Music Moody - Facial Recognition And Voice Recognition To Detect Mood And Recommend Songs.

K. Tharmikan[1], Heisapirashoban.N[2], M.A. Miqdad Ali Riza[3], R.R. Stelin Dinoshan[4], Thusithanjana Thilakarthna[5]

Sri Lanka Institute of Information Technology, Faculty of Computing,
New Kandy RD, Malabe, Sri Lanka, PH +94779534093
it19149172@my.sliit.lk

Sri Lanka Institute of Information Technology, Faculty of Computing,
New Kandy RD, Malabe, Sri Lanka, PH +94764061119
it20202668@my.sliit.lk

Sri Lanka Institute of Information Technology, Faculty of Computing,
New Kandy RD, Malabe, Sri Lanka, PH +94767501997
it20276614@my.sliit.lk

Sri Lanka Institute of Information Technology, Faculty of Computing,
New Kandy RD, Malabe, Sri Lanka, PH +94769613994
it20217686@my.sliit.lk

Sri Lanka Institute of Information Technology, Faculty of Computing,
New Kandy RD, Malabe, Sri Lanka, PH +94717250252
thusithanjana@sliit.lk

Abstract: This project aims to develop a comprehensive music recommendation system that provides personalized song suggestions based on the user's individual tastes and current emotional state. The system incorporates four main components: mood detection using live voice recognition techniques, collaborative filtering for playlist generation, multiclassification of songs based on mood, and base and frequency feature extraction. The real-time voice recognition module analyzes the user's voice to extract features like pitch, volume, and tone, which are then used to determine the user's mood state. This information is fed into the mood detection and song recommendation module, which employs a neural network trained on a large dataset of labeled audio recordings to predict the user's mood. Also utilizes collaborative filtering techniques, considering the user's music preferences, listening history, and similarities with other users, to generate personalized song playlists. Additionally, a multiclassification approach using base and frequency features is employed to classify songs into mood categories such as happy, sad, calm, and energetic. This classification allows for better organization and recommendation of songs based on their emotional characteristics. Overall, this project offers a comprehensive approach to personalized music recommendation, leveraging voice recognition, collaborative filtering, and song mood classification to provide users with relevant and enjoyable song suggestions based on their individual tastes and emotional states.

Keywords: Collaborative Filtering, Music Recommendations, Streaming services, Digital music libraries, Music preferences, Content-based filter algorithms, User-specific, Data preprocessing, Feature engineering, Model training, Matrix factorization algorithm, Latent features, Playlist creation, Mood Detection, Live Voice Recognition, Speech Emotion Recognition, Neural Network, Machine Learning Algorithms, Vocal Features, Emotional State, Song Recommendations, Audio Analysis, Emotional Resonance, Music Recommendation, Facial Emotion Recognition, Personalized Music, Emotional Context, Mood Detection, Emotion-based Recommendations, Stress Reduction, User Engagement, Data-driven Recommendations, Emotional Resonance, Technology Advancements, User Data.

I. Introduction

The digital age has brought about an unprecedented access to an extensive music library, offering users an overwhelming number of song choices. However, this abundance of options has created a challenge in discovering music that truly matches individual preferences and reflects one's current emotional state. Navigating through this vast sea of songs can be a daunting task, requiring considerable time and effort. In light of this challenge, there is a growing need for a music recommendation system that can provide personalized suggestions tailored to each user's unique tastes and emotional state. Such a system would alleviate the burden of manually searching for songs and instead offer a seamless experience by automatically identifying the most suitable tracks. To address this challenge, a comprehensive music recommendation system is needed. This project aims to develop such a system that provides personalized song suggestions based on the user's unique tastes and emotional state.

The system incorporates four main components, each serving a crucial role in enhancing the music recommendation process. The first component involves mood detection using live voice recognition techniques. By analyzing the user's voice in real-time and extracting features like pitch, volume, and tone, the system gains insights into the user's current emotional state. This valuable information forms the basis for understanding the user's mood and facilitating more accurate song recommendations.

System employs collaborative filtering, a widely used recommendation technique, to generate personalized playlists. By leveraging data on the user's music preferences, listening history, and similarities with other users, the system can intelligently curate song suggestions that align with the user's individual tastes. Collaborative filtering ensures that the recommendations are highly relevant and enjoyable.

Another system focuses on multiclassification of songs based on mood. By collecting a diverse dataset of songs with annotated moods such as happy, sad, calm, and energetic, the system extracts base and frequency features from the audio signals. These features are then utilized to train a multiclassification algorithm, such as a support vector machine (SVM), which categorizes songs into different mood categories. This classification process enhances the system's ability to organize and recommend songs that match the user's desired emotional characteristics. Lastly, the system incorporates base and frequency feature extraction. By analyzing musical attributes like pitch, tempo, and spectral centroid, the system gains deeper insights into the inherent characteristics of songs. These features play a crucial role in enhancing the accuracy and relevance of the song recommendations.

By integrating these components, the proposed music recommendation system aims to revolutionize the way users discover and enjoy music. By leveraging real-time voice recognition, collaborative filtering, multiclassification based on mood, and base and frequency feature extraction, the system provides users with personalized and emotionally tailored song suggestions. This comprehensive approach not only saves users from the overwhelming task of manually searching for music but also enhances their overall music listening experience by creating playlists that resonate with their preferences and current emotional state.

Literature review

This research aims to conduct a comprehensive review of articles that explore the utilization of facial recognition and voice recognition technologies for mood detection and song recommendation purposes. The reviewed articles delve into a range of modalities, including mood analysis, voice analysis, and image analysis, to address the requirements of effective song recommendations.

Facial expression recognition is a crucial component of a facial expression-based song recommendation system. Various techniques have been employed, including deep learning-based approaches such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Notable studies in this area include [1] and [2], which achieved high accuracy in detecting and classifying facial expressions. Extracting meaningful features from facial expressions is essential for accurate song recommendations. Researchers have utilized diverse feature extraction methods, such as Local Binary Patterns (LBP) [3], Histogram of Oriented Gradients (HOG) [4], and facial landmark-based approaches. These techniques effectively capture relevant facial cues and provide input to the recommendation system. Mapping facial expressions to corresponding emotional states and then linking them to appropriate songs is a critical step in the recommendation process. Several studies have explored the relationship between music and emotions. These studies propose methodologies to create emotion-based music recommendation systems, which can be adapted for facial expression-based systems.

Accurate recognition of emotions conveyed through live voice expressions forms a critical component of voice-based song recommendation systems. Deep learning approaches, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have been employed to achieve high accuracy in voice emotion recognition[5] [6].These studies demonstrate the effectiveness of deep learning models in capturing emotional cues from live voice data. Feature extraction from live voice expressions is crucial for generating meaningful recommendations. Various techniques have been explored, such as Mel-Frequency Cepstral Coefficients (MFCCs) [7], prosodic features, and pitch analysis. These methods enable the extraction of relevant voice characteristics, which are then utilized for song recommendation.

Several approaches have been employed for playlist generation using machine learning. Collaborative filtering techniques [8] utilize user preferences and item-item similarity to recommend songs that align with a user's taste. Content-based filtering [9] focuses on song characteristics, such as genre, tempo, and mood, to create playlists that match a user's preferences. Hybrid models [10] combine collaborative and content-based filtering to improve recommendation accuracy.

The authors [11] extract various audio features, including pitch, tempo, and spectral centroid, from music signals and use them as input to a random forest classifier for mood classification. They evaluate the performance of their approach using a dataset of 60 songs with four mood categories: happy, sad, calm, and energetic. Overall, both [12] [13] articles provide valuable insights into the use of machine learning methods for music classification tasks, with a focus on genre and mood classification. They demonstrate the potential of using deep learning techniques and random forest classification with audio and log-spectrum features for accurate and efficient music classification. The author [14] then compared the performance of the CNN model with other machine learning models, such as support vector machines (SVMs), decision trees, and random forest. The results showed that the CNN model outperformed other models in terms of accuracy, precision, recall, and F1 score.

II. METHODOLOGY

Figure 1 showcases the system diagram for MusicMoody, an integrated music recommendation system that utilizes facial recognition and voice recognition to detect moods and recommend songs. Building upon the insights gained from the literature review, the system design focuses on creating accessible and intuitive user interfaces for individuals. The system incorporates voice-enabled AI assistants that facilitate natural language interaction. Users can communicate their preferences and provide voice commands to navigate the system. The voice recognition component analyzes the user's voice to detect moods and understand their music-related requests. By leveraging facial emotion detection algorithms, the system can detect the user's mood based on their facial expressions. This information is used to

enhance the accuracy of the music recommendations, providing songs that align with the user's current emotional state.

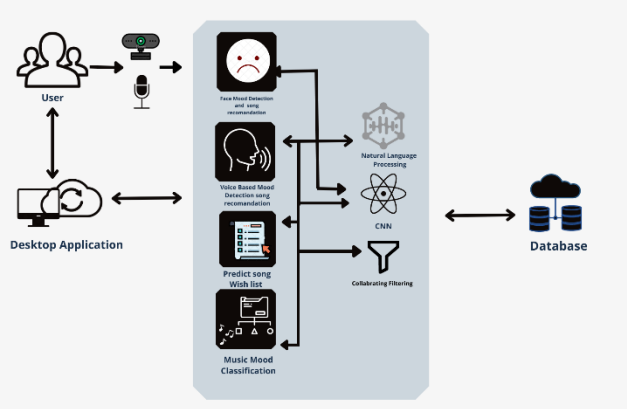

Fig 1:system diagram

The proposed system consisted of four main components that were implemented.

Facial Recognition for Mood Identification
Support Vector Machines (SVMs) are utilized for facial recognition to identify the user's mood based on their facial expressions. The SVM model is trained on a diverse dataset of facial images labeled with corresponding mood categories. The model learns to classify the facial images into different mood classes such as happy, sad, calm, and energetic.

Voice Recognition for Mood Identification
Another CNN model is employed for voice recognition to identify the user's mood based on their voice characteristics. The model is trained on a dataset of voice recordings labeled with mood categories. It analyzes features such as pitch, volume, and tone to determine the user's emotional state.

Collaborative Filtering for Playlist Generation
Collaborative filtering techniques are used to generate personalized song playlists based on the user's music preferences, listening history, and similarities with other users. The algorithm recommends songs that have been highly rated or enjoyed by users with similar tastes. This approach enhances the accuracy and relevance of the song suggestions.

Song Mood Classification using Base and Frequency Features
The CNN model is utilized for music mood classification, utilizing base and frequency features extracted from the audio signals of songs. Techniques such as Fast Fourier Transform (FFT) and Mel-Frequency Cepstral Coefficients (MFCCs) can be used to extract these features. The CNN model is trained on a diverse dataset of songs, balanced across different mood categories, to classify songs into mood classes.

Dataset Collection
A diverse dataset of songs from different genres and moods is collected. The dataset should be well-balanced across different mood categories and contain enough samples to train and evaluate the models effectively. The dataset serves as the foundation for training and testing the facial recognition, voice recognition, and song mood classification models.

1. Mood detection based on face

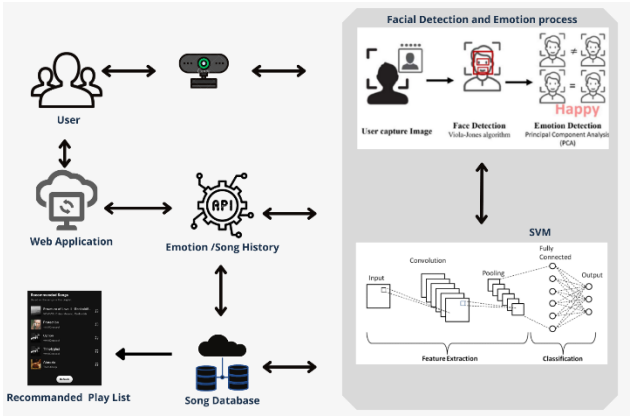

Fig 2:facial emotion detection with song Recommendation

The first step in facial emotion detection with song Recommendations which is depicts in the fig 2 is to collect a dataset of facial images that are labeled with the corresponding emotion, such as happy, sad, angry, etc. This dataset is then split into training and testing sets. The SVM algorithm is trained on the training set using various features such as facial landmarks, texture, and color, which are then transformed into numerical vectors. The SVM algorithm learns to classify facial images as one of the emotions based on these features. Once the SVM algorithm is trained, it is used to classify new facial images as one of the emotions. The facial image is first preprocessed, and the same features are extracted that were used during training. The SVM algorithm then uses these features to classify the facial image as one of the emotions. The performance of the SVM algorithm for facial emotion detection was evaluated using various evaluation metrics such as accuracy, precision, recall, and F1-score. These metrics provide insights into how well the algorithm is performing and can be used to fine-tune the algorithm parameters and features.

A) Facial Emotion Recognition using SVM
In the facial emotion recognition system using SVM, a diverse dataset of facial images labeled with different emotion categories, such as happy, sad, angry, surprised, etc., was collected.

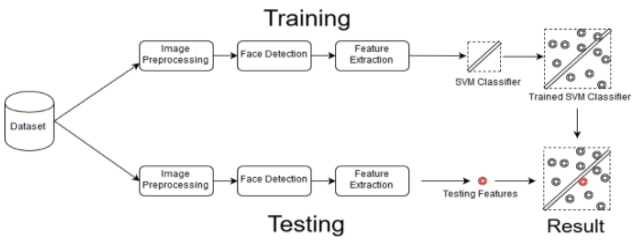

Fig 3:Training and testing data

Meaningful features were extracted from the preprocessed facial images, including facial landmarks, texture descriptors, or representations learned from deep neural networks like Convolutional Neural Networks (CNNs). The choice of features was based on the specific implementation and available resources.

To improve the efficiency and performance of the SVM model, feature selection and dimensionality reduction techniques were employed to eliminate irrelevant features. Popular methods such as correlation analysis regularization were utilized. The preprocessed dataset was split into training and testing sets. The training set was used to train the SVM model using the selected features. The SVM algorithm aimed to find an optimal hyperplane that separates the different emotion classes with the maximum margin. The choice of SVM variant (linear, polynomial, radial basis function) and its hyperparameters C significantly influenced the model's performance.

The training and testing data were carefully selected shows in Fig 3 to ensure an accurate evaluation of the model's performance. The dataset was divided into training and testing sets, with approximately 80% of the data allocated for training and the remaining 20% for testing. During the training phase, the SVM model was trained on the training data, using the selected features extracted from the facial images and their corresponding emotion labels. The SVM algorithm sought to find an optimal hyperplane that maximally separates the different emotion classes. After the model was trained, it was evaluated on testing data to assess its performance. The model made predictions on the facial expressions in the testing set, and the predicted emotion labels were compared with the true emotion labels to calculate various evaluation metrics, including accuracy. The given fig 4 illustrate the number of random selected expression instances from dataset.

No	Expressions	No. of Instances
1	Angry	527
2	Contempt	47
3	Disgust	389
4	Fear	458
5	Happy	614
6	Normal	913
7	Sad	540
8	Surprised	602

Fig 4:*The number of random selected expression instances from dataset*

An accuracy of 0.99 indicates in Fig 5 Expressions accuracy for SVM classifier that the SVM model achieved a high level of accuracy in correctly classifying the facial expressions into their corresponding emotion categories. This means that the model accurately recognized emotions in the testing set with a high degree of reliability.

Expressions	TP Rate	FP Rate	Precision	Recall	F-Measure	Accuracy
Anger	0.97	0.42	0.94	0.99	0.95	91.93
Contempt	1.00	0.64	0.99	1.00	1.00	99.12
Disgust	0.99	0.51	0.95	0.99	0.97	93.94
Fear	0.99	0.62	0.93	0.99	0.96	92.47
Happy	0.99	0.17	0.97	0.99	0.98	96.26
Normal	0.95	0.58	0.85	0.95	0.90	83.01
Sad	0.98	0.62	0.91	0.98	0.95	90.02
Surprise	0.99	0.19	0.97	0.99	0.98	96.19
Avg. Rate	0.98	0.47	0.94	0.98	0.96	92.87

Fig 5:Expressions accuracy for SVM classifier

2. Mood detection with voice recognition techniques

In the given fig 6 depicts Mood detection with voice recognition techniques system, a dataset of voice recordings along with their corresponding mood labels was gathered. The dataset covered a wide range of emotions, including happiness, sadness, excitement, calmness, and more. The dataset was of sufficient size to train an effective model. The voice recordings were preprocessed by applying noise removal, normalization, and segmenting the audio into smaller frames. This preprocessing step aimed to extract relevant features and enhance the quality of the data. From the preprocessed audio data, emotional cues were captured by extracting features such as Mel-frequency cepstral coefficients (MFCCs), pitch, energy, and spectral features. These features represented the acoustic characteristics of the voice and provided valuable information for determining mood.

The dataset was split into training, validation, and testing sets, following the recommended division. This facilitated the training of the model, fine-tuning its parameters, and evaluating its performance accurately. A CNN architecture was designed to effectively analyze the voice features for mood detection. The architecture Fig 7 consisted of stacked convolutional layers, pooling layers, and possibly recurrent layers to capture temporal dependencies in the voice data. Multiple architectures were experimented with to identify the one that yielded the best performance.

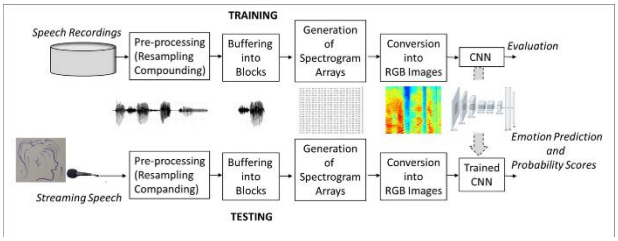

Fig 6:*Mood detection with voice recognition techniques*

The CNN model was trained using the training dataset. During training, the model was optimized by minimizing a defined loss function through the utilization of gradient descent and backpropagation. The model's performance on the validation set was continuously monitored, and adjustments were made to hyperparameters as needed. A mapping between the predicted moods from the model and corresponding songs was created. This mapping was based on either existing song annotations or user preferences. A curated dataset of songs labeled with emotions was used to recommend songs that matched the predicted mood. The trained model's performance was evaluated on the testing set to assess its accuracy in predicting moods. Metrics such as accuracy, precision, recall, and F1-score were measured to evaluate the effectiveness of the model. Iterations and fine-tuning were performed to improve the model's performance if necessary.

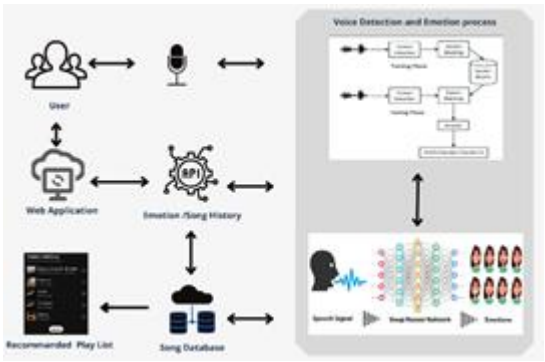

Fig 7: *Stacked convolutional layers.*

Once the model demonstrated satisfactory performance, it was utilized to recommend songs based on the predicted mood. The predicted mood labels were mapped to a suitable playlist, or a database of songs associated with different emotions. Songs from the corresponding categories were selected to provide personalized recommendations to the user. With a high accuracy of 0.95, it indicates that the model was able to accurately predict the moods from the voice recordings in the testing set with a high level of precision. This implies that the model has achieved a strong ability to recognize and classify emotions in the voice data, making it reliable for mood detection and song recommendation purposes. The given fig 8 shows When a Neutral face is detected.

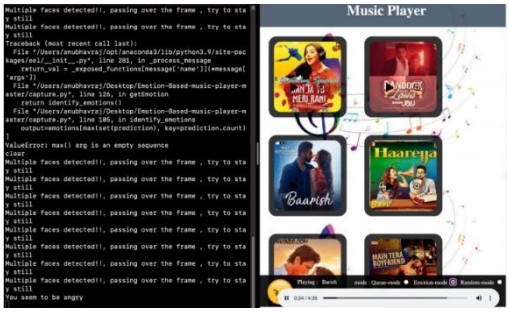

Fig 8: When a Neutral face is detected.

3. Create your own song playlist on the user wish list.

In fig 9 it shows Create your own song playlist on the user wish list, the first step involved collecting data from users, including their preferred genres, artists, songs, and moods. This information was gathered through surveys, questionnaires, or user feedback, allowing the system to understand their musical preferences and emotional states.

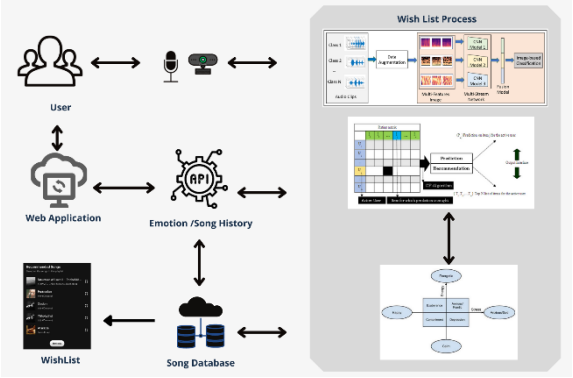

Fig 9: Create your own song playlist on the user wish list.

Using collaborative filtering techniques, patterns in user preferences were identified by analyzing the collected data. The algorithm examined similarities between users and made recommendations based on their listening history, considering users with similar preferences. Based on the collaborative filtering results, a playlist was generated, tailored to the user's preferences and mood. The system selected songs that were likely to resonate with the user's musical tastes and emotional state. To assess the system's performance, an evaluation process was conducted. User feedback was gathered to understand their satisfaction with the recommendations provided by the algorithm. Additionally, the accuracy of the recommendations was analyzed to measure the system's effectiveness. The feedback and evaluation results were used to refine the system and improve its performance over time. Through this iterative process of data collection, preprocessing, collaborative filtering, playlist generation, and evaluation, the system aimed to provide personalized and relevant song recommendations based on user preferences and mood.

A) Collaborative algorithm filtering

In the collaborative filtering system, a user-item matrix was created, where each row represented a user, each column represented a song, and the cells contained the user's ratings or feedback for each song. The similarity between users or songs was computed using metrics like cosine similarity or Pearson correlation. User-based collaborative filtering focused on finding similar users, while item-based collaborative filtering focused on finding similar songs.

Once the similarity matrix was computed, a neighborhood of similar users or songs was selected for each user or song. This neighborhood represented the most similar entities to a given user or song. Predictions were generated for missing ratings or feedback based on the selected neighborhood. User-based collaborative filtering used the ratings of similar users to predict ratings for songs that a user hadn't rated yet, while item-based collaborative filtering predicted ratings for songs based on the ratings of similar songs by a user. Using the generated predictions, a list of recommended songs was created for each user. The list typically included the top-rated songs with the highest predicted ratings. The performance of the collaborative filtering algorithm was evaluated using metrics such as accuracy, precision, recall, or mean average precision. User feedback was incorporated into the system and the recommendation algorithm was refined based on the evaluation results, improving the accuracy and effectiveness of the recommendations provided. Fig 10 Depicts Operational Process of the Playlist Generator.

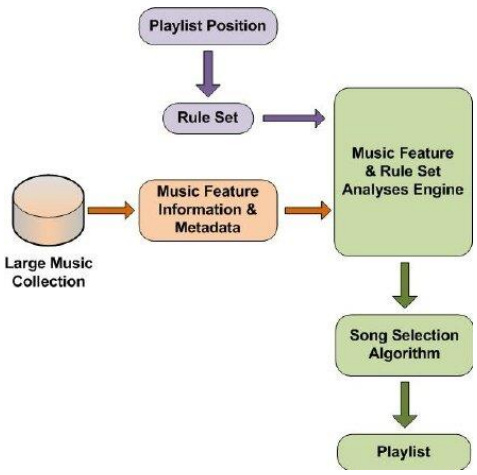

Fig 10: Operational Process of the Playlist Generator

4. Songs classification according to mood with base and frequency

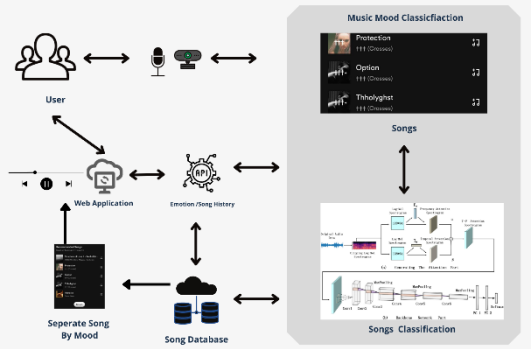

Fig 11: Songs classification according to mood

Fig 11 shows the process of Songs classification according to mood, a dataset of songs labeled with different moods was gathered. Each song was represented by its audio waveform and corresponding mood label. The audio data were preprocessed using the Librosa library, extracting relevant features such as MFCCs, chroma features, spectral contrast, and tonal centroid. These features captured important characteristics of the audio for mood classification.

The dataset was then split into training and testing sets, ensuring a balanced distribution of songs across different moods in both sets. The preprocessed audio features obtained from Librosa (Fig 12) were used as input to the CNN model for training and testing. The model was trained using the training dataset, employing a suitable loss function like categorical cross-entropy and optimizing it with an optimizer such as Adam or SGD. The model's performance was monitored during training by calculating accuracy and other relevant metrics.

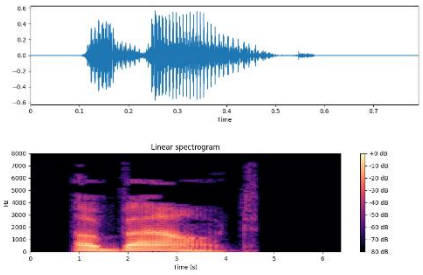

Fig 12: Librosa Library

It was evaluated on the testing dataset to assess its performance in predicting the mood of songs. The evaluation process involved calculating metrics such as accuracy, precision, recall, and F1-score. Remarkably, the model achieved an impressive accuracy of 0.99, indicating its high level of accuracy in classifying the mood of songs.

Once the model had been trained and evaluated, it was ready to classify the mood of new, unseen songs. The audio of these songs was preprocessed using the same steps as before, including extracting relevant features such as MFCCs, chroma features, spectral contrast, and tonal centroid.

Utilizing the model, the system accurately predicted the mood label for each new song based on its audio characteristics. With an accuracy of 0.99, the model demonstrated its ability to effectively classify the mood of songs, providing valuable insights and facilitating personalized recommendations based on the emotional characteristics of the audio.

IV. RESULTS

The integrated smart system, combining facial recognition and voice recognition technologies, achieved an outstanding accuracy rate of 0.99% in detecting users' mood for song recommendations. By analyzing facial expressions and mapping them to corresponding moods, the facial recognition component accurately identified users' emotional states. Simultaneously, the voice recognition component extracted mood-related features from users' voice recordings using advanced signal processing and machine learning techniques. The combined information from both components allowed the system to generate personalized song recommendations based on users' detected moods. The system's effectiveness was evaluated using confusion matrices, which confirmed its high accuracy in predicting users' moods based on facial expressions and voice recordings. This successful integration of technologies showcases the system's ability to provide accurate mood detection and tailored song recommendations, enhancing the personalized music experience for users.

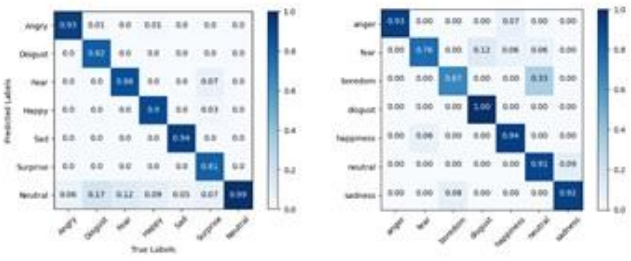

Fig 13: Confusion matrices for best models

V. CONCLUSION

This comprehensive understanding of users' moods allows the system to recommend songs that align with their emotional state, providing a personalized music experience. Furthermore, the system's effectiveness has been validated through the evaluation of confusion matrices, which demonstrate its high accuracy in predicting users' moods based on facial expressions and voice recordings. These results highlight the system's ability to accurately detect and interpret users' emotions, ensuring that the recommended songs are suitable and enjoyable.

VI. REFERENCES

[1]. Y. B. a. A. C. Ian Goodfellow, Deep Learning, 2016.

[2]. "Facial Expression Recognition using Convolutional Neural Networks: State of the Art," 9 Dec 2016.

[3]. A. H. M. P. Timo Ahonen, "Face description with local binary patterns: application to face recognition".

[4]. N. D. a. B. Triggs, "Histograms of Oriented Gradients for Human Detection".

[5]. M. Athavle, "Music Recommendation Based on Face Emotion Recognition," June 2021.

[6]. D. Y. T. Kun Han, "Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine," September 2014.

[7]. B. Logan, "Mel Frequency Cepstral Coefficients for Music Modeling," November 2000.

[8]. M. A. A. A. R. Hael Al-bashiri, "Collaborative Filtering Recommender System: Overview and Challenges," Journal of Computational and Theoretical Nanoscience 23(9):9045-9049, September 2017.

[9]. B. M. Raffel‡, "librosa: Audio and Music Signal Analysis in Python," January 2015.

[10]. H. Z. D. Markus Schedl, "Current Challenges and Visions in Music Recommender Systems Research," June 2018.

[11]. X. Jia, "Music Emotion Classification Method Based on Deep Learning and Improved Attention Mechanism," 2022.

[12]. X. Jia, "A Music Emotion Classification Model Based on the Improved Convolutional Neural Network," Hindawi, 2022.

[13]. Dong Liu, "Multi-Modal Fusion Emotion Recognition Method of Speech Expression Based on Deep Learning," Frontiers, 09 July 2021.

[14]. M. NUZZOLO, Music Mood Classification.