iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.81664

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

Multi-Modal Emotion Based Music Recommendation System

Mrs. N Jaya Santhi[1], Sree Mallika Maila[2], Venkata Swetha Gollamudi[3], Haveela Akunuri[4]
Department of Computer Science and Engineering, Bapatla Women's Engineering College, Srinivas Nagar, Bapatla, Andhra Pradesh, India

Abstract: ***The issue of comprehending human emotions has a significant influence on the user experience in the digital world. The conventional technique of recommending music is primarily based on the previous listening behavior of users, ratings, or playlists. But this approach does not take into account the current emotional status of the user. Several current emotion detection frameworks are based on one modality only, such as text or facial expression. Such systems can be ineffective in practical applications due to the complicated nature of human emotions that vary according to context and personal behavior.***
This limitation can be addressed through the proposed use of artificial intelligence, machine learning, and deep learning technologies for building an emotion recognition framework that will incorporate all the aforementioned elements to recognize emotions in a user in an improved manner. The combination of several inputs makes it easier for the system to detect emotions. Facial expressions, voice tone, and text analysis can be used together to give better insights about emotions and reduce possible inaccuracies arising from single-input recognition methods.
This system is developed into an interactive web application through the use of Flask programming language. Once the emotion of the user is detected, he/she is suggested personalized music that correlates to his/her mood state. In this way, user experience can be significantly enhanced through personalized suggestions of music depending on their current emotional state.
Keywords: Multi-modal Emotion Recognition, Music Recommendation System, Artificial Intelligence, Deep Learning, Personalized Recommendation.

I. INTRODUCTION

Music is a significant part of human life, and it is closely related to people's emotions. Individuals usually select music according to their emotional state, for example, choosing cheerful music when they are joyful and soothing music when they are depressed or anxious. Unfortunately, current music recommender systems rely on past data and user profiles. They ignore users' emotional states, which means that recommendations are unlikely to be aligned with individuals' current moods.

Advancements in AI, ML, and DL in recent years have enabled researchers to develop algorithms to recognize human emotions more accurately. Most of the existing algorithms concentrate on one type of data, including face expression recognition, speech signal analysis, and text sentiment analysis, respectively. Nevertheless, human emotions are intricate and dynamic. Hence, the algorithms relying on one data source are insufficient to represent humans' emotional states. It leads to low accuracy and inappropriate predictions [6], [8].

However, the use of a single type of data for analysis does not make it possible to accurately understand the emotional state of users. In order to improve the accuracy and completeness of the detection process, it is proposed to implement a multi-modal emotion recognition-based music recommendation system, which analyzes emotions through facial expressions, voice tone, and written text. Machine learning and deep learning algorithms will be applied for emotion recognition from multiple types of input data. Facial expression analysis will be performed based on deep learning Convolutional Neural Network models, while audio and text-based analysis of emotions will be carried out using machine learning methods.

Music recommendation based on user emotions will be implemented via a Flask web-application, thus allowing to create a convenient and efficient tool to detect user emotions and make personalized music suggestions. Based on the detected emotion of the user, a set of suitable music tracks will be suggested to the user – cheerful music when the user is happy, and relaxing tracks when the user feels sad and stressed.

Overall, the development of a multi-modal emotion recognition-based music recommendation system is aimed at improving the quality and efficiency of user interaction with the machine.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

II. LITERATURE SURVEY

In their research work, Kumar et al. [1] investigated the application of Artificial Intelligence in emotion classification through social media posts. The findings indicated that sentiment analysis of emotions through AI techniques was highly effective. The results of this study indicate the significance of the text-based emotion classification approach since it can be applied in developing mood-based music recommendation systems.

Islam et al. [3] examined emotion-related topics by analyzing the text data provided by users of online platforms such as Reddit. In their study, the authors confirmed that the data generated by users can serve as an essential means for detecting emotions. This finding confirms the usefulness of using text sentiment analysis in multi-modal approaches for emotion detection.

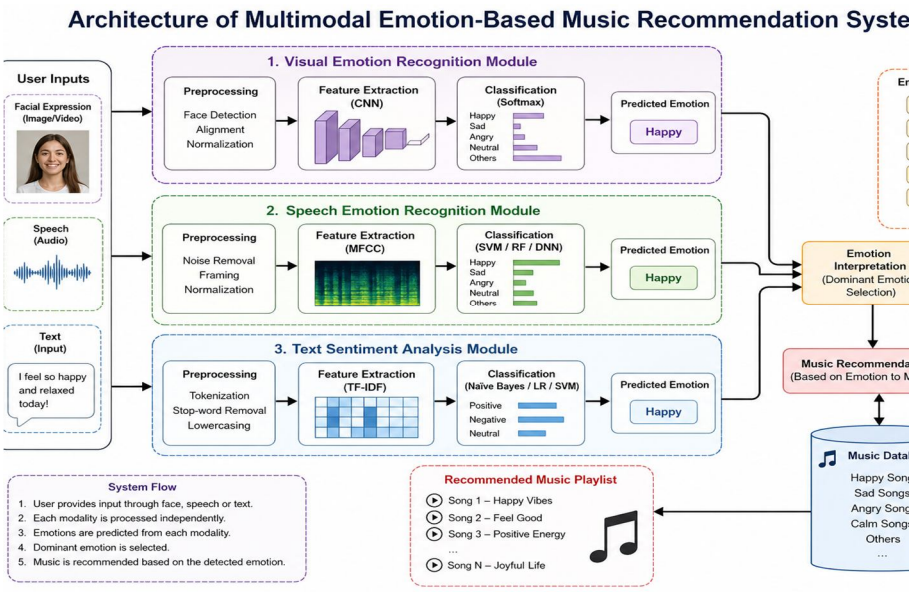
In their research paper, El Ayadi et al. [5] proposed a real-time speech emotion recognition framework through deep learning and data augmentation. The study revealed that speech signals contained significant features such as pitch, tone, and intensity, which were highly correlated to emotions.

In the same manner, studies done by Abadi et al. [6] also explored emotion detection by using machine learning techniques on speech analysis. They demonstrated that differences in tone and pitch have high correlation with emotions. It also correlates well with using speech-based models such as the CNN-LSTM used in this system.

Additionally, Gupta et al. [8] proved the applicability of deep neural networks in speech emotion recognition in real time. The authors point out the significance of employing deep learning techniques in order to increase the accuracy in emotion detection.

Moreover, Poria et al. [9] made a great contribution to affective computing through the introduction of multi-modal emotion recognition. The research conducted by the authors proves the effectiveness of multi-modal inputs, such as text, speech, and facial expressions, in increasing the performance of emotion detectors. This technique is implemented in the current project.

III. METHODOLOGY



Proposed diagram

The suggested approach provides a platform for the detection of multi-modal emotions using Artificial Intelligence. The approach caters to three different modes of data:

A. Analysis of Emotions on the Face

Facial emotion recognition helps to understand the emotions of the users based on their facial expressions. In this approach, Convolutional Neural Network algorithms including ResNet and VGG16 are employed for the classification of emotions from faces. Emotions like happy, sad, angry, and neutral are detected using facial characteristics including eye motion and lips' positions. Facial emotions are detected by training the system on various databases like FER2013.

B. Emotion Based on Speech

Emotion recognition through speech involves the analysis of emotions in the voice of the user. This technique uses audio recordings to create either spectrogram or MFCC features that are further used for emotion recognition through deep learning algorithms such as CNN or CNN-LSTM. The system recognizes changes in emotions based on factors such as pitch, tone, and speech rate. This process is helpful as emotions can be conveyed through voice despite unclear facial expressions.

C. Emotion Analysis Based on Text

The text-based emotion detection technique is one that involves detecting emotions from the text provided by the users. In this case, the system applies NLP algorithms together with SVM and CNN algorithms to analyze the text. Emotions such as positive, negative, or neutral are detected depending on the wording in the text.

1) Data Collection Layer

These three sources include the expression of the face through a webcam, speech through a microphone, and text through typing. This multi-channel strategy enables the system to detect emotions more precisely while avoiding reliance on just one mode of information input.

2) Preprocessing Layer

Once the data has been collected, it undergoes a process known as preprocessing. This involves cleaning the data and getting it ready for use. In the case of facial images, normalization and resizing are done to emphasize certain features. Similarly, audio data undergoes pre-processing such as filtering out of noise to convert the data into more meaningful features such as MFCCs. Textual data is preprocessed by removing unnecessary words through NLP.

Once the data has been collected, it undergoes a process known as preprocessing. This involves cleaning the data and getting it ready for use. In the case of facial images, normalization and resizing are done to emphasize certain features. Similarly, audio data undergoes pre-processing such as filtering out of noise to convert the data into more meaningful features such as MFCCs. Textual data is preprocessed by removing unnecessary words through NLP.

3) Ai-Based Feauture Extraction And Classification

The critical information that needs to be mined out of the data is identified and is used for classifying emotions in this layer. ResNet and VGG16 are the kinds of CNN models that will be used to recognize emotions from faces. The information in speech will be analyzed using CNN-LSTM models in order to classify emotions on the basis of pitch and tone in voice.

4) Decision Making Layer

The system then analyses the output from the facial, speech, and text analysis modules and identifies the final emotional state of the user. As opposed to measuring the levels of stress, the system measures different emotions, which include happiness, sadness, anger, or calmness.

5) Music Recommendation Layer

The recommendation for suitable music depends on the detected mood of the user. When the user feels happy, the system suggests upbeat music. On the other hand, when the user is sad, then the music will be slow, and when the user is angry, the music becomes soothing.

6) Deployment And User Interface Layer

The detection of emotions of the users along with offering them music recommendations in real-time can be done via a web application framework like Flask. The Flask framework will serve the purpose of the backend and deal effectively with inputs from the user. For the development of a front-end for the interface, the technologies used will be HTML, CSS, and JavaScript.

Through this, the users will be able to submit image or video of their faces, voice recordings, or any other input in the form of text that may help in the process of detecting their emotions. The emotions detected from the inputs are then analyzed by the system to give recommendations regarding music instantly.

7) Output And Real-Time Feedback Layer

Quickly after user input, the system identifies the user's emotional state such as happy, sad, angry, or calm and displays the result on the screen. Based on the detected emotion, the system provides suitable music recommendations that match the user's mood. Users can immediately listen to the suggested songs, which helps improve their overall experience.

By using multimodal emotion detection and real-time processing, the system delivers fast, accurate, and personalized music suggestions. This makes the system easy to use, interactive, and effective in enhancing user satisfaction through emotion-based music recommendations.

IV. IMPLEMENTATION

A. Data Collection

The three major sources of information that the system relies on are facial expressions, speech data, and textual input. The former two sources are recorded in the form of an image or video and analyzed with the use of CNN models, focusing on the features of the face and finding patterns related to certain emotions. Speech data are recorded with a microphone, and then they are analyzed by applying the methodology of deep learning along with features like spectrograms or Mel Frequency Cepstral Coefficients. Textual input is gathered through the users' interactions with the system and analyzed by using natural language processing.

B. Machine Learning Models

In the emotion classification module, several machine learning and deep learning algorithms can be used in the present system. In case of text-based emotion detection, support vector machine (SVM) algorithm can be used in sentiment analysis for classifying emotions. In case of audio-based input, the speech signals need to be processed using certain feature extractions such as spectrograms and Mel-frequency cepstral coefficients (MFCC) and machine learning models for detecting emotions. On the other hand, convolution neural networks (CNN) can be applied in both cases including face image-based emotion detection using datasets like FER2013.

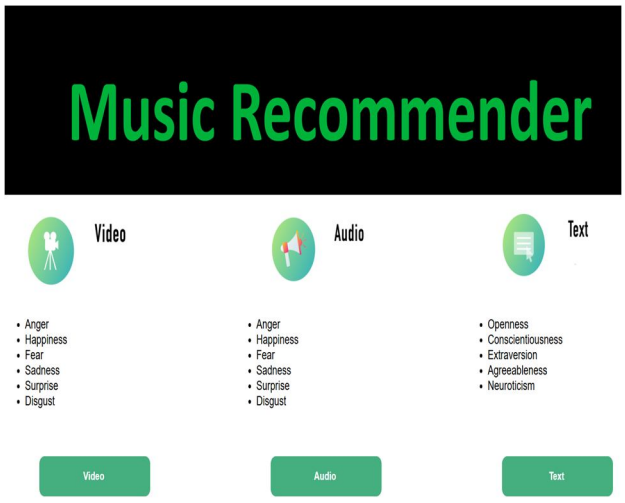
It can be said that a more robust system can be developed through the application of a multimodal method, taking into consideration the outputs of the algorithms such as SVM, CNN, and NLP.

C. SVM

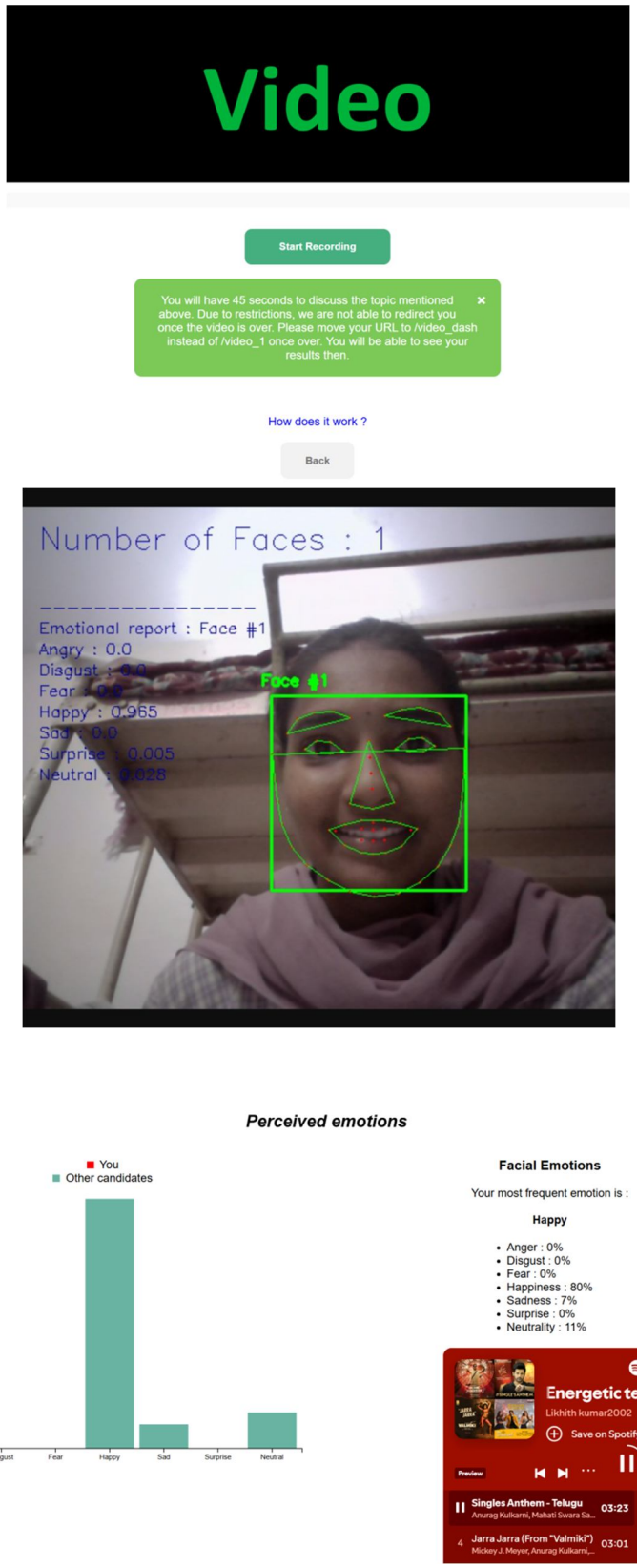
For text-based emotion detection, Support Vector Machine (SVM) classifiers were employed. At first, the SVM classifier model was designed without performing any process of dimensionality reduction, where several kinds of kernel functions including linear, polynomial, and Radial Basis Function (RBF) were experimented. From the first experimentation result, the performance of classifying emotion using this technique could be considered as moderate, and the RBF kernel showed better result compared to other kernels.

Several improvements had been made by adding some more processes such as feature selection and reduction. By applying Chi-square test, it is possible to eliminate less important features from the model, leading to the elimination of unnecessary words from the text-based dataset. Then, PCA was conducted, where three different variances including 90%, 95%, and 98% were tested, which represent the numbers of dimensions. After applying all these processes, an improvement in performance occurred; however, the difference was not too much. Polynomial and RBF kernels showed improvement at a higher level of feature dimension, particularly for 98%. As before, the best kernel remained RBF.

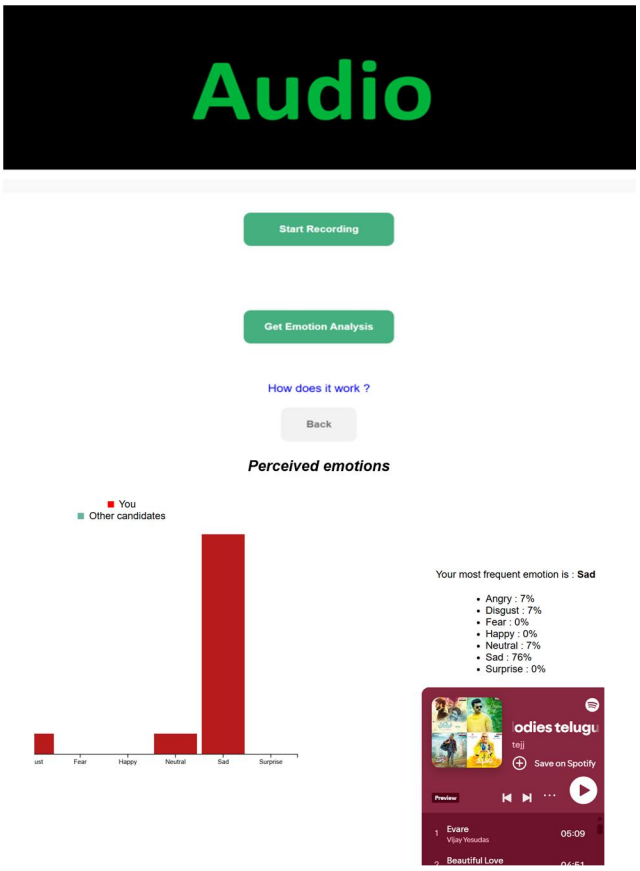
V. RESULTS



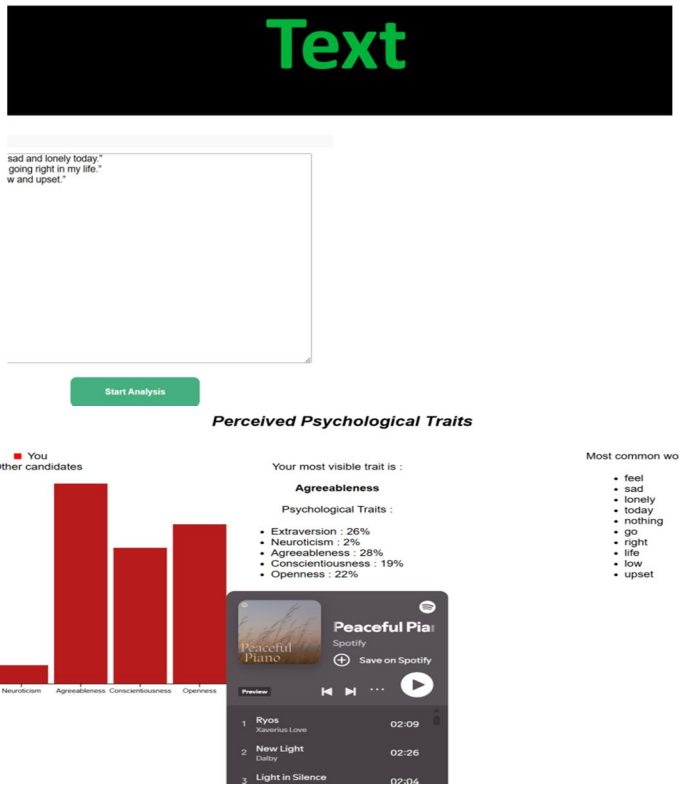
Home page



Video result for emotion



Audio result for emotion



Text result for emotion

VI. CONCLUSION

This paper presents a multimodal emotion-based music recommendation system that can recognize emotions of users and recommend appropriate music through analysis of their facial expressions, voice patterns, and text inputs. This method increases the accuracy and effectiveness of emotion recognition since it relies on multiple sources rather than just one source.

This multimodal emotion-based music recommendation system has been developed as a web application with Flask. It enables real-time recognition of the emotion of a user and provides recommendations for appropriate music instantly. Depending on the emotion recognized, the system recommends personalized music tracks for the user based on their emotions. The simplicity and efficiency of this system make it user-friendly and effective.

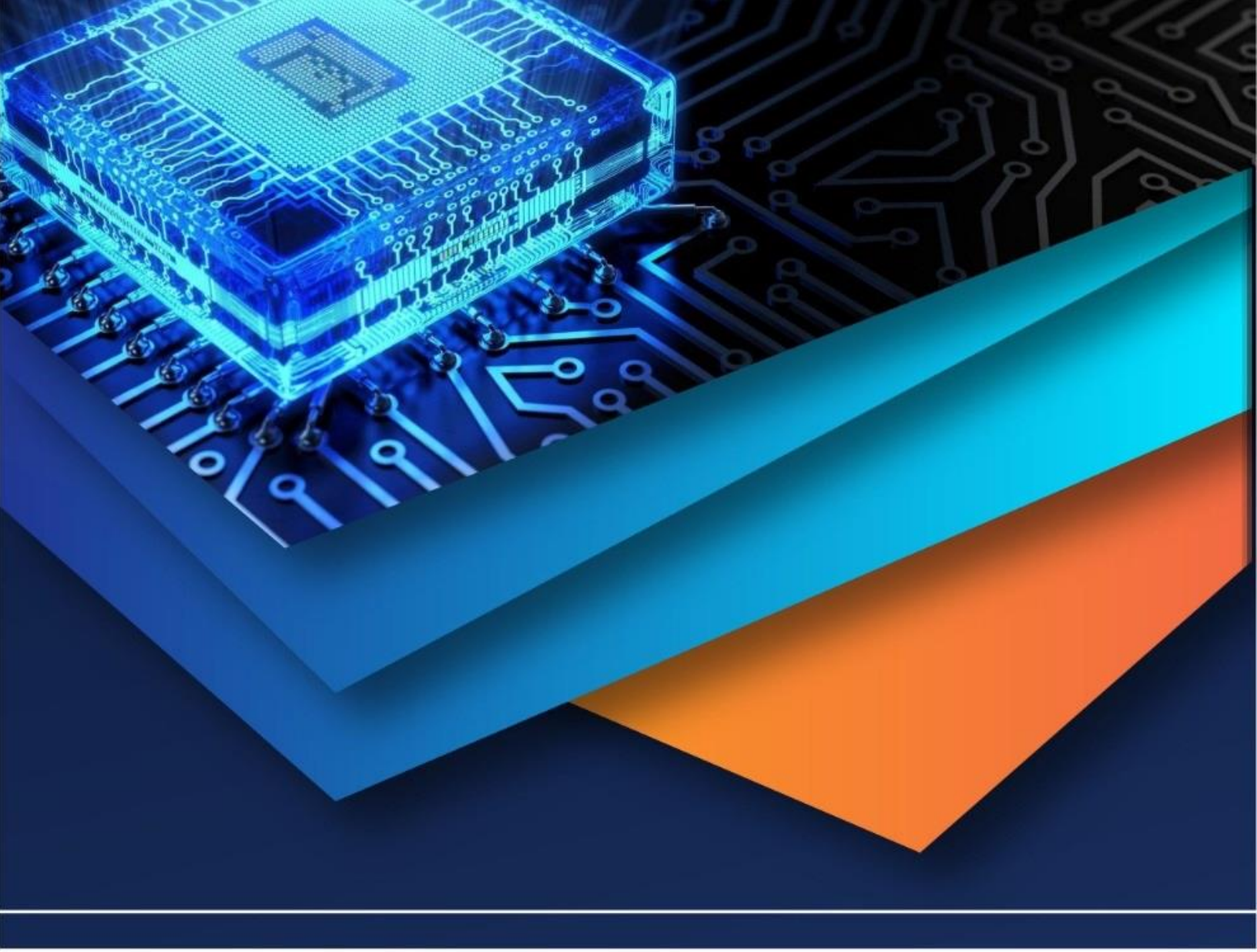
VII. FUTURE SCOPE

Music recommendation based on individual users' preferences and behavior constitutes a significant and emerging domain within intelligent systems, as AI learns and evolves through experience based on users' data. Future developments of the system will include the use of machine learning algorithms to develop user profiles, which will enable the intelligent system to understand user behavior, emotional states, and responses in order to provide personalized music recommendations.

Integration of the music streaming platform will contribute to the enhancement of the system through the provision of song recommendations in real-time from vast libraries of music collections. Further developments of the system will include the use of deep learning algorithms to improve the precision of the system's ability to detect emotions. Support for multiple languages will also be integrated into the system.

REFERENCES

[1] A. Abdul, J. Chen, H. Y. Liao, and S. H. Chang, “Emotion-Aware Personalized Music Recommendation System Using Convolutional Neural Networks,” *Applied Sciences*, vol. 8, no. 7, 2018.
[2] W. Deng, “Application of Multimodal Emotion Recognition Technology in Recommendation Systems,” *Highlights in Science, Engineering and Technology*, 2025.
[3] Y. Wu, Q. Mi, and T. Gao, “A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions,” *Biomimetics*, 2025.
[4] V. S. G. S. Phaneendra and K. Ragavan, “Emotion-Based Music Recommendation System Integrating Facial Expression Recognition and Lyrics Sentiment Analysis,” *IEEE Access*, 2025.
[5] D. Ayata, Y. Yaslan, and M. E. Kamasak, “Emotion Recognition from Multimodal Physiological Signals for Emotion-Aware Systems,” *Journal of Medical and Biological Engineering*, vol. 40, pp. 149–157, 2020.
[6] S. Wang, “Music Emotion Recognition and Modeling Based on Multimodal Signal Fusion,” *Traitement du Signal*, 2025.
[7] M. Athavle, D. Mudale, U. Shrivastav, and M. Gupta, “Music Recommendation Based on Face Emotion Recognition,” *Journal of Informatics Electrical and Electronics Engineering*, 2021.
[8] “IoT-Based Approach to Multimodal Music Emotion Recognition,” *Alexandria Engineering Journal*, 2024.
[9] Y. Wu, Q. Mi, and T. Gao, “A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions,” *Biomimetics*, 2025.
[10] R. Pillalamarri and U. Shanmugam, “A Review on EEG-Based Multimodal Learning for Emotion Recognition,” *Artificial Intelligence Review*, 2025.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY