



VocaLyze – AI Based Emotion Recognition from Voice

**Sadigale Shravani Santosh¹, Mali Vaishnavi Sukhdev², Gotavle Sushma Sanjay³,
Linge Sandeep Ganpat⁴**

UG Student, Department of Computer Science and Engineering, SKN Sinhgad College of Engineering, Pandharpur,
India¹

UG Student, Department of Computer Science and Engineering, SKN Sinhgad College of Engineering, Pandharpur,
India²

UG Student, Department of Computer Science and Engineering, SKN Sinhgad College of Engineering, Pandharpur,
India³

Professor, Department of Computer Science and Engineering, SKN Sinhgad College of Engineering, Pandharpur,
India⁴

Abstract: When a Person emotion change Changes can be seen in their voice tone, pitch speaking speed and intensity. The name of our project is VocaLyze. It is an AI-based web application that identifies a person emotion from their voice. In this system the user can provide their voice in two ways first by recording their voice using a microphone and second by uploading an existing audio file. After receiving the voice, the system first removes background noise and unwanted silent parts from the audio. Then the system extracts voice features such as MFCC, pitch, energy and spectral features. These features are then given to machine Learning and Deep Learning models such as Random Forest, SVM, CNN and LSTM. These models help the system to identify and classify the emotion present in the user voice. The system also provides a confidence score for the detected emotion. Our project does not only identify the emotion. It also provides recommendations according to the detected emotion. These recommendations can include quotes, calming music, breathing, meditation and wellness activities.

Keywords: Artificial Intelligence, Machine Learning, Voice Emotion Recognition, MFCC, CNN, LSTM, Emotion Classification, Personalized Recommendation.

I. INTRODUCTION

Human feelings heavily shape how we interact every day. Beyond the words spoken the way someone talks their pitch, volume, space and vocal energy offers clear indicators of how they feel. Core emotional states like joy, sorrow, frustration, fear, shock or calm all leave fingerprints on a person's voice. Being able to read these cues automatically opens up incredible possibilities for mental health tracking, responsive software and interactive tools. That said, deciphering feelings from sound is tricky mostly because everyone has a unique voice and a personal way of speaking. Earlier methods relied heavily on feature extraction and older basic statistical tools to guess mood. Thanks to artificial intelligence, machine learning and deep learning computers can now automatically analyze complex speech signals to catch nuanced emotional patterns. Key vocal traits such as Mel-Frequency Cepstral coefficients (MFCCs) fundamental frequency, energy levels and spectral characteristics when fed into models like Support Vector Machines (SVM) Random Forests, Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks make accurate mood detection possible.

To solve these challenges this project introduces VocaLyze, a smart platform designed for voice-driven mood analysis. The software lets users either speak directly into a microphone or upload a -recorded audio file. Once received the system cleans up the sound by stripping background noise trimming pauses applying digital filters and normalizing the volume. After cleaning it extracts acoustic markers and feeds them to the classification engine, which predicts the user's emotional state along with a percentage-based confidence rating. To protect data VocaLyze features secure account logins and logs past results saving timestamps, emotions and certainty levels for future viewing.



Beyond identifying mood VocaLyze acts as a supportive companion by offering tailored suggestions based on the output. Depending on how the user feels the platform might recommend calming music, guided meditations, uplifting quotes, breathing drills, mindfulness habits or everyday wellness tips. Everything comes together in a dashboard where users can view their current mood, reliability score, custom advice and historical trends at a glance. By combining analysis secure data storage, mood tracking and personalized wellness suggestions VocaLyze delivers an accessible intuitive hub for understanding emotional well-being through speech.

II. RELATED WORK

The name of the proposed system VocaLyze derives from these four research vectors; (i) "Augmented vision through Natural Language Processing/NLP and Deep Machine Learning (DeepML) (including Text Mining) (ii) "Intelligent Prediction (iii) "Voice-Based Emotion Detection and Recognition (VCER) and (iv) "AI-based Digital Healthcare and personalized support (Lifelifeltd).

A. Artificial Intelligence and Natural Language Processing

BERT (introduced by Devlin et al. in 2018) is a model based on the Transformer that has the capacity of capturing the contextual teacher' of the words in both directions (2018). It was designed to tackle the question answering, sentiment analysis, and text classification tasks [1]. Despite their logical advantages, BERT-based techniques require huge data and page memory.

B. Deep Learning and Intelligent Prediction

Liu and Lang (2022) have proposed a deep learning intrusion detection system to detect operations in computer networks [2]. They considered different models CNN RNN, and LSTM (after the data was pre-processed). It is observed that deep learning method could be useful for detecting complicated patterns and aid in real-time forecast; though, the data used for training, and the computational requirement as well as the tuning of models are important for implementation.

Pallewela and Alahakoon (2024) Suggested a blockchain security system for the internet of things device [3]. It focused on the IoT secure data and authentication. Yet, it has some challenges like scalability, processing delay, resource consumption.

C. Voice-Based Emotion Recognition

Jain and Shrimali (2024) were designed and tested the computer voice-based study in combination with artificial knowledge-based system [4]. The calculated phoneme characteristics including the number of common mode components and the frequency of pitch served as raw data for the research. The systems also used LSTM and hybrids of CNN-LSTM for the determination of the presence or absence of emotion that were assessed on the test speech data sets, including TESS and RAVDESS. The authors found that in different studies in the literature, AI can be used for emotion recognition for human computer interaction, mind reading, healthcare, and virtual assistants. But, the training data was shown to be a big part to the model's success and that real human voices and a wide source of languages still pose challenges.

D. AI-Based Healthcare

Mohammed Mohsin et al. (2024) designed the AI based health prediction model by implementing various machine learning algorithms [5]. They used different algorithms including SVM, Random Forest, and ANN for prediction on health factors. It was found that the machine learning can assist in making health predictions, this way adding an extra layer of support to decision making. But, the performance of such systems relies heavily on the training data.

E. Research Gap

Based on previous researches, we have identified that Artificial Intelligence and Deep Learning approach is better when analyzing the speech, emotions, and prediction for the health applications. In the market, no such integrated solution is available, which the system will do all these activities like voice input, audio preprocessing, feature extraction prediction confidence scoring emotion history and personalized recommendations.



TABLE I
MAPPING OF LITERATURE TO THE PROPOSED SYSTEM

Reference	Focus	Contribution to Proposed System
[1]	BERT and Natural Language Processing using Transformer-based models	Provides a foundation for using AI-based models to identify complex patterns and supports the use of intelligent classification techniques.
[2]	Deep Learning-based intrusion detection using CNN, RNN and LSTM	Supports the use of Deep Learning models such as CNN and LSTM for identifying complex patterns from input data.
[3]	Blockchain-based security framework for IoT systems	Highlights the importance of secure authentication and data management, supporting secure user authentication and data protection in VocaLyze.
[4]	Voice-based emotion recognition using MFCC, pitch, LSTM and CNN-LSTM	Directly supports the proposed voice emotion recognition approach, including speech feature extraction and Deep Learning-based emotion classification.
[5]	AI-based healthcare prediction using SVM, Random Forest and ANN	Supports the use of Machine Learning algorithms for prediction and classification and highlights the importance of quality training data.
[1-5]	AI, Deep Learning, voice emotion recognition, security and intelligent prediction	Identifies the need for an integrated system that combines voice input, audio preprocessing, feature extraction, emotion prediction, confidence scoring, emotion history and personalized recommendations.

III. PROBLEM STATEMENT

Problem Statement: “To develop an AI-based web application that can detect and classify emotions, from voice input using audio feature extraction and Machine Learning/Deep Learning techniques while providing secure user authentication, emotion history and personalized recommendations based on the detected emotion.”

IV. PROPOSED SYSTEM AND METHODOLOGY

The software that will be designed to will follow the Agile SDLC process method. This will be used because of various aspects of the software such as voice processing, detected emotion, security measures, use logs and tailored recommendations. This will also ensure that the software remains flexible, extensible, and adaptable to any change.

A. Planning

The main objective of the application is to create a program that can study human speech and search for mood. The user will be able to provide some sound samples using their voice and microphone or upload audio file. Once the user inputs, the application will be able to process the sound by filtering normalizing trimming and noise reduction.

Secondly, the system also analyses of the acoustic parameters of the recorded audio like pitch and sound energy, MFCCs and spectral information. It will employ AI and machine learning algorithms such as Support Vector Machine (SVM), Convolutional Neural Networks (CNN), Long Short Term Memory (LSTM) and Random Forest for classification and prediction of emotion. The system will record the prediction including the confidence score and timestamp. The app will then generate personalized recommendations like music, guided meditations, breath work, positive prompts and mindfulness meditation.

B. Analysis

The following main functions of the proposed system are discussed:

1. **User Authentication** – Ensuring secure registration and login to protect user's data privacy.
2. **Voice Input Collection** – Supporting real-time microphone capture and files uploads.
3. **Audio Preprocessing** – Filtering noise, cutting empty Pauses and normalizing audio volume level.
4. **Feature Extraction** – Extracting important sound metrics such as MFCC, pitch, intensity, and spectral features.



5. **Emotion Prediction** – Performing classification of emotions using Machine Learning and Deep Learning models including Random Forest, SVM, CNN, and LSTM.
6. **Confidence Scoring** – Calculating probability or confidence of the predicted emotion.
7. **Emotion History Management** – Storing detected emotion, confidence score and timestamp for future use.
8. **Personalized Recommendation** – Recommending appropriate music, meditation, quotes, breathing exercises, mindfulness and wellness tips based on detected emotion.
9. **Dashboard and Output Display** – Showing the results of the emotion prediction, confidence score, recommendations and emotion history

The system is expected to produce high accuracy, reliability and speed of the emotion predictions. Pre-processing and feature extraction are important due to the impact of noise, silence and voice variation on prediction accuracy.

C. Design

This system includes different modules. The User Interaction Module contains the web/mobile interface that allows registration, login, voice recording, and audio file upload. The Audio Preprocessing Module is responsible for noise removal, silence trimming, normalization, filtering, and standardization of the data.

Then the preprocessed audio is sent to the Feature Extraction Module, where MFCC, pitch, energy, spectral features and Mel Spectrograms are calculated. The obtained features are used as an input for the Emotion Prediction Module, where the user's emotional state is classified using machine learning/deep learning models and confidence score is calculated. Emotion, confidence score, and time stamp are stored in the Emotion History Module. In addition, Personalized Recommendation Module gives recommendations to the user, and the Dashboard Module shows emotion, confidence score, recommendations, and emotion history.

D. Implementation

Deployment of the suggested system involves a step-by-step development process. Firstly, web/mobile application and user authentication system are developed. Further, recording using microphone and uploading of audio files functionalities are implemented. Audio pre-processing subsystem is developed that will involve noise reduction, silence trimming, normalization, filtering and standardization. The next step includes the extraction of audio features such as MFCC, pitch, energy, and spectrum features. Machine learning/deep learning models such as Random Forest, SVM, CNN, and LSTM are trained and tested for classification of emotions. Then, selected model is incorporated into the backend for real-time emotion prediction and confidence scoring.

Emotion history and recommendation subsystems are then developed and linked to the dashboard. At last, system is tested for accuracy, performance, usability and reliability. User feedback can help in improving the system and adding other functionalities in future versions.

V. SYSTEM DESIGN

In this section, the five key components of the proposed VocaLyze system are seen. These are: User input & Authentication; Audio Preprocessing; Feature Extraction; Emotion Prediction; & Emotion history with Personalized Recommendations. User input & Authentication, allows users to register and login into the platform securely, while also providing an option to input voice signals by either recording in this environment using a microphone or by uploading an existing audio file. The audio preprocessing component takes the voice signals and pre-processes it by removing noise, trimming of the silent regions, filtering, and normalizing the audio signals. The feature extraction component extracts what comes next key parameters of the speech; Mel Frequency Cepstral Coefficients (MFCC) Pitch Energy, Spectral features & Mel Spectrograms. These extracted features are sent as input to the emotion prediction component, which would use several machine learning algorithms such as Random Forest, Support Vector Machine, Convolution Neural Network and Long Short Term Memory. Emotions will then be classified with an associated confidence score. The last components are the Emotion history and suggestions, which logs the detected emotion and the confidence score and timestamp of detection, and then offers personalized suggestions of music meditation motivational quotes, breathing exercises, mindfulness exercises and wellness tips.

VI. ARCHITECTURE

The proposed system, VocaLyze - AI Based Voice Emotion Recognition System should analyze user's voice and find out the emotion of user from the voice. It allows user to register or login in the system and gives their voice in the form of microphone or upload audio file. Once user gives the voice input, the Audio Preprocessing module preprocesses the input audio by removing noise, silences, normalizes, and then applies the filtering process. This preprocessed audio is



given to Feature Extraction module, which extracts the important features of audio like MFCC pitch energy, spectral, and many others. These generated features are provided as an input to Machine Learning and Deep Learning algorithms like Random Forest SVM CNN, LSTM for finding the emotion.

Next, those predicted emotion and the info related to them are handled by the Emotion History Management module where the emotion, the confidence score and the timestamp are saved. Based on the predicted emotion the Personalized Recommendation module gives the appropriate recommendations like music meditation motivational quotes, breathing exercises, mindfulness, and wellness suggestions, etc. All the information of the users, emotion history and the recommendations are been saved in the database, which is the last module as well. Finally, it is the User Dashboard, which displays the predicted emotion, confidence score, emotion history, and the personalized recommendations. This all modules together build a combined web-based system for voice emotion detection and giving personalized wellness recommendations.

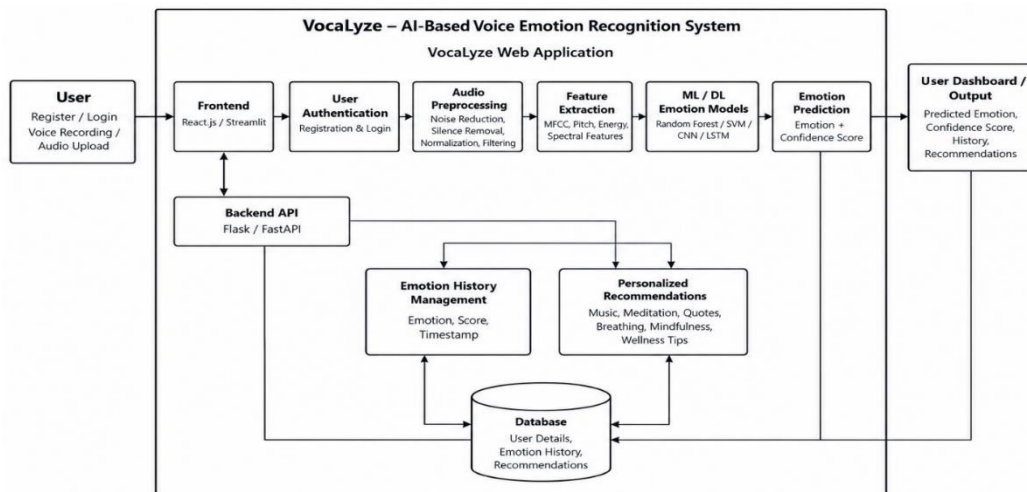


Fig. 1 System Architecture

VII. RESULTS

A. Create Account

Fig. 2. Create Account: The registration page allows new users to enter their name, email, and password to create an account and securely access VocaLyze.

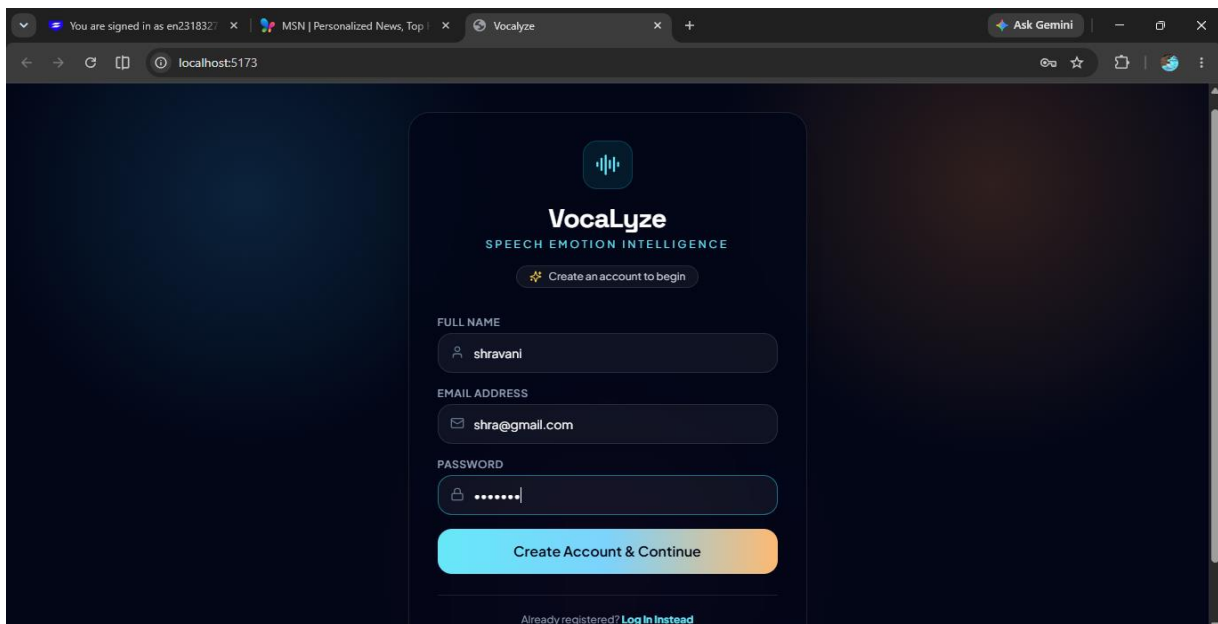


Fig. 2 Create Account



B. Sign in Page

Fig. 3 Sign in Page: The page allows registered users to enter their email and password to securely sign in to the VocaLyze dashboard.

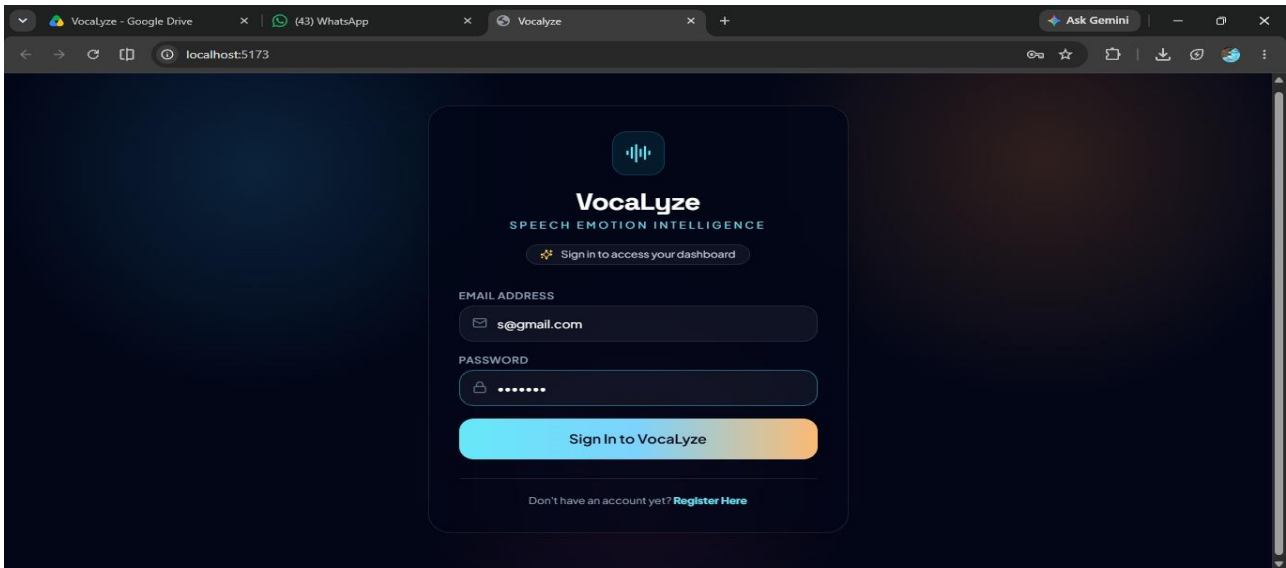


Fig. 3 Sign In

C. Dashboard Page

Fig. 4 Dashboard Page: The dashboard allows users to record or upload voice, analyze the emotion, view the predicted emotion with its confidence score, and receive personalized recommendations.

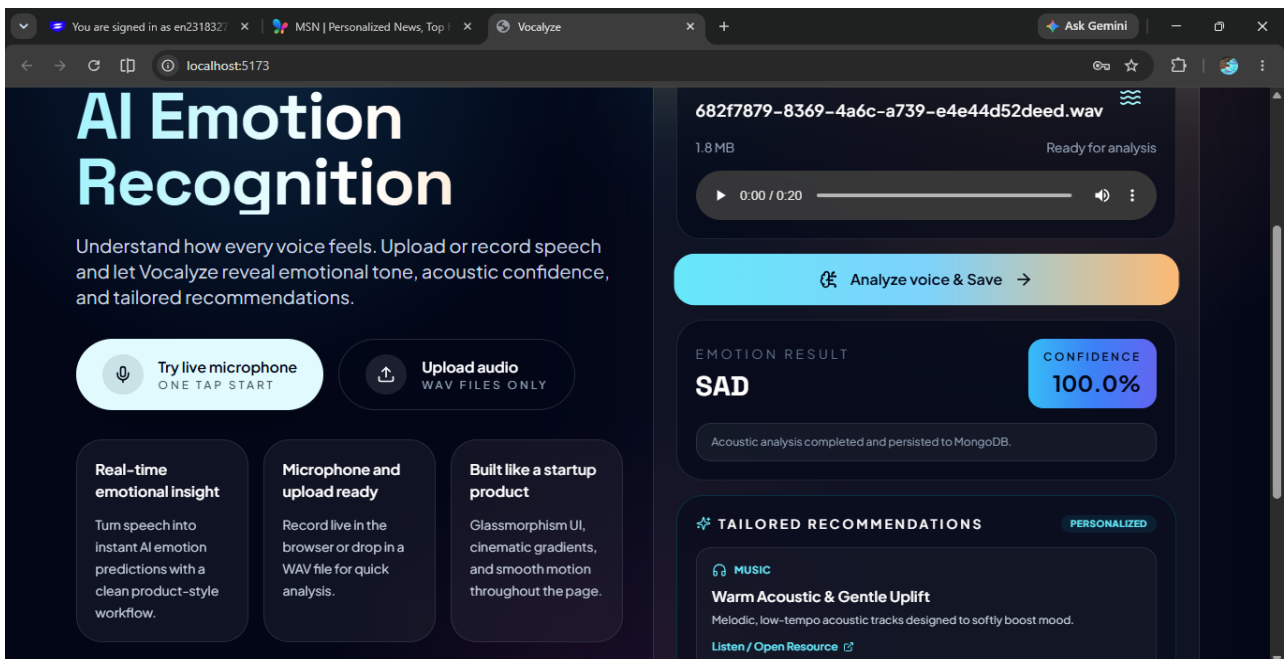


Fig. 4 Dashboard Page

D. Voice Processing

Fig. 5 Voice Processing & History Page: This page displays the user's previous voice emotion analysis results, including the detected emotion, confidence score, voice recording, and personalized recommendations.

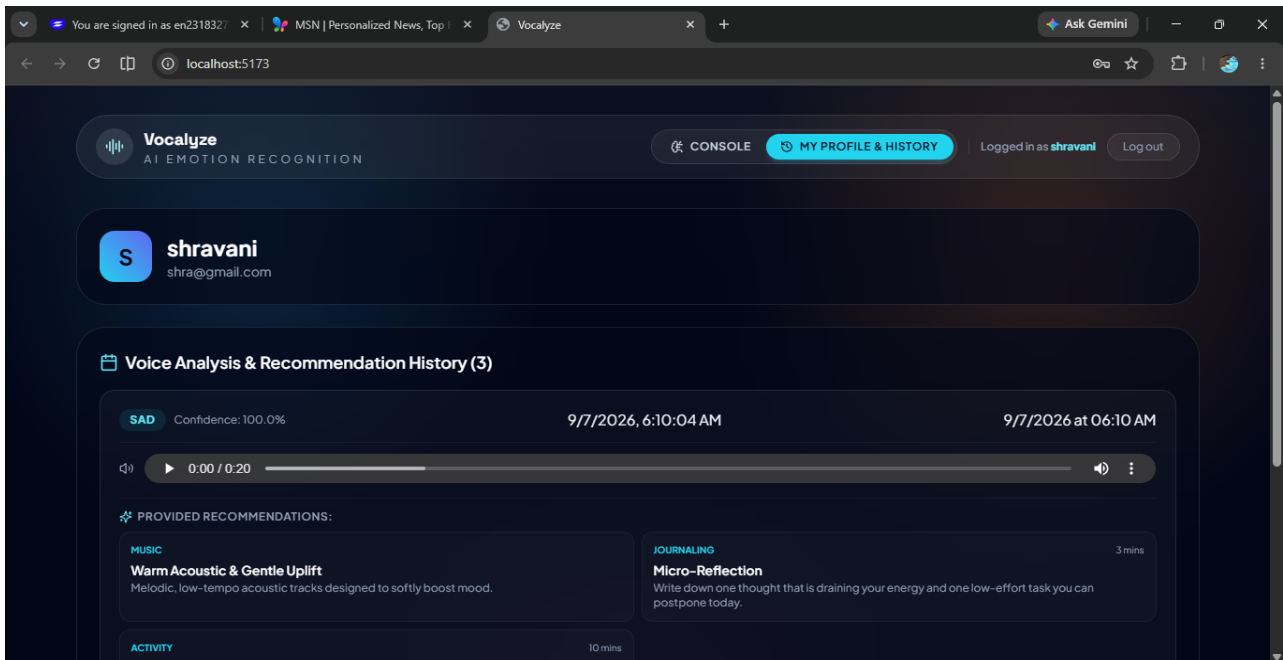


Fig. 5 Voice Processing

E. Recommendation Page

Fig. 6 Recommendation Page: This page displays personalized recommendations based on the detected emotion, such as music, journaling, and physical activities to support the user's emotional well-being.

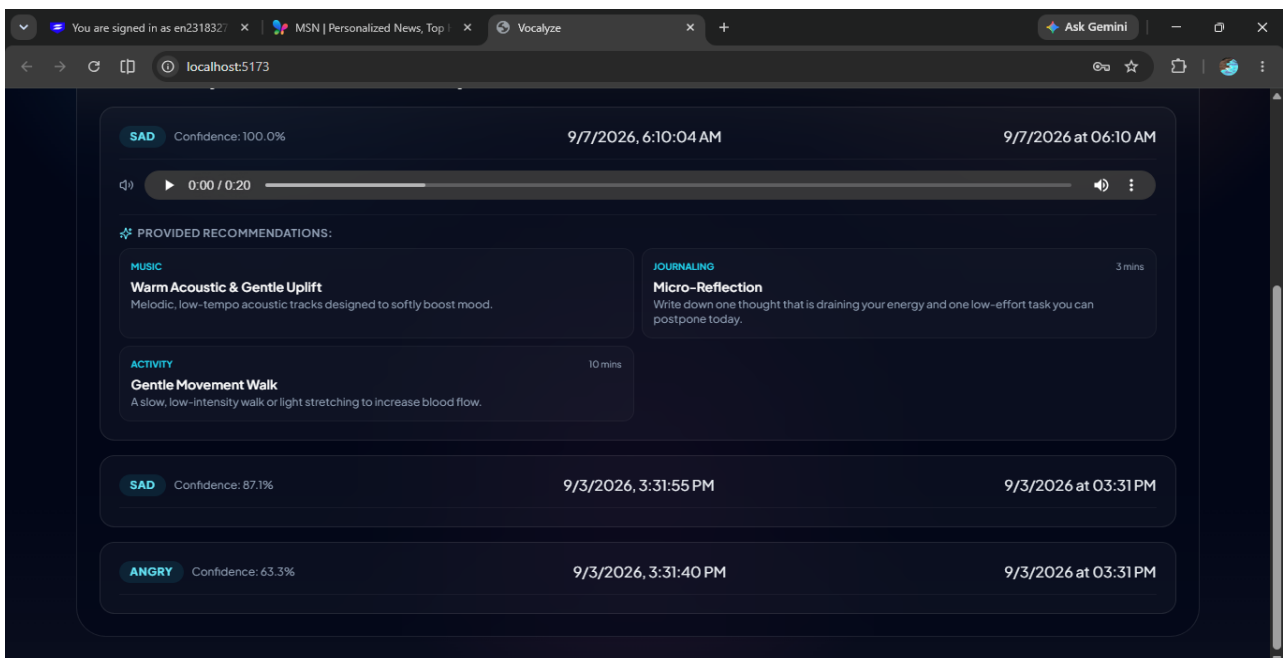


Fig. 6 Recommendation Page

F. Result Page

Fig. 7 Result Page: The result page displays the detected emotion, confidence score, voice analysis status, and personalized recommendations based on the user's voice.

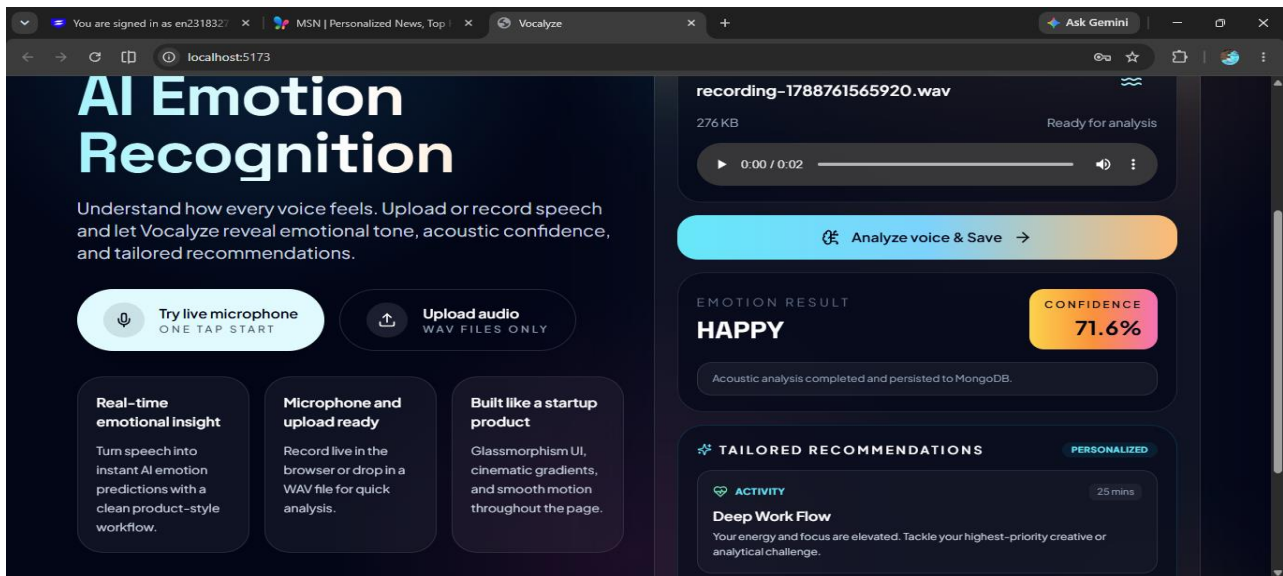


Fig. 7 Result Page

G. Login Database Page

Fig. 8 Login Database Page: This page shows the MongoDB database storing registered user details, including user names, email addresses, encrypted passwords, and account creation dates.

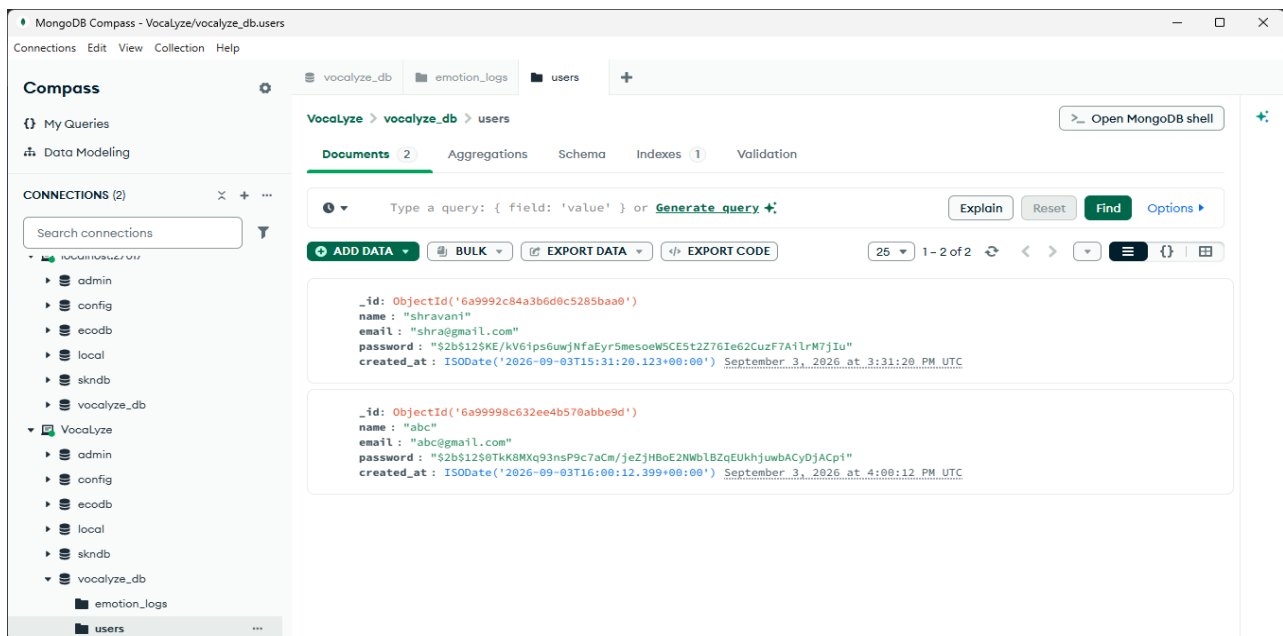


Fig. 8 Login Database Page

H. Result Database Page

Fig. 9 Result Database Page: This page shows the stored voice analysis results, including the detected emotion, confidence score, user ID, and timestamp for each prediction.

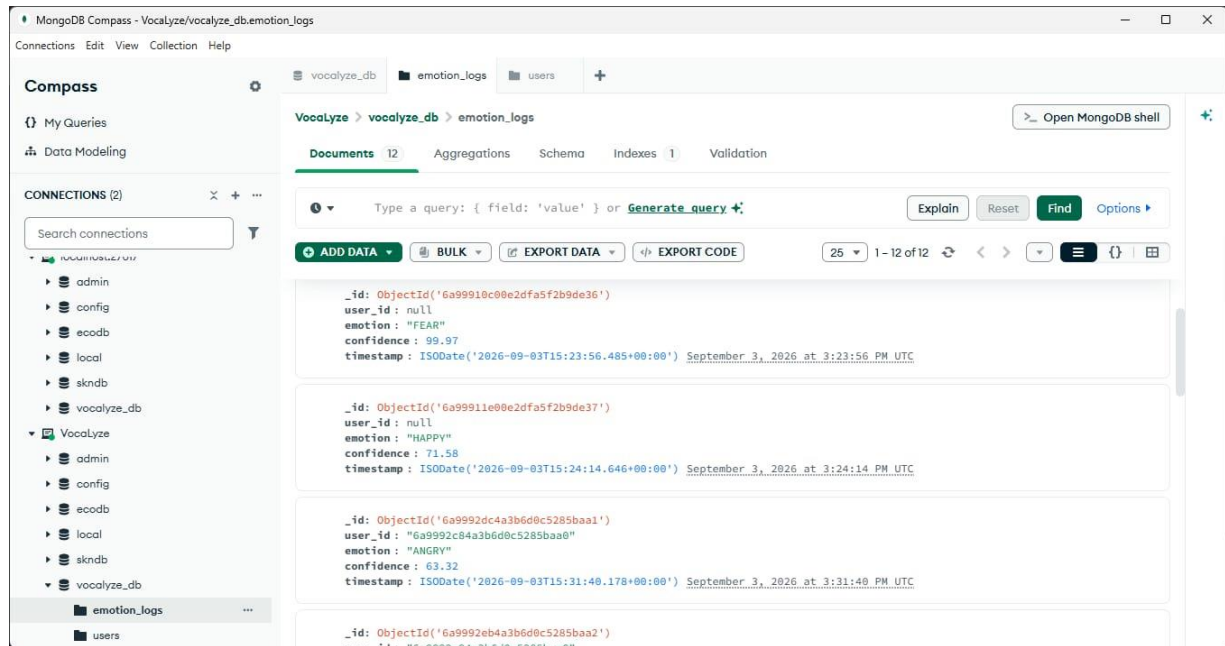


Fig. 9 Result Database Page

VIII. FEASIBILITY ANALYSIS AND DISCUSSION

A. Technical Feasibility

VocaLyze solution is feasible with existing technology. React.js structure shall be used for frontend development and Flask setup for backend development. This machine learning shall make use of scikit-learn, NumPy, and Pandas Python packages. User data, emotion data, and predictions shall be stored in MongoDB database.

B. Economic Feasibility

Cost-effective as the system is composed principally of free software - React.js, Flask, Python, and MongoDB. No additional hardware is required, and the software can be installed on existing cloud infrastructure if needed.

C. Operational Feasibility

The system allows the end user to record or upload voice to the system, and a very user-friendly interface that visualizes the analysis results. End users can check what the predicted emotion, confidence score, emotion history and recommendations from the dashboard.

D. Expected Outcomes

It is claimed that the VocaLyze will detect voices' emotions and offer a confidence figure for the prediction. The platform will also record emotion history and generate personal recommendations for detected emotion such as songs journaling breathing etc.

IX. CONCLUSION AND FUTURE SCOPE

The developed system VocaLyze is able to input voice and analyze speech features to suggest the user's emotion. To input voice, the system uses either microphone recording or file upload of the audio files. The preprocessing we used in this process are denoising, silence removal, filtering, and normalization. Some of the vital speech features, which will be beneficial to determine an emotion are MFCC pitch energy and spectral features.

The system also produces a log of all previous emotion predictions done by it that includes the emotion it detected, confidence score, and timestamp. It also offers tailored advice based on the predicted emotion such as music journaling, deep breathing mindfulness wellness tips, and so on.

More variations of voice data, languages, and real voice conditions may also be introduced to increase the robustness of the system. Other algorithms with Machine Learning and Deep Learning techniques could be improved and tested to



improve the accuracy of initial and prevalent emotion prediction. There may also be future enhancements for in the aspects of emotion history, feedback, and customization. These potential system improvement ideas could make VocaLyze more accurate adaptive user-friendly and functional.

REFERENCES

- [1]. Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding."
- [2]. Liu, Hongpo, and Bo Lang. "Deep Learning-Based Intrusion Detection System."
- [3]. Pallewela, Nuwan, and Damminda Alahakoon. "Blockchain-Based Security Framework for IoT Systems."
- [4]. Jain, Savita, and Dr. Tarun Shrimali. "Intelligent Simulation and Design of Voice-Based Emotion Recognition Using Optimized Methods."
- [5]. Mohammed Mohsin, S. K., and Mohammed. "Artificial Intelligence-Based Healthcare Prediction Model."