Music Recommendation System Using Advanced CNN And Face Expression Recognition

Prof. Renuka Kajale[1], Ayushi Kale[2], Asawari Khairnar[3],Vaishnavi Mavale[4]

Computer Engineering Department[1,2,3,4]

Nutan Maharashtra Institute of Engineering and Technology, Pune[1,2,3,4]

Abstract—In the ever-evolving landscape of music consumption, the development of intelligent recommendation systems has become imperative to enhance user experience. This research paper introduces a pioneering approach to music recommendation by integrating advanced Convolutional Neural Networks (CNN) with face expression recognition. The proposed system aims to personalize music suggestions by analyzing users' facial expressions, extracting emotional cues, and aligning them with the corresponding auditory preferences. The convolutional neural network component of the system is designed to learn intricate patterns and features from music spectrograms, capturing both the audio content and underlying emotional nuances. Simultaneously, facial expression recognition technology is employed to discern users' emotional states during music listening sessions. By fusing these two modalities, our system strives to create a holistic understanding of users' preferences, considering both explicit musical features and implicit emotional responses. To achieve this integration, we leverage machine learning architectures for music analysis and facial expression recognition. A wide variety of facial expressions and musical genres are included in the dataset that the model is trained on. Additionally, the research explores the challenges and opportunities associated with combining these distinct modalities, such as data preprocessing, feature extraction, and model fusion. This research contributes to the ongoing discourse on the fusion of multimodal technologies for more nuanced and personalized recommendation systems, paving the way for innovative applications in the intersection of music and affective computing.

***Keywords*—**Music, CNN, Expression, Feature Extraction

I. INTRODUCTION

In the era of digital music consumption, the demand for personalized and intelligent music recommendation systems has surged. As users are inundated with an overwhelming abundance of musical content, the need for algorithms that can understand individual preferences and emotions becomes paramount. This research paper delves into the innovative integration of advanced Convolutional Neural Networks (CNNs) and facial expression recognition techniques to create a novel paradigm in music recommendation systems. Traditional music recommendation systems primarily rely on user listening history, collaborative filtering, and content-based approaches. However, these methods often fall short in capturing the dynamic and nuanced nature of human emotions, which play a crucial role in shaping musical preferences. [1] It is a primary study direction in the field of computer vision and plays a key role in intelligent data collecting and processing based on images

To address this limitation, our research explores the fusion of cutting-edge technologies, namely advanced CNN architectures and facial expression recognition, to enrich the recommendation process. [2] Facial expressions indicate the variations of facial appearance in reaction to a person's inner emotional states, social communications, or goals.[9] The user's local music collection is initially clustered depending on the emotion conveyed by the song, often known as its mood. [10] The Emotion Module detects the user's mood (90.23%) accurately by using deep learning algorithms and a snapshot of the user's face as input.[14] We employ facial expressions to present a recommender system for emotion recognition that can recognize user moods and provide a list of acceptable music. Convolutional Neural Networks, known for their proficiency in image analysis, are adapted to extract intricate patterns and features from audio spectrograms. By treating music as a visual entity, we aim to capture both low-level and high-level features that contribute to the emotional and perceptual aspects of music.[3] Deep learning consists of supervised learning, semi-supervised learning, and unsupervised learning.[5] The various expression properties demonstrate that distinct expressions have similar and different characteristics, and even the same type of expression might have several expressions.

Facial expression analysis is integrated into the system, allowing it to adjust recommendations in real-time according to the user's emotional state. This dynamic and responsive aspect of the proposed system ensures a more immersive and personalized music recommendation experience. Through the synergistic combination of advanced CNNs and facial expression recognition, our research aims to contribute to the evolution of music recommendation systems. The proposed model not only considers the audio features of the music but also takes into account the user's emotional responses, creating a holistic and tailored recommendation experience. [8] It is a primary study direction in the field of computer vision and plays a key role in intelligent data collecting and processing based on images

This paper will delve into the technical details of the architecture, the methodology employed, and the experimental results, providing insights into the potential of this hybrid approach to revolutionize the landscape of personalized music recommendations. [16] Machine learning, a subset of artificial intelligence, involves techniques such as data analysis and pattern identification

This system will bridge a gap between traditional music browsing system and the users' needs. Using this system the songs can be categorized by various emotions like sad, happy, neutral, angry.

II. Methodology

Procedures for creating the system design Training datasets and test photos are taken into consideration throughout system design, and the following steps are taken to achieve the intended outcomes for each. [6] The two ways will be introduced and summarized, respectively. [13] Every task had a Kaggle host. The test set is the input provided for recognition purposes, while the training set is the raw data with a lot of data saved in it. The entire system is created in five stages:

A. *Image Acquisition*

Getting the image from the source is the first step in every image processing procedure. These pictures can be obtained via a camera or through freely accessible online standard databases.

B. *Pre-processing*

Pre-processing is mostly used to remove unnecessary information from captured images and adjust certain values so that the value stays constant. [12] However, a user's music selection is not just based on their prior preferences or song content. The photos are scaled to 256*256 pixels and transformed from RGB to grayscale during the pre-processing stage. All photographs that are taken into consideration are in the.jpg format; files in any other format won't be processed further. The mouth, nose, and eyes are regarded as the zone of interest during pre-processing.

C. *Facial Feature Extraction*

Feature extraction comes next after pre-processing. During the training and testing phases, the extracted facial features are saved as usable information in the form of vectors. [7] Finding the position of a human face in an image is what we described as the face detection (FD) key point. The mouth, forehead, eyes, skin tone, cheek, chin, example eyebrows, nose, and facial wrinkles are all regarded to be part of the face. Since they portray the most endearing expressions, the eyes, nose, mouth, and forehead are taken into consideration for feature extraction in this work.

It is simple to tell whether someone is astonished or afraid by the lines on their forehead or their parted lips. However, one's complexion can never be accurately captured. The PCA approach is used to extract the face features. [4] In theory, CNN can gain a more representative and abstract concept of a picture if it penetrates deeper.

D. *Expression Recognition*

The Euclidean distance classifier is used to identify and categorize an individual's expressions. It finds the training data set's closest match for the test data, providing a better match for the currently identified expression. Essentially, the Euclidean distance is the separation between two points and is calculated as “(3.1)”. It is computed using the training dataset's eigenface mean.

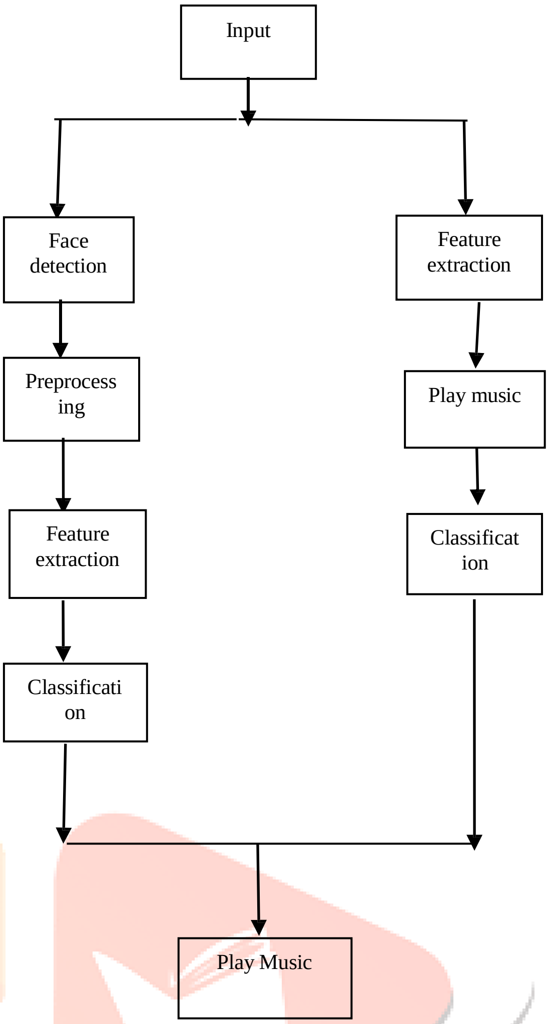


Fig.1. System Architecture of music recommendation system

III. ALGORITHM

There are multiple processes involved in developing a face detection music recommendation system employing Convolutional Neural Networks (CNN) and the Haar Cascade technique.[11] The suggested system identifies emotions; if the subject has a negative emotion, a special playlist will be offered that contains the most suitable forms of music to boost his mood.

A. Gathering of Data

Compile a dataset of face-containing photos and the matching attributes or music preferences.

B. Prior to processing

Adjust each image to a standard size that the CNN can process. Set the values of the pixels to a range of 0 to 1. Based on traits or musical tastes, label the pictures.

C. Face Recognition

Use the Haar Cascade technique to identify faces in the pictures. Take out the faces, or regions of interest, from every picture.

D. Feature Extraction

To extract features from the face pictures, use the N model. Using the gathered dataset, fine-tune a pre-trained CNN model to discover discriminative characteristics associated with musical tastes.

E. Suggestion Algorithm

To compare the feature vectors of faces and musical preferences, use similarity metrics such as cosine similarity and Euclidean distance.

F. Evaluation

Use measures like precision, recall, and F1-score to evaluate their commendation system. Verify the model's generalizability using methods like cross-validation. Below is an explanation of the architecture*:*

A. Input Layer

The input layer receives images containing faces captured from a camera or uploaded by the user.

B. Face Detection Module:

This module utilizes the Haar Cascade algorithm to detect faces within the input images. The Haar Cascade algorithm identifies regions of interest (faces) based on predefined features and patterns[20].

C. Preprocessing Module

Once faces are detected, the preprocessing module performs tasks such as cropping and resizing to prepare the face images for further processing. Normalization may also be applied to ensure consistency in pixel values across images.

D. Feature Extraction using CNN

Using CNN for feature extraction is a common practice in machine learning. The preprocessed face images are fed into a CNN model specifically trained for feature extraction from faces. The CNN extracts high-level features from the face images, capturing important patterns and characteristics. Convolutional layers, pooling layers, flattening and fully connected layers helps to extract the features of given image.

E. Music Preference Representation Module

Simultaneously, the system gathers music preference data associated with each user. This data could include past listening history, genre preferences, mood preferences, etc.

F. Similarity Calculation Module

The extracted features from the face images and the music preference data are used to compute similarities between users. It is possible to measure the degree of similarity between feature vectors that reflect users' faces and their musical tastes using methods like cosine similarity or Euclidean distance.

G. Recommendation Algorithm

Based on the computed similarities, a recommendation algorithm suggests music items that are preferred by users with similar facial features and music preferences. To provide individualized recommendations, this system may use content-based filtering, collaborative filtering, or hybrid techniques.

H. Output Layer

The output layer presents the recommended music items to the user. Overall, the architecture integrates face detection using the Haar Cascade algorithm, feature extraction with CNNs, and music preference analysis to provide personalized music recommendations [17].

IV.MATHEMATICAL MODEL

Input (I): The input to the system consists of images capturing facial expressions. These images serve as the primary data source for the recommendation system.

Procedure (P): The procedure describes the operations performed by the system using the input data. Given the input image, the system processes it to extract relevant features, likely using facial recognition techniques. The system then analyzes these features to infer the user's emotional state or preference regarding music. Based on the analysis, the system predicts suitable music recommendations.

Output (O): The output of the system is the recommended music selection. The system plays music according to the facial expression detected and the corresponding inference made about the user's preference or mood. For example, if the user is deemed to be happy based on their facial expression, the system might recommend upbeat or cheerful music. Conversely, if the user appears sad, the system may suggest calming or soothing music. Overall, the model outlines a straightforward process where facial expressions serve as the input, undergo analysis to determine the user's mood or preference, and generate music recommendations accordingly.

Now,
Let S represent the entire system S = {I, P, O}
I-input
P-method
O-output
Input (I)
I ={Image}
Where,
Dataset-;
Process (P),
P = {I, Utilizing I System to carry out activities and
compute the forecast}
output(O)-
O= {Play music based on facial expression in

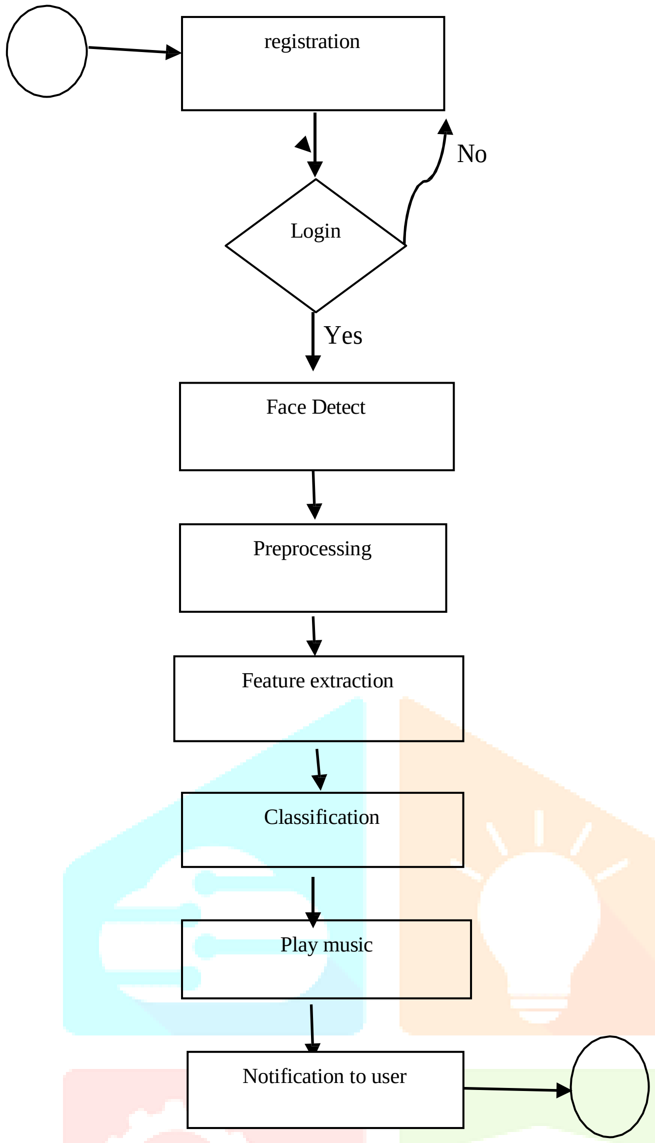


Fig.2. Activity diagram of music recommendation system

Acitvity Diagram:

The activity diagram for the music recommendation system using facial detection with CNN and Haar cascade algorithm**:**

A.Start

This represents the system's starting state. It indicates the beginning of the process.

B.Capture Facial Image:

The system initiates the process by capturing a facial image of the user using a camera or any other image-capturing device available in the system environment**.**

C.Preprocessing

The captured facial image may undergo preprocessing to ensure its quality and suitability for facial detection. Preprocessing steps may include resizing the image to a standard size, normalization to adjust lighting conditions or color balance, or other enhancement techniques.

D.Facial Detection

This step involves using a Convolutional Neural Network (CNN) model and the Haar cascade algorithm to detect faces within the preprocessed image [20]. The CNN model is trained to recognize facial features and patterns, while the Haar cascade algorithm detects specific patterns like the presence of eyes, nose, and mouth.

E.Check for Detected Faces:

The system checks whether any faces are detected in the captured image. If no faces are detected, the system may prompt the user to adjust their position or lighting conditions and repeat the capturing process.

F.Feature Extraction:

After a face is successfully identified, the algorithm collects pertinent features from the face picture. These features could include facial landmarks, such as the position of eyes, nose, mouth, as well as overall facial expressions, gender, or age estimation. [19] Identification image Based on the extracted features, the system identifies or authenticates the user. This identification process may involve comparing the extracted features with a database of known users or profiles to find a match.

G.Retrieve Music Preferences:

Once the user is identified, the system retrieves the user's music preferences from their profile stored in the system's database. These preferences may include genres, artists, albums, or specific tracks that the user has previously shown interest in or listened to.

H.Generate Recommendations:

Using the retrieved music preferences and potentially other factors such as the user's current mood inferred from facial expressions, the system generates personalized music recommendations. This could involve algorithms such as collaborative filtering, content-based filtering, or hybrid approaches to recommend music tracks or playlists tailored to the user's tastes and preferences.

I.Display Recommendations:

The recommended music tracks or playlists are presented to the user through the system's interface, which could be a web application, mobile app, or any other user interface**.**

J.End

This indicates the completion of the process. The system has successfully captured the user's facial image, detected their face, identified the user, retrieved their music preferences, generated personalized recommendations, and presented them to the user.

IV. Result

The results of a music recommendation system using face detection with CNN and the Haar Cascade algorithm would depend on various factors including the quality of the dataset, effectiveness of the face detection and CNN models, similarity calculation method, and recommendation algorithm. Here are some potential results

A. Accuracy of Face Detection:

The accuracy of the Haar Cascade algorithm in detecting faces within images affects the system's performance. Higher accuracy leads to more reliable facial feature extraction.

B. Feature Extraction:

The CNN model's effectiveness in extracting meaningful features from facial images influences the system's ability to capture relevant information about users' emotional states or preferences.

C. Similarity Calculation

The similarity calculation between facial features and music preferences determines the relevance of recommended music items. More accurate similarity measures result in better personalized recommendations.

D. Recommendation Quality

The quality of music recommendations depends on how well the system can match users' facial expressions with appropriate music genres, moods, or preferences. [18] Users may provide feedback on the relevance and accuracy of the recommendations, which can be used to improve the system over time.

*E. User Satisfaction***:**

Ultimately, the success of the system is measured by user satisfaction. If users find the recommended music aligns well with their preferences or mood, they are more likely to engage with the system positively.

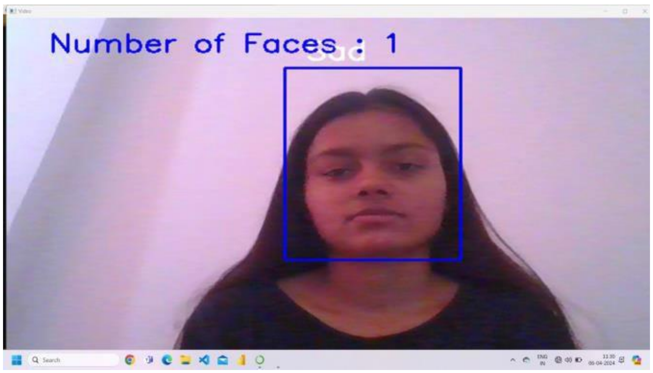
F. Evaluation Metrics:

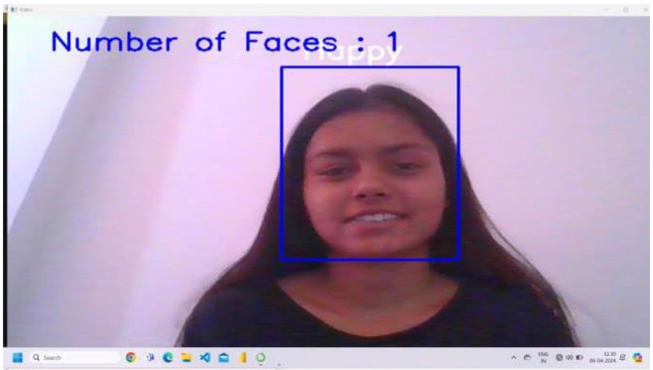
The effectiveness of the system may be objectively assessed using metrics like accuracy, precision, recall, and user engagement.

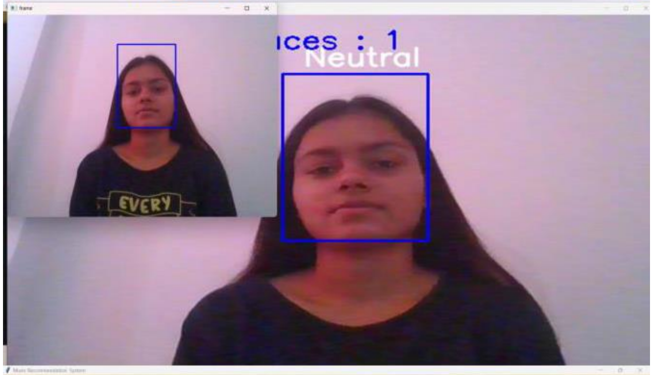
F. Scalability and Efficiency

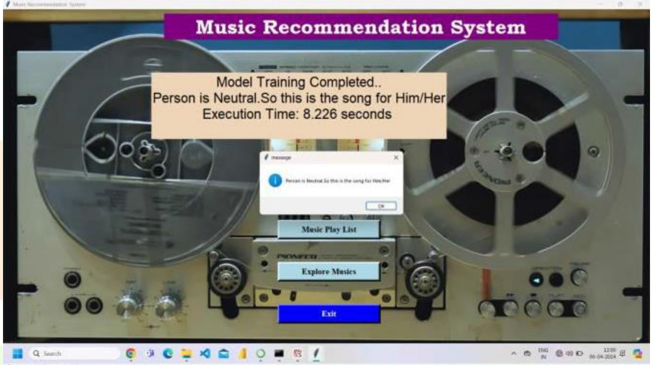
The system's ability to handle a large volume of users and adapt to real-time changes in preferences or facial expressions is also crucial for its effectiveness.[13] The suggested system identifies emotions; if the subject has a negative emotion, a special playlist will be offered that contains the most suitable forms of music to boost his mood.

G. Output of System

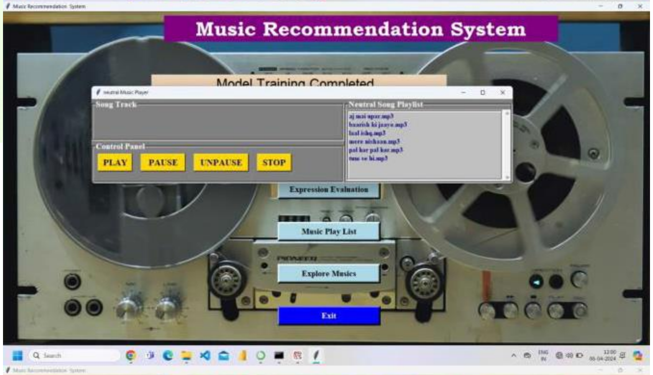






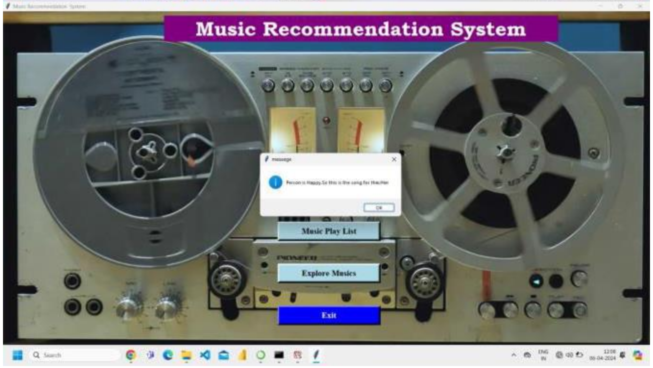


Model Training Completed..
Person is Neutral.So this is the song for Him/Her
Execution Time: 8.226 seconds



Neutral Song Playlist

aj mai upar.mp3
baarish ki jaaye.mp3
laal ishq.mp3
mere nishaan.mp3
pal har pal har.mp3
tum se hi.mp3



Person is Happy.So this is the song for Him/Her

V. Future Scope

In the future personalization and context-aware music recommendation systems can become even more personalized by taking into account the user's current context, such as their location, activity, mood, or time of day. Combining several data sources, including audio analysis, lyrics, and user-generated content (such comments and reviews), can improve suggestion and offer a more comprehensive understanding of music. AI-generated music is a growing field. Recommender systems could incorporate AI-generated music alongside human-created music, opening up new creative possibilities and helping users discover unique compositions. Moving beyond genre-based recommendations, systems can identify and recommend music based on various musical characteristics, including tempo, instrumentation, lyrical content, and emotional qualities. Expanding the use of implicit feedback such as user skips, pauses, and repeat listens, can provide richer insights into user preferences.

VI. Conclusion

The suggested work offers a facial expression recognition system that can both categorize and play music based on the expression that is identified. It extracts features using the CNN technique, and then classes these expressions using the Euclidean distance classifier. Real photos, or user-dependent photographs, are taken with the built-in camera in this work.

VII References

[1] Yao L S, Xu G M, Zhap F. Facial Expression Recognition Based on CNN Local Feature Fusion[J]. Laser and Optoelectronics Progress, 2020, 57(03): 032501.

[2] Li Huihui. Research on facial expression recognition based on cognitive machine learning [D]. Guangzhou: South China University of Technology, 2019.

[3] Li Yong, Lin Xiaozhu, Jiang Mengying. Facial expression recognition based on cross-connection LeNet-5 network [J]. Journal of Automation, 2018,44 (1): 176-182.

[4] Xie S, Hu H. Facial expression recognition with FRR-CNN [J]. Electronics Letters, 2017, 53 (4): 235-237.

[5] Li Yong, Lin Xiaozhu, Jiang Mengying. Facial expression recognition based on cross- connection LeNet-5 network [J]. Journal of Automation, 2018,44(1): 176-182.

[6] Yao L S, Xu G M, Zhap F. Facial Expression Recognition Based on CNN Local Feature Fusion[J]. Laser and Optoelectronics Progress, 2020, 57(03):032501.

[7] Xie S, Hu H. Facial expression recognition with FRR-CNN [J]. Electronics Letters, 2017, 53 (4): 235-237.

[8] ou Jiancheng, Deng Hao. An automatic facial expression recognition method based on convolutional neural network [J]. Journal of North China Universityof Technology, 2019,31 (5): 51-56

[9] Ramya Ramanathan, Radha Kumaran, Ram Rohan R, Rajat Gupta, and Vishalakshi Prabhu, an intelligent music player based on emo-tion recognition, 2nd IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions 2017.

[10] Shlok Gilda, Husain Zafar, Chintan Soni, Kshitija Waghurdekar, Smart music player integrating facial emotion recognition and music mood recommendation, Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India, (IEEE),2017

[11] Ahlam Alrihail, Alaa Alsaedi, Kholood Albalawi, Liyakathunisa Syed, Music recommender system for users based on emotion detection through facial features, Department of Computer Science Taibah University, (DeSE), 2019.

[12] Deger Ayata, Yusuf Yaslan, and Mustafa E. Kamasak, Emotion-based music recommendation system using wearable physiological M. Athavle et al. ISSN (Online) : 2582-7006 International Conference on Artificial Intelligence (ICAI-2021) 11 Journal of Informatics Electrical and Electronics Engineering (JIEEE) A2Z Journals sensors, IEEE transactions on consumer electronics, vol. 14, no. 8, May

[13] Ahlam Alrihail, Alaa Alsaedi, Kholood Albalawi, Liyakathunisa Syed, Music recommender system for users based on emotion detection through facial features, Department of Computer Science Taibah University, (DeSE), 2019

[14] Preema J.S, Rajashree, Sahana M, Savitri H, Review on facial expression-based music player, International Journal of Engineering Re-search & Technology (IJERT), ISSN-2278-0181, Volume 6, Issue 15, 2018.

[15] AYUSH Guidel, Birat Sapkota, Krishna Sapkota, Music recommendation by facial analysis, February 17, 2020.

[16] CH. sadhvika, Gutta.Abigna, P. Srinivas reddy, Emotion-based music recommendation system, Sreenidhi Institute of Science and Technology, Yamnampet, Hyderabad; International Journal of Emerging Technologies and Innovative Research (JETIR) Volume 7, Is-sue 4, April 2020

[17] Vincent Tabora, Face detection using OpenCV with Haar Cascade Classifiers, Becominghuman.ai,2019.

[18] Frans Norden and Filip von Reis Marlevi, A Comparative Analysis of Machine Learning Algorithms

[19] Ahmed Hamdy AlDeeb, Emotion- Based Music Player Emotion Detection from Live Camera, ResearchGate, June 2019.

[20] K. ShanthaShalini et al.“Facial emotion based music recommendation system using computer vision and machine learning techniques”Turkish J. Comput. Math.Edu.(2021)