptability. Feature extraction techniques transform raw speech signals into meaningful input for ML models. Various algorithms, including deep neural networks and support vector machines, are explored for emotional classification. Special attention is given to address challenges like data imbalance and overfitting. Successful implementation holds implications across domains. In human-computer interaction, it can enhance user experience by enabling devices to respond empathetically. In healthcare, it could aid in early detection of emotional distress by analyzing speech patterns. In education, the technology might contribute to personalized learning experiences based on students' emotional engagement. Ongoing concerns include cultural variability, ethical considerations, and the need for continuous model adaptation. Future directions involve refining models with larger and more diverse datasets, exploring real-time applications, and addressing interpretability challenges associated with complex ML models. Speech Emotional Recognition using Machine Learning represents a promising frontier in artificial intelligence. The project aims to advance understanding of emotional cues in speech, paving the way for practical applications that could redefine human-machine interactions. As technology evolves, imbuing machines with emotional intelligence opens new possibilities for a more responsive and empathetic technological landscape.*

**Keywords***: Speech Emotional Recognition (SER), Machine Learning (ML), Emotional Cues, Human-Computer Interaction (HCI), Emotional Intelligence, Machine Learning models*

## 1. INTRODUCTION

In the digital society we live in today, better communication requires an understanding of human emotions. Our project's main goal is to use machine learning to create a spoken emotion recognition system. With the

use of this cutting-edge technology, spoken words can be used to evaluate and understand emotions. Through the use of cutting-edge machine learning algorithms, our goal is to develop a system that can recognize a variety of emotions, including happiness, sadness, and rage, from audio cues. Artificial intelligence is being used to investigate applications in a variety of industries, including customer service and mental health, in addition to recognizing emotions in speech. Come along on this adventure with us as we construct an advanced emotion identification system with the goal of improving human-machine interactions.

### 1.1 Background of the project:

The need to comprehend and respond to human emotions is becoming increasingly important in the technologically driven world of today. Our venture spins around the production of a Discourse Feeling Acknowledgment Framework through the utilization of AI. The potential for decoding and comprehending the feelings expressed in spoken language exists with this technology. We want to train the system to accurately identify a range of emotions, including joy, sadness, and anger, by analyzing distinct vocal patterns. To accomplish this, we will use cutting-edge machine learning algorithms.

The vast potential of artificial intelligence fuels this endeavor, with the ultimate objective of developing a tool that not only interprets speech emotions but also has practical applications in a variety of fields. Our project aims to harness the transformative power of emotion-aware technology for the benefit of individuals and industries alike by improving mental health diagnostics, improving interactions with customer service, and improving human-computer engagement. The combination of state of the art innovation and human inclination opens ways to another outskirts in correspondence and connection.

### 1.2 Motivation (Scope of the Proposed Initiative)

The transformative potential of the Speech Emotion Recognition System made possible by machine learning is the driving force behind our initiative. Understanding emotions is essential for effective communication in today's dynamic world, which requires more than just words. The goal of our project is to develop a tool that can not only decipher the intricate emotions that are expressed through speech but also provide practical applications in a variety of fields. We want to create a system that can recognize a wide range of emotions, from joy to anger, using advanced machine learning, which will improve user experiences and interactions between computers and humans.

The extent of this drive reaches out to fields like psychological wellness, where feeling mindful innovation can assume a vital part in diagnostics. In addition, it may aid in more individualized and sympathetic interactions in customer service. Our goal is to contribute to a world in which machines can intelligently and empathetically comprehend and respond to human emotions. The significant impact that this technology has had on communication exemplifies the motivation behind our project.

## 2. LITERATURE SURVEY

The literature review conducted for the project provides a comprehensive overview of recent advancements in speech emotion recognition using machine learning techniques. This review critically assesses the current state of research, identifies areas with limitations, and proposes potential solutions. Here, we'll delve deeper into the reviewed studies and expand on the central issues and challenges highlighted in the literature.

This comprehensive review by Kott ursamy and Kott ilingam (2021) emphasizes eXnet library's significance in improving facial expression recognition accuracy. However, problems like overfitting with large models and limitations in memory and computation persist. The study proposes eXnet, a novel Convolutional Neural Network (CNN) that uses parallel feature extraction to improve accuracy while significantly reducing the number of parameters, as a solution to overfitting. The most recent version of eXnet maintains a balance between effectiveness and model size while demonstrating significant accuracy advancements over its predecessor. The generalized eXnet addresses overfitting by incorporating data augmentation techniques. As a result, it offers a comprehensive strategy for advancing facial expression recognition in deep learning for social science and human-computer interaction.

A. Thakur, P. Budhathoki et al., (2020) focuses on developing a user-friendly and accurate Sign Language Recognition System to address the communication difficulties faced by the deaf and mute community. The system, which makes use of neural networks and image processing, aims to provide a cost-effective way to make it easier for people with speech impairments and the rest of society to communicate. The goal is to develop a system that can use input gestures to generate speech and text without the need for a translator and encourage two-way communication. This research contributes to the larger objective of enhancing

inclusivity for the deaf-mute population worldwide and is a significant step toward the commercialization of a reliable and accessible sign language recognition system.

This review by J. Kaur and A. Kumar (2021) uses various machine learning algorithms, including k-nearest neighbor, multi-layer perceptron (MLP), convolutional neural network (CNN), and random forest, to investigate emotion recognition from speech. The classifiers distinguished seven emotions with admirable precision using features from the Berlin database, such as short-term Fourier transform spectrograms and mel frequency cepstral coefficients. Eminently, the MLP classifier arose as the best, bragging an amazing generally speaking exactness 90.36%. This study sheds light on the potential of machine learning to accurately decipher emotional states from speech signals and provides valuable insights into the comparative performance of various classification algorithms for emotion recognition.

K. W. Gamage, V. Sethu et al., utilizing i-vectors that encapsulate the distribution of frame-level MFCC features, this review introduces a back-end for utterance-level emotion classification using Gaussian Probabilistic Linear Discriminant Analysis (GPLDA). Through investigates the IEMOCAP corpus, our proposed GPLDA back-end shows better execution looked at than a SVM-based partner. Importantly, the GPLDA model has lower sensitivity to i-vector dimensionality, facilitating parameter tuning during system development and increasing robustness. The framework's ability to improve the accuracy and adaptability of emotion classification systems was demonstrated in this study, which points to a promising future for speech-based emotion recognition technologies.

This review by J. Han, Z. Zhang et al., uses a novel reconstruction-error-based (RE-based) learning framework for improving speech-based continuous emotion recognition. The framework makes use of memory-enhanced recurrent neural networks (RNN) by employing two RNN models in succession. The first model is an autoencoder for reconstructing the original features, and the second model uses the complementary descriptor of reconstruction error (RE) to predict emotions. Improved Concordance Correlation Coefficients (.729 vs..710 for arousal,.360 vs..237 for valence) show that experimental results on the RECOLA database significantly outperform baseline systems. Outstandingly, this proposed structure outperforms other best in class strategies, highlighting its true capacity in refining the exactness and viability of consistent feeling acknowledgment frameworks.

Albornoz et al., investigate a new spectral feature in order to determine emotions and
to characterize groups. In this study based on acoustic features and a novel hierarchical classifier, emotions are grouped. Different classifier such as HMM, GMM and MLP have been evaluated with distinct configuration and input features to design a novel hierarchical techniques for classification of emotions. The innovation of the proposed method is two things, first the election of foremost performing features and second is employing of foremost class-wise classification performance of total features same as the classifier. Experimental result in Berlin dataset demonstrates the hierarchical approach achieves the better performance compare to best standard classifier, with decuple cross-validation. For example, performance of standard HMM method reached 68.57% and the hierarchical model reached 71.75%

## 3. OBJECTIVE AND METHODOLOGY

This undertaking means to involve AI to comprehend and decipher feelings in communicated in language. By breaking down the manner in which individuals talk, like their pitch, tone, and speed, we need to foster calculations that can precisely distinguish feelings like annoyance, disdain, dread, satisfaction, lack of bias, trouble, and shock progressively. Our primary goals incorporate tracking down the main elements for feeling acknowledgment to work on the presentation of our models and make the cycle more proficient. We likewise need to foster a model that can deal with different sorts of feelings, process sound continuously, and adjust to various speakers. In less difficult terms, we are attempting to make a framework that can comprehend and decipher feelings from how individuals talk. This includes investigating the manner in which individuals talk and creating PC programs that can perceive various feelings, like annoyance, bliss, or bitterness, as individuals are talking. We will likely make the framework work rapidly and precisely, in any event, while managing various speakers and complex discourse designs.

### 3.1 Objectives of the Proposed Work

1. **Effective Feature Selection:**
   Executing procedures to recognize the main highlights for feeling acknowledgment from discourse, like pitch, tone, and speed, to work on the precision and productivity of the model.

2. **Resistance Emotion Classification Model:**
   Fostering a model that can precisely characterize feelings like resentment, disdain, dread, joy, nonpartisanship, misery, and shock, even within the sight of foundation commotion or different unsettling influences.
3. **Enhanced Audio Processing:**
   Working on the handling of sound information to separate pertinent elements that can be utilized to recognize and group feelings continuously.
4. **Feature Extraction:**
   Creating strategies to separate acoustic highlights from discourse information that are generally pertinent for recognizing various feelings, and incorporating these elements into the feeling acknowledgment model.
5. **Real-time Processing:**
   Making calculations and frameworks that can examine and characterize feelings in communicated in language continuously, considering prompt criticism or reaction.
6. **Adaptability to Speaker Variability:**
   Guaranteeing that the feeling acknowledgment model can adjust to varieties in discourse examples and attributes across various speakers, dialects, and vernaculars.
7. **Handling Complex Datasets:**
   Creating methods to deal with the intricacy of discourse information, remembering varieties for accents, pitches, and profound articulations, to work on the strength and precision of the feeling acknowledgment model.

### 3.2 Proposed Methodology

This task is tied in with helping PCs to comprehend and answer human feelings in a more modern manner. The group has made a framework that utilizations trend setting innovation called fake brain organizations to perceive seven distinct feelings from individuals' voices. The framework has been tried broadly and has demonstrated to be extremely exact in understanding and ordering feelings from voice accounts. To prepare the framework, it needs a ton of information that incorporates insights regarding the loads utilized in the preparation cycle and marks for various feelings. At the point when th e framework is given a sound record, it is changed in accordance with process the sound reliably. This assists the framework with figuring out how to perceive various feelings. During testing, the framework cautiously takes a gander at the contribute and energy the voice information and utilizations what it has figured out how to precisely distinguish the feelings being communicated. The fundamental objective of this venture is to make a framework that can truly comprehend and answer human feelings when individuals talk. This could make it a lot simpler and more normal for individuals to collaborate with PCs, prompting better and more instinctive encounters while utilizing innovation.

### 3.2 Block Diagram

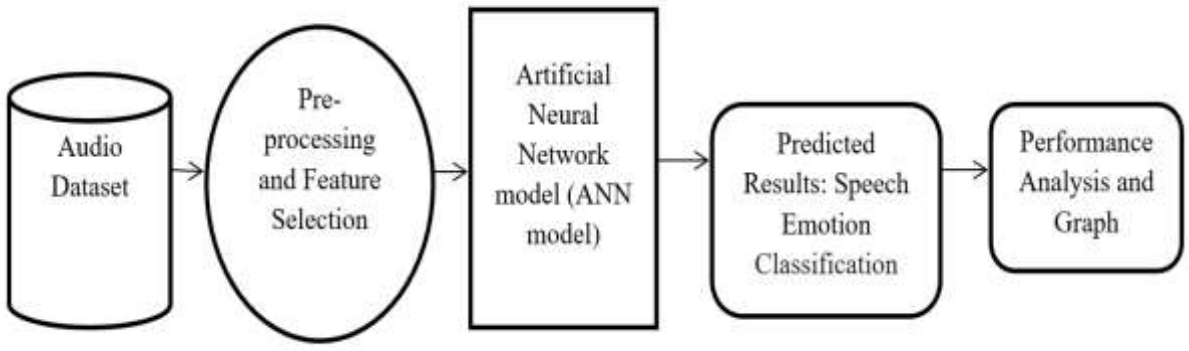


Figure 1.Process flow chart

### 3.3 Flask Frameworks

Flask is integral to the disease prediction project for several key reasons:

1. **Web Development:** Flask simplifies web application creation, forming the basis for the user interface (UI) and user interactions.

2. **User-Friendly Interface:** It aids in designing a user-friendly interface, making web pages, forms, and navigation intuitive for users.
3. **Dynamic Web Pages:** Flask supports dynamic web pages, enabling real-time data presentation, user input, and immediate prediction and metric display.
4. **Routing:** Flask's routing mechanism defines URLs and associates them with Python functions, creating distinct sections for login, data upload, prediction, and analysis.
5. **Python Integration:** Its alignment with Python allows seamless data transfer between the frontend and backend components.
6. **Efficient Data Handling:** Flask streamlines data handling, crucial for tasks like dataset uploads, symptom data transmission, and metric display.
7. **Security:** Flask offers robust security features, essential for safeguarding healthcare data and user privacy.
8. **Scalability:** It suits both small-scale and large-scale applications, accommodating project growth.
9. **Community and Ecosystem:** Flask has an active developer community and a rich ecosystem, providing pre-built components for enhanced functionality.
10. **Deployment Flexibility:** It can be deployed on various platforms, meeting specific requirements.
11. **Rapid Prototyping:** Flask's simplicity supports swift prototyping and iterative development.
12. **Customization:** Developers can customize the application's appearance and behaviour to match project requirements.

In summary, Flask is essential for UI development, data exchange, security, and user interaction in the disease prediction project. Its adaptability, Python integration, and community support make it invaluable for creating a functional and user-centric web application.

## 4. PROPOSED WORK MODULES

### 4.1 MODULES DESCSRIPTION:

#### 1. Dataset:

In the principal module, we fostered the framework to get the information dataset for the preparation and testing reason. Dataset is given in the model envelope. The dataset comprises of 2,800 Discourse Feeling sound dataset. The dataset comprise of classes like: irate, disdain, Dread, blissful, nonpartisan, Miserable and shock. The dataset is alluded from the kaggle site.

#### 2. Importing the necessary libraries:

In the subsequent module, we import the essential libraries for our discourse feeling discovery framework. The vital and extraordinary library that upholds sound and music investigation is Librosa. Just utilize the Pip order to introduce the library. It gives building blocks that are expected to develop a data recovery model from music. Another extraordinary library we will utilize is for profound getting the hang of displaying objects is TensorFlow, and I trust everybody has proactively introduced TensorFlow.

#### 3. Exploratory Data Analysis of Audio data

• We have various envelopes under the dataset organizer. Prior to applying any preprocessing, we will attempt to comprehend how to stack sound documents and how to envision them in type of the waveform. To stack the sound document and stand by listening to it, then, at that point, you can utilize the IPython library and straightforwardly give it a sound record way. We have taken the primary sound document in the crease 1 envelope.
• Presently we will utilize Librosa to stack sound information. So when we load any sound record with Librosa, it gives us 2 things. One is test rate, and the other is a two-layered cluster. Allow us to stack the above sound document with Librosa and plot the waveform utilizing Librosa.
• Test rate - It addresses the number of tests that are recorded each second. The default testing rate with which librosa peruses the document is 2,800. The example rate contrasts by the library you pick.
• 2-D Exhibit - The principal pivot addresses recorded examples of adequacy. What's more, the subsequent pivot addresses the quantity of channels. There are various kinds of channels - Monophonic (sound that has one channel) and sound system (sound that has two channels).
• We load the information with librosa, then, at that point, it standardizes the whole information and attempts to give it in a solitary example rate. Similar we can accomplish utilizing scipy python library too. It will likewise give us two snippets of data - one is test rate, and the other is information. At the point when you print the example rate utilizing scipy-it is not the same as librosa. Presently let us imagine the wave sound

information. Something significant to comprehend between both is-the point at which we print the information recovered from librosa, it tends to be standardized, however when we attempt to peruse a sound document utilizing scipy, it can't be standardized. Librosa is presently getting famous for sound sign handling in light of the accompanying three reasons.

**1.** It attempts to unite the sign into mono (one channel).

**2.** It can address the sound sign between - 1 to +1 (in standardized structure), so a customary example is noticed.

**3.** It is likewise ready to see the example rate, and as a matter of course, it changes it over completely to 22 kHz, while on account of different libraries, we see it as indicated by an alternate worth.

**4. Imbalance Dataset check:**

Presently we realize about the sound records and how to picture them in sound arrangement. Moving arrangement to information investigation we will stack the CSV information document accommodated every sound record and check the number of records we that have for each class. The information we have is a filename and where it is available so allowed us to investigate first record, so it is available in crease 7 with class. Presently utilize the worth counts capability to really look at records of each class. At the point when you see the result so information isn't imbalanced, and the vast majority of the classes have a roughly equivalent number of records

**5. Data Preprocessing:**

- A few sounds are getting recorded at an alternate rate-like 44KHz or 22KHz. Utilizing librosa, it will be at 22KHz, and afterward, we can see the information in a standardized example. Presently, our undertaking is to remove some significant data, and keep our information as independent(Extracted highlights from the sound sign) and ward features(class names). We will utilize Mel Recurrence Cepstral coefficients to extricate autonomous elements from sound signs.
- MFCCs - The MFCC sums up the recurrence appropriation across the window size. In this way, breaking down both the recurrence and time attributes of the sound is conceivable. This sound portrayal will permit us to distinguish highlights for arrangement. Thus, it will attempt to change over sound into some sort of highlights in view of time and recurrence attributes that will assist us with doing characterization. To discover and peruse more about MFCC, you can watch this video and can likewise peruse this exploration paper by springer. To show how we apply MFCC practically speaking, first, we will apply it on a solitary sound record that we are as of now utilizing.
- Presently, we need to extricate highlights from all the sound documents and set up the dataframe. In this way, we will make a capability that takes the filename (document way where it is available). It stacks the record utilizing librosa, where we get 2 data. To start with, we'll track down MFCC for the sound information, and to figure out scaled highlights, we'll track down the mean of the render of an exhibit.
- Presently, to remove every one of the elements for every sound record, we need to utilize a circle over each column in the dataframe. We additionally utilize the TQDM python library to follow the advancement. Inside the circle, we'll set up a modified record way for each document and call the capability to separate MFCC includes and add includes and comparing marks in a recently framed dataframe.

**6. Splitting the dataset:**

Split the dataset into train and test. 80% train data and 20% test data.

**7. Audio Classification Model Creation:**

We have removed highlights from the sound example and splitter in the train and test set. Presently we will carry out an ANN model utilizing Keras consecutive Programming interface. The quantity of classes is 7, which is our result shape (number of classes), and we will make ANN with 3 thick layers and design is made sense of beneath.

1. The principal layer has 100 neurons. Input shape is 40 as indicated by the quantity of highlights with enactment capability as Relu, and to keep away from any overfitting, we'll utilize the Dropout layer at a pace of 0.5.
2. The subsequent layer has 200 neurons with enactment capability as Relu and the drop out at a pace of 0.5.
3. The third layer again has 100 neurons with initiation as Relu and the drop out at a pace of 0.5.

**8. Compile the Model**

To accumulate the model we want to characterize misfortune capability which is downright cross-entropy, precision measurements which is exactness score, and a streamlining agent which is Adam.

**9. Train the Model**

We will prepare the model and save the model in HDF5 design. We will prepare a model for 100 ages and group size as 32. We'll utilize callback, which is a designated spot to realize what amount of time it required to prepare over information.

**10. Check the Test Accuracy**

Presently we will assess the model on test information. We became close to around 100% precision on the preparation dataset and 99 percent on test information.

**11. Saving the Trained Model:**

When we are sufficiently sure to take our prepared and tried model into the creation prepared climate, the initial step is to save it into a .h5 or .pkl document utilizing a library like pickle . Then, we import the module and dump the model into .h5 record.

## 5. RESULTS AND DISCUSSION

The venture zeroed in on saddling the abilities of fake brain organizations (ANNs) for discourse feeling acknowledgment (SER). The outcomes showed that ANNs displayed effective equal handling, empowering them to deal with numerous undertakings all the while, which is critical for ongoing feeling acknowledgment from verbally expressed words. Notwithstanding, the concentrate likewise uncovered the vulnerability of ANNs to obstruction, where the deficiency of brain cells can influence their exhibition. This finding accentuates the significance of strength and flexibility in the plan and execution of SER frameworks in light of ANNs.

Moreover, the venture featured the capacity of ANNs to hold data inside the organization, guaranteeing that they can in any case produce results even without a trace of explicit information inputs. This trademark is especially important for constant uses of SER, as it considers persistent acknowledgment and reaction to profound signs, even in testing or dynamic conditions.

The study also noted the gradual degradation of ANNs, indicating that they exhibit a progressive decline in performance over time rather than an abrupt cessation. Understanding this behavior is essential for the long-term reliability and maintenance of SER systems, as it prompts the need for ongoing monitoring, adaptation, and potential retraining of the networks to sustain their effectiveness. The proposed system achieves the target train accuracy of 100% and the target test accuracy of 99%.

Moreover, the project demonstrated the potential for training ANNs to learn from past events and make decisions, highlighting their adaptability and learning capabilities. This aspect is critical for SER, as it enables the networks to continuously improve and refine their ability to recognize and respond to diverse emotional expressions in spoken language.

Overall, the results and discussions from this project underscore the potential of ANNs for SER, while also highlighting the importance of addressing challenges such as resistance and gradual degradation to ensure the robustness and reliability of emotion recognition systems based on artificial neural networks. These insights provide valuable guidance for the continued development and enhancement of SER technology.

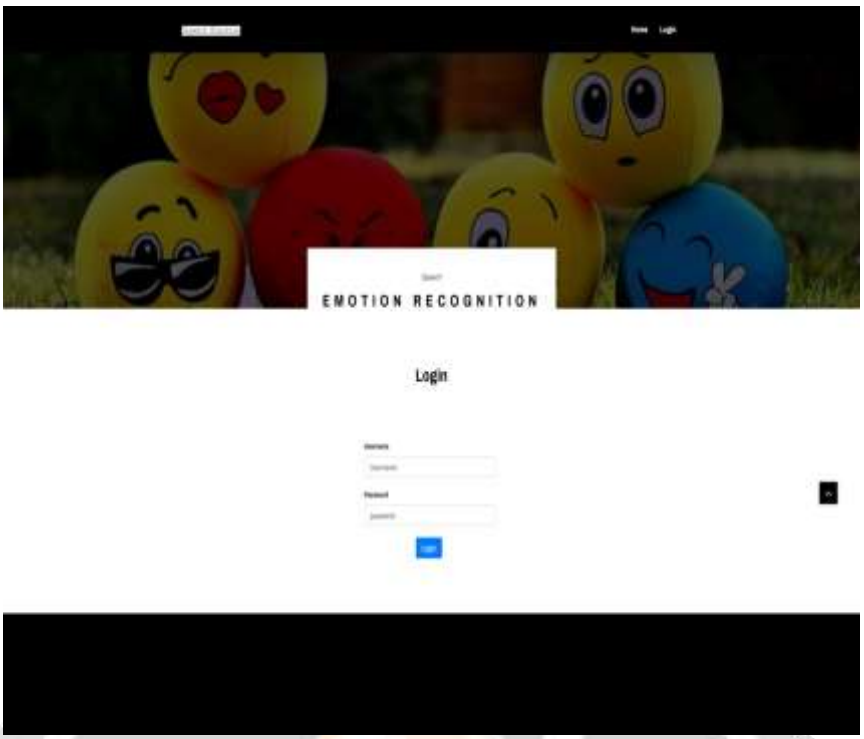


Figure 2. Login page

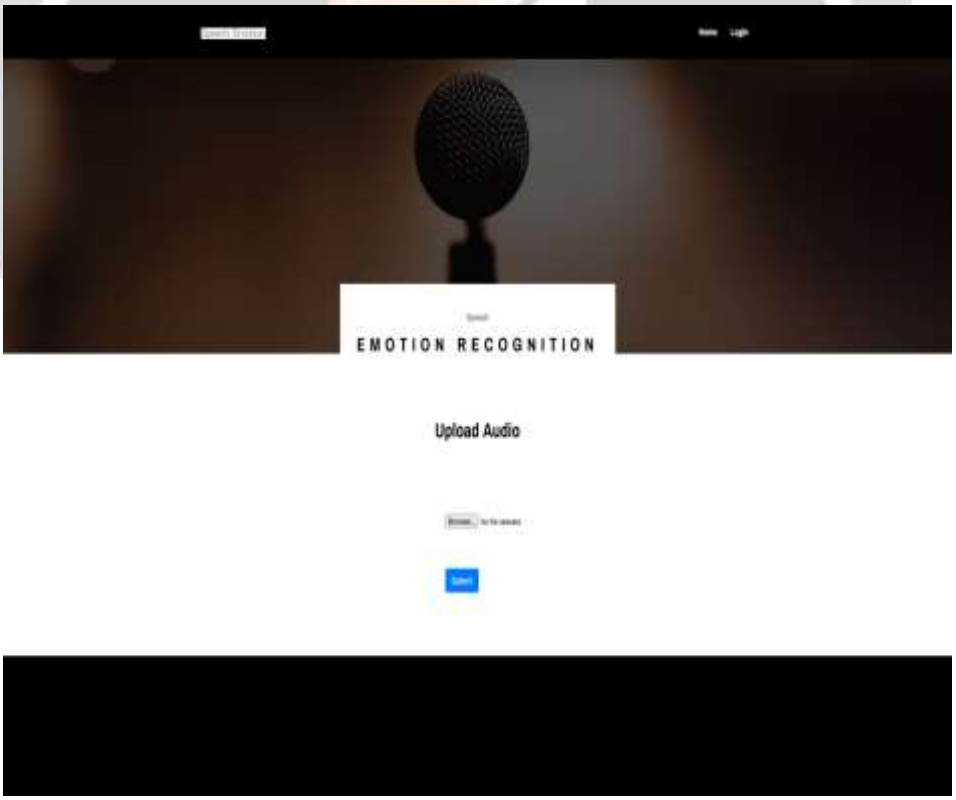


Figure 3. Audio Uploading page

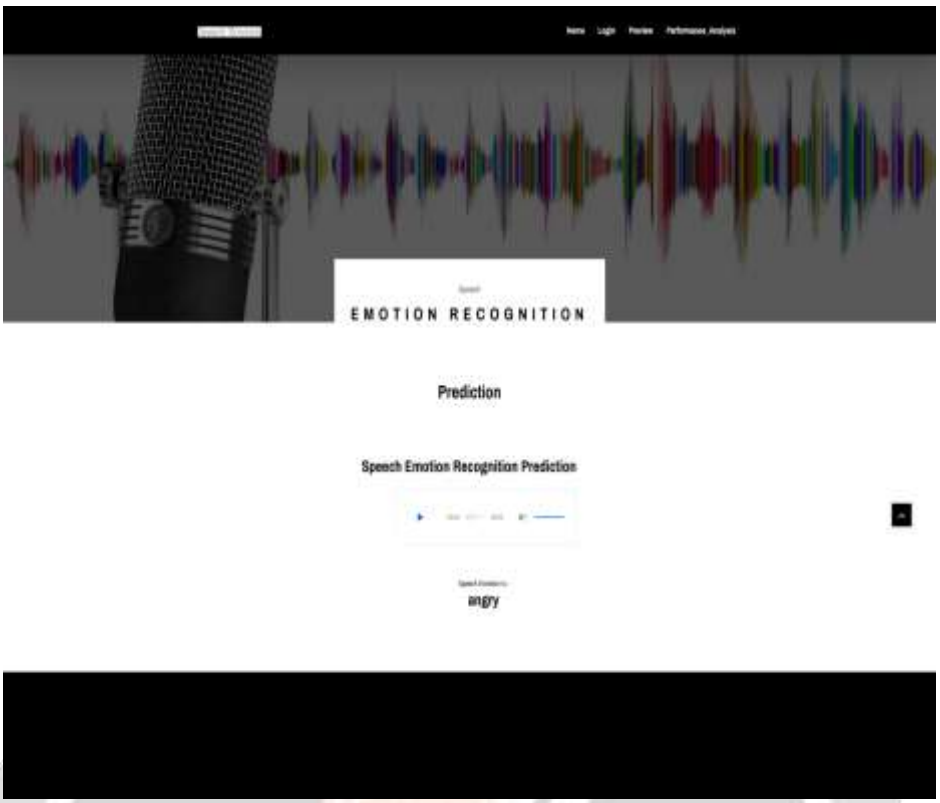


Figure 4. Prediction page

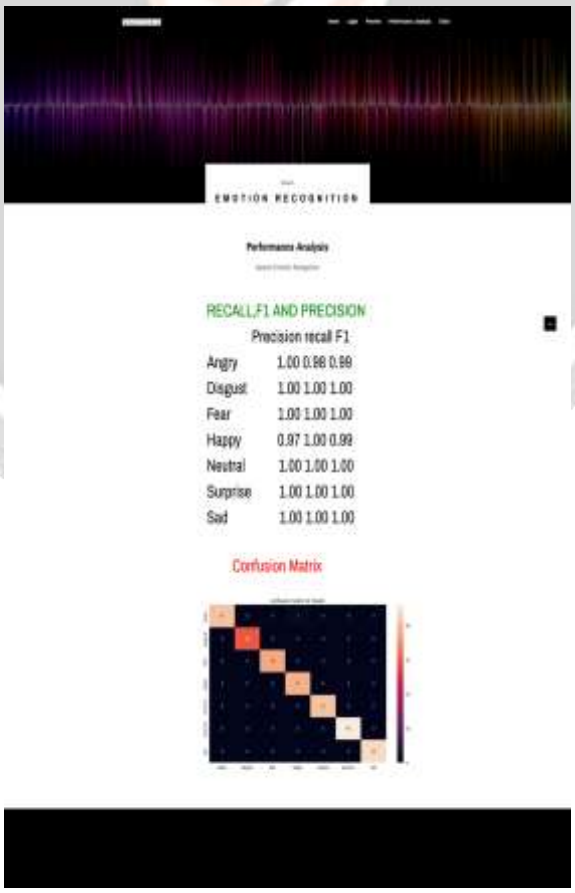


Figure 5. Accuracy Score page

## 6. CONCLUSION

Speech Emotion Recognition (SER) is a compelling area of study within software engineering research. The proposed framework aligns with the current state of SER calculation. Furthermore, there is potential to expand the framework to support multilingual emotion recognition. Additionally, there is an opportunity to further refine emotion characterization at more granular levels and in diverse contexts.

The potential for extending the framework to encompass multilingual emotion recognition is an exciting prospect, as it could enable the system to understand and interpret emotions expressed in different languages, thereby enhancing its versatility and applicability across diverse linguistic contexts. Moreover, the refinement of emotion characterization at minute levels and in various design scenarios holds promise for capturing and understanding nuanced emotional expressions, leading to more comprehensive and accurate emotion recognition and response within the framework.

Overall, the proposed framework offers a solid foundation for SER and presents promising opportunities for future advancements, including multilingual emotion recognition and enhanced characterization of emotions at finer levels and in diverse contexts.

## 7. REFERENCES

**[1]** Kottursamy, Kottilingam. "A review on finding efficient approach to detect customer emotion analysis using deep learning analysis." Journal of Trends in Computer Science and Smart Technology 3, no. 2 (2021): 95-113.

**[2]** Thakur, Amrita, Pujan Budhathoki, Sarmila Upret i, Shirish Shrestha, and Subarna Shakya. "Real Time Sign Language Recognition and Speech Generation." Journal of Innovative Image Processing 2, no. 2 (2020): 65-76.

**[3]** Kaur, Jasmeet , and Anil Kumar. "Speech Emotion Recognition Using CNN, k-NN, MLP and Random Forest." In Computer Networks and Inventive Communication Technologies, pp. 499-509. Springer, Singapore, 2021.

**[4]** Gamage, Kalani Wataraka, Vidhyasaharan Sethu, Phu Ngoc Le, and Eliathamby Ambikairajah. "Ani-vector gplda system for speech based emotion recognition" In 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp. 289-292. IEEE, 2015.

**[5]** Han, Jing, Zixing Zhang, Fabien Ringeval, and Björn Schuller. "Reconstruction-error-based learning for continuous emotion recognition in speech."In 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 2367-2371. IEEE, 2017.

**[6]** Akrami, N., F. Noroozi, and G. Anbarjafari. "Speech based emotion recognition and next reaction prediction." In 25th Signal Processing and Communications Applications Conference, Antalya, pp. 1-6. 2017.

**[7]** Rieger, S. A., Muraleedharan, R., and Ramachandran, R. P. (2014, September). Speech based emotion recognition using spectral feature extraction and an ensemble of kNN classifiers. In The 9th International Symposium on Chinese Spoken Language Processing (pp. 589-593). IEEE.

**[8]** Tabatabaei, Talieh S., and Sridhar Krishnan. "Towards robust speech based emotion recognition."2010 IEEE International Conference on Systems, Man and Cybernetics. IEEE, 2010.

**[9]** Z. Li, "A study on emotional feature analysis and recognition in speech signal," Journal of China Institute of Communications, vol. 21, no. 10, pp. 18–24, 2000.

**[10]** T. L. Nwe, S. W. Foo, and L. C. de Silva, "Speech emotion recognition using hidden Markov models," Speech Communication, vol. 41, no. 4, pp. 603–623, 2003.

**[11]** Vaijayanthi, S., and J. Arunnehru. " Synthesis Approach for Emotion Recognition from Cepstral and Pitch Coefficients Using Machine Learning." In International Conference on Communication, Computing and Electronics Systems, p. 515.

**[12]** Y. Kim, H. Lee, and E. M. Provost , "Deep learning for robust feature generation in audio-visual emotion recognition," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '13), Vancouver, Canada, 2013.

**[13]** Livingstone, Steven R., and Frank A. Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English." PloS one 13, no. 5 (2018): e0196391

**[14]** Chourasia, Mayank, Shriya Haral, Srushti Bhatkar, and Smita Kulkarni. "Emotion recognition from speech signal using deep learning." Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2020 (2021): 471-481.

**[15]** R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review", IEEE Access, vol. 2, no. 7, pp. 117327-117345, 2019.