iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: V Month of publication: May 2025

DOI: https://doi.org/10.22214/ijraset.2025.70901

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Music Recommendation System based on Facial Emotion Detection using Spotify API

Prasanjit Singh[1], Machanuru Suresh Babu[2]
Deptof CSE, Narsimha Reddy Engineering College

Abstract: In the world of entertainment, music holds considerable importance, especially for those who find joy in rhythmic experiences. Despite the abundance of streaming platforms that enable access to favorite songs, they often fall short in capturing the intricate emotional nuances of users. This research recognizes a spectrum of emotions, including fear, happiness, sadness, anger, and neutrality. Its goal is to enrich the user experience by developing a recommendation system that proposes songs based on the user's current emotionalstate.Theemotion-driven recommendationengine has seamlessly integrated into Spotify, a well-known music streaming service, providing users with a smooth and individualizedjourneyinexploringmusic.Thesystemaimsto simplify the user experience by eliminating the necessity for manualsongsearchesand,instead,intuitivelysuggeststracks thatresonatewiththeuser'semotions.TheSpotifyAPIserves as a crucial tool for accessing curated playlists, enabling the retrieval of desired music from thoughtfully organized collections centered around specific themes or titles.
Keywords: Music,SpotifyAPI,UserExperience,Emotion- based, Time-saving.

I. INTRODUCTION

Everybody is surrounded by music in their daily lives. People may now access and enjoy a vast array of music thanks to the emergence of streaming services. Music varies based on location and cultural norms. People also differ in likes, dislikes, and choices. This people's musical preferences also differs. Thus, to determine what kind of music someone could suchashearingandcreatingasystemofrecommendations to
assist in reaching a variety of acts, tunes, and genres, both new and old to people. Determining the connections between different music is a laborious undertaking. It's possible that one song, a specific person enjoys or is his favorite genre is disliked by another user.

The system will determine users' musical preferences by analyzing their interactions with the Spotify app, specifically utilizing Spotify's web API to query informationaboutusers'recenttoptracks.Additionally,the work explores the potential benefits of incorporating the user's emotional context into a music recommendation system (RS) to enhance accuracy and provide superior recommendations compared to existing models. While someresearchprojectshavesuggesteddirectemotion-based playlists as recommendations, prevalent recommendation systemslikeSpotifyemployahybridmodelthatcombines content-basedandcollaborativefiltering.Objectiveofthepaper is to introduce a innovative approach where songs, irrespectiveof theirageor popularity, willberecommended with equalimportance,takingintoconsideration theuser's overall preferences.

II. RELATED WORKS

Work examines the following papers as part of literature evaluation, and here is a quick summary of the work that was done:

The paper compares the performance of domain-specific networks and image classification networks using two datasets: the widely-used GTZAN benchmarking dataset and a newly created, much larger dataset. Our findings indicate that the image classification network requires significantlyfewerresourcesandoutperformsthedomain- specific network in our test conditions. This suggests that usingtheimageclassificationnetworkeliminatestheneed for expert effort in designing specialized networks.

Themusicindustryhasexperiencedasignificantincreasein newchannelsforbrowsinganddistributingmusic,butthis growth comes with challenges. As the volume of data rapidly expands, manual curation becomes increasingly difficult. Audio files contain numerous features that could streamline this process, though the best methods for utilizing these features for various tasks are not always clear. This thesis evaluates two deep learning models, convolutionalneuralnetworks(CNNs)andlongshort-term memorynetworks(LSTMs),formusicgenreclassification using mel-frequency cepstral coefficients (MFCCs). The goalofpaper[2]istomaximizetheutilityofaudiodatafor futureapplications.ThemodelsweretestedontheGTZAN and FMA datasets, with the CNN achieving prediction accuracies of 56.0% and 50.5%, respectively. This performance surpassed that of the LSTM model, which achieved prediction accuracies of 42.0% and 33.5%.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Automaticemotionrecognitionbasedonfacialexpressions is a fascinating research area with applications in safety, health, and human-machine interfaces. Researchersaim to develop techniques to interpret and encode facial expressions, and to extract these features for improved computer-based predictions. With the significant advancements in deep learning, various architectures have been leveraged to enhanceperformance.Thispaper[3]aims to review recent works on automatic facial emotion recognition (FER) using deep learning. We focus on the contributions,architectures,anddatabasesused,comparing theproposedmethodsandtheirresults. Thegoalofthispaper is to guide and inform researchers by reviewing recent developments and offering insights for further advancements in the field.

Soleymani, Aljanaki, Wiering, Veltkamp, et al. [4] developedarecommendationsystem(RS)thatincorporated psychological aspects of users' musical preferences, supplementing the conventional genre-based suggestions. The system employed regression analysis to identify key qualitiesusingfeaturesfromauditorymodulationanalysis. Unlike other recommendation systems relying on user- based or genre-based approaches, this unique research demonstrated superior performance, as indicated by lower root-mean-square error values. As a result, work opted to select parameters distinct from genre for recommendation system.

To simplify the complex process of explicit feature extraction in traditional facial expression recognition, a method based on a convolutional neural network (CNN) andimageedgedetectionisproposedinpaper[5].Initially, facial expression images are normalized, and the edges of each image layer are extracted during the convolution process. The extracted edge information is superimposed onto each feature image to maintain the edge structure of the texture image. Dimensionality reduction of the extracted implicit features is then performed using the maximum pooling method. Finally, a Softmax classifier is used to classify and recognize the expressions in the test sample images. To test the robustness of this method for facial expression recognition in complex backgrounds, a simulation experiment was conducted by combining the Fer-2013facialexpressiondatabasewiththeLFWdataset. The experimental results demonstrate that the proposed algorithm achieves an average recognition rate of 88.56% with fewer iterations and has a training speed approximately1.5times fasterthan thatofthecomparison algorithm.

Paper[6] utilizes deep learning to categorize human facial expressions, enabling the filtering and mapping of corresponding emojisor avatars.Thegoalisnotto solvea real-world problem but to make communication more vibrant. Emojify is software designed to streamline the creation of emojis and avatars.

A hybrid technique was utilised by Akrati et al. [7]. The user's context—that is, whether they are working or dancing—isanother factor that the algorithm considers. A playlistincludingthetop-Nsongswassuggested,andfour distinctrecommendationalgorithmswereapplied.Thereis nodetaileddescriptionofthestrategies,anda96%accurate supervised kNN model is employed.

III. PROPOSED METHOD

Therearethreemodules:EmotionDetectionModule,Face Detection Module and Song Recommendation Module.

TheFaceDetectionmodulelooksforHaar-likefeaturesin the webcam input to identify faces using the Viola-Jones technique.ThelibraryusedforthisiscalledOpenCV.Once the face has been identified, the CNN facial expression recognition model from the Emotion Recognition Module is used to categorise the facial input into an emotion category.

The music recommendation engine takes the identified emotioncategoryasaninput.Usersarepromptedtologin to their Spotify account, following which the Spotify web API is utilized to fetch information about their latest favorite songs. An additional set of ten songs is recommendedfromthepoolofsongsbasedontheemotion deduced from the facial expression.

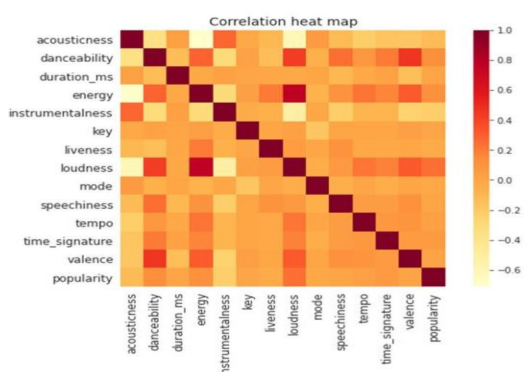

Fig.1:CorrelationHeatMap

In above Fig. 1,the heat map shows the correlation between 13 audio features ofa song. Each feature is listed ontheleftandrightsidesoftheheatmap,andthestrength of the correlation between each pair of features is represented by the color intensity in the corresponding square. The heat map reveals several interesting patterns:

1) *Acousticness and valence* show a positive correlation, indicating that acoustic songs generally evoke more positive emotions. This aligns with the common perceptionthatacousticmusicisoftenlinkedtoasense of tranquility and serenity.
2) *Danceability and energy* exhibit a positive correlation, suggesting that danceable songs typically possess high energylevels.Thisfindingisinlinewiththeexpectation that dance music is crafted to inspire movement and excitement.
3) *Speechiness and instrumentalness*display a negative correlation, indicating that content with a significant amount of spoken word (like audiobooks or podcasts) tends to have fewer accompanying instruments. This is logical,consideringspoken wordcontentusuallydoesn't require a musical backdrop.
4) *Loudness and valence* show a slight negative correlation, implying that louder songs tend to have lowerpositiveemotions.Thiscouldbeattributedtothe association of loud music with emotions like anger or aggression.

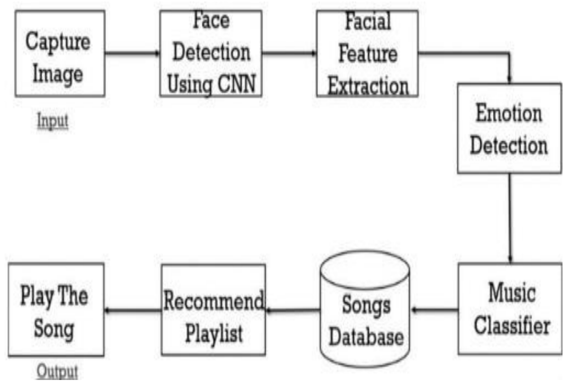

Fig.2:FlowDiagramoftheSystem

The above Fig .2 depicts a system that captures the user's face using a camera. A CNN detects facial features like eyes and mouth. Information like eye distance and mouth shape is extracted.A machine learning model identifies emotions (e.g., happiness, sadness) based on extracted features.The system suggests songs or playlists matching the detected emotion (e.g., upbeat pop for happiness, melancholic ballad for sadness).The recommended music is played through speakers or headphones, offering a personalized listening experience.

IV. MODULE DESCRIPTION

A. *EmotionDetectionModule*

1) *Get the Data:* This approach functions on the FER2013dataset,consistingofimageswitharesolutionof 48*48 pixels. Each image is associated with a label indicating one of six emotions: anger, disgust, fear, happiness, sadness, surprise, or neutrality. The dataset is organized into three columns: emotion, pixels, and usage. The two primary applications of the dataset are for testing and training purposes.

Fig.3:SevenEmotions

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

2) *Prepare the Data:* To make sure the data is in the right format, preprocessing is done on it. Two subsets of the dataset have been identified: X_train, X_test and y_train, y_test. The former is for strings of pixels, and the latter is for emotion labels (integer encoded labels). There are seven emotion classes in use, according to the value num_classes. The following parameters are added to the data to create a 4D tensor for training: row_num, width, height, and channel.
3) *Build and train the model:* Conv2D layer, Batch- Normalization, Max-Pooling2D, Dropout, and Flatten are usedtoconstructblocksfortheCNNmodel.Theseblocks are then piled one on top of the other. The final output is produced using Dense Layer. Adam optimises the model during compilation. 0.001 is the learning rate maintained. On an Intel i3 Windows PC, training takes 30 minutes for one epoch.

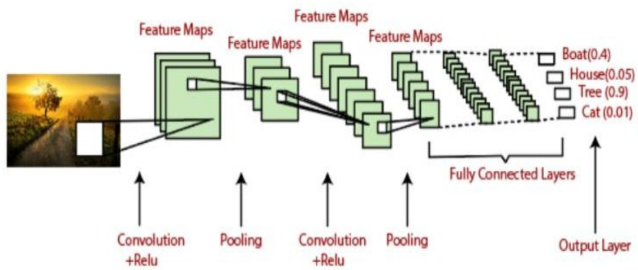

Fig.4:LayerofConvolutionNeuralNetwork

B. FaceDetectionModule:
1) *Load the model:* Theinitialstep involvesimporting the weights and the trained model architecture. Once the Haar-cascade method is employed to identify the position of the face, the faces are then cropped.
2) *Data Preprocessing:* The emotion label is obtained through the variable "emotion_prediction," and the OpenCV Python Library for image processing is responsible for reading it. To rescale the test image, it is dividedby255.Thecoordinatesoftheidentifiedfaceinthe input are represented as (x, y, w, h) within 4D tensors formed from 3D matrices.

C. SongRecommendationModule:
Users will initially receive their credentials (client ID andclientsecretkey)throughtheSpotifyonlinedashboard. Uponobtainingcredentials,theuserwillbepresentedwith theirmost-playedtracks.Subsequently,thesetrackswillbe fed into therecommenderengine, whichwillanalyzetheir attributes such as dance, tempo, mode, valence, etc.
To identify the most effective features for the recommendation engine, a heat map will be utilized as a visualizationtechnique(refertoFig1).Inthisscenario,the analysis revealed four features that demonstrated optimal outcomes,highlightingeitherthefeaturewiththestrongest correlationwithothersortheoneexhibitingthemost positive association. Once the feature selection process is concluded, the chosen values will be input into a separate dataframe,andanyunnecessaryoneswillbeeliminatedto form the song data frame.

D. Dataset
TheFER2013datasetcomprisesaround30,000facialRGBimagesrepresentingvariousexpressions,eachconstrained toasizeof 48×48 pixels. Theprimarylabelsinthedataset can be categorized into seven types: 0 for Angry, 1 for Disgust, 2Neutral.

E. ConvolutionalNeuralNetwork
The architectureof CNN is almost exactly the same as the neuronal communication patterns found in the human brain.Fig. 4 depicts the visualization of the layers of a convolutional neural network (CNN) feature map. It uses local receptive fields to process human perception. Deep learning typically uses the neural network idea for speech processing and image recognition. The CNN approach is employed to recognize images by utilizing a range of attributes that enable differentiation. Compared to other classificationtechniques,CNNrequireslesspreprocessing. To ensure accurate predictions, the CNN design is structured to resize the image in a manner that facilitates easy processing without compromising essential features. The CNN method functions by passing the input image through a sequence of layers, including Convolutional layer, Rectified Linear Unit, pooling layer, and Fully Connected layer

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

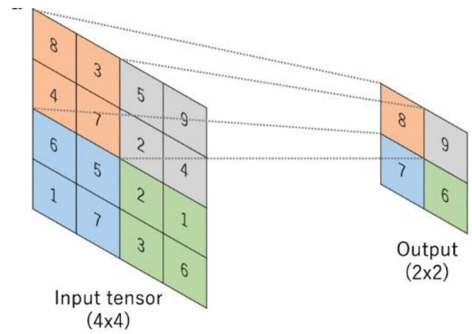

Fig.5:Max-PoolinginCNN

1) *Convolutional Layer*

Convolutionallayerstransmittheoutcometothenextlayer by performing a convolution operation on the input. Each pixel within the receptive area of a convolution is amalgamatedintoasingularvalue.Toillustrate,employing aconvolutiononan imageleadstoa decreaseinimagesize andtheamalgamationofalltheinformationwithinthefield intoasinglepixel. Theultimateoutputoftheconvolutional layer is a vector. The selection of convolution types dependsonthefeaturestargetedforlearningandthenature of the problem being addressed.

2) *Relu Layer*

Thelayerusesactivationfunctionstosendtheconvolution layer'soutputasinput,makingitnonlinear.Theconvolved feature'snoisewaseliminatedand replacedwith0.Thebest answertotheabsenceofgradientproblemshasbeenshown to be corrected linear units.

3) *Pooling Layer*

By summing or averaging values across the convolved feature maps, this layer aimed to isolate the feature maps. The pooling layer reduces the spatial dimensionality in order to provide flexible convolved features. The convolved feature in which the convolved map's size exceeds the pooling filter is covered by the filter as it is draggedoveritinthislayer.AverageandMaxpoolingare twopopularpoolingmethods.InorderforAveragePooling to function, each patch's average from the convolved feature is determined. The convolved feature is used to determine each patch's ceiling value in order for max pooling to function.

F. *Haar Cascade*

To capture the image from the user's live stream, a cascadeclassifierisemployedtodetecttheuser'sfacein the canvasprovided by the JavaScript object. Subsequently, the imageisconvertedintobytesusingarectanglebounding box through base64 and the haar function. This process enables
the video to be saved in an XML file, highlighting the contrastbetweendarkerandlighterregionsbytraversing rectangles over pixels from the two-stage features to the 38
stages, effectively minimizing the false negative ratio. Haarvalueistheproductofthesumofthepixelsinthe lighter and darker regions.

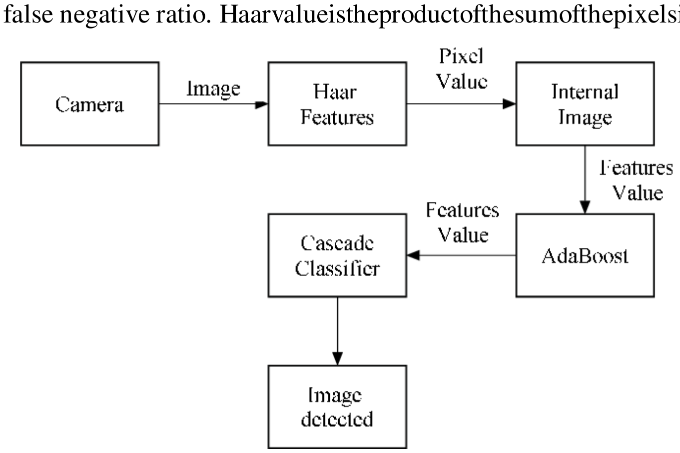

Fig.6:FlowDiagramofHaar-cascade

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

V. RESULTS AND DISCUSSION

Emotionderivedthenmappedtothespotify APItofetch the random song from the respective emotional playlist from the database. Asthesongfromthehappygenrehastobe recommended.

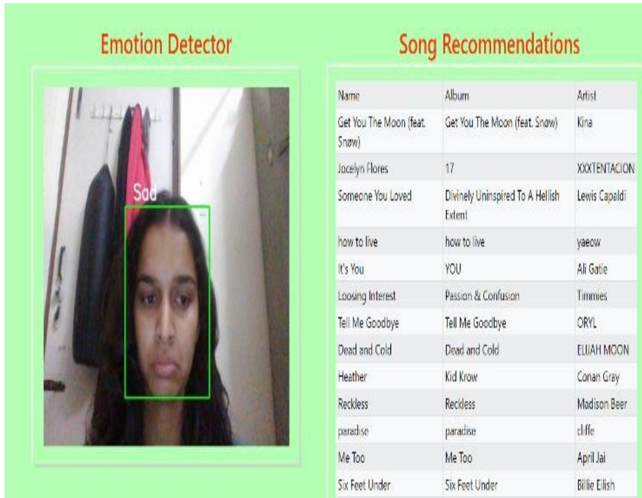

Fig.7:ResultsonSadEmotion

InaboveFig.7,useremotesasadexpressionresultingin a playlist containing sad songs.

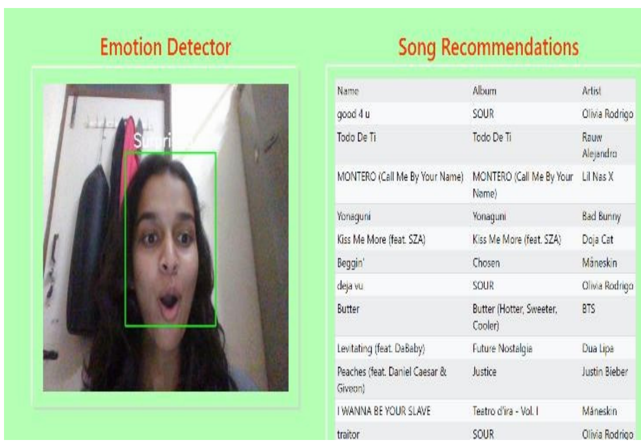

Fig.8:ResultsonSurprisedEmotion

Fig. 8, depicts results on SurprisedExpression by user. Fig.9and10depictsangryandHappyexpressionofthe user.

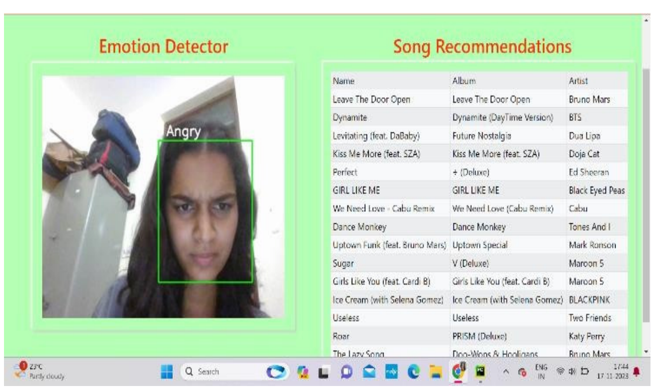

Fig.9:ResultsonAngryExpression

Fig.10.ResultsonHappyExpression

Recommendation systems play a vital role in enriching a user'smusicalpreferencesandintroducingthemtodiverse music sources, given the widespread enjoyment of music through online streaming services worldwide. Concurrently, these algorithms aid in the exploration of songs spanning different time periods. In this project, the primary objective was to create a recommendation engine utilizing Spotify's Web API to identify related music from a database of 1.2

VI. CONCLUSION AND FUTURE SCOPE

Insummary,theamalgamationoffacialemotiondetection with the Spotify API in our music recommender system represents a unique and captivating approach to enriching user interactions. Employing the FER13 dataset ensures precise identification of emotions, allowing our system to not only leverage advanced technology but also interpret users' nuanced facial expressions, providing music recommendations that are not just personalized but emotionally resonant.

Our recommender system strengthens its bond with users by interpreting facial cues and linking them to emotional states,aligningtheirmusicalpreferenceswiththeircurrent moods.Thereal-timematchingofsongstousers'evolving emotional states, coupled with the seamless integration with the extensive Spotify API, ensures a diverse and expansive music selection.

This innovative methodology goes beyond traditional recommendation systems, offering a responsive and dynamicmusicdiscovery experience.Asusersconveytheir emotions through facial expressions, our system adapts, curating playlists and suggesting songs that mirror the shifting emotional landscape. The harmonious interplay between Spotify's vast music library and facial emotion detection transforms the system into more than just a recommendation tool—it becomes a companion in the user's emotional journey.

Positioned at the intersection of emotion, technology, and music within the dynamic realm of personalized technology, this music recommender system promises a comprehensive and immersive user experience. As the work progress, continual enhancements and the incorporation of state-of-the-art technologies will ensure thatthissystemremainsafrontrunnerindeliveringtailored musicalexperiences,redefininghowusersengagewithand appreciate the impact of music in their lives.

REFERENCES

[1] Hassen,AlanKai,etal."Classifyingmusicgenresusing image classification neural networks."Archives of Data Science, Series A (Online First) 5.1 (2018): 20.

[2] Gessle, Gabriel, and Simon Åkesson. "A comparative analysisofCNNandLSTMformusicgenreclassification." (2019).

[3] Mellouk,Wafa,andWahidaHandouzi."Facialemotion recognition using deep learning: review and694.

[4] Erdal, Barış, et al. "The magic of frequencies-432 Hz vs. 440 Hz: Do cheerful and sad music tuned to different frequencies cause different effects on human psychophysiology?Aneuropsychologystudyonmusicand emotions: Frekanslarınsihri–432 Hz 440 Hz'ekarşı: Ayrıfrekanslaragöreakortlanmışneşelivehüzünlümüziklerinsanpsikofizyolojisiüzerindefarklıetkileryaratırmı? Müzikveduygularüzerinebirnöropsikolojiaraştırması."JournalofHumanSciences18.1(2021):12-33.

[5] M.J.Awan,A.Raza, A.Yasin,H.M. F.Shehzad,andI.Butt,"TheCustomizedConvolutionalNeuralNetworkof Face Emotion Expression Classification," Annals of the Romanian Society for Cell Biology, vol. 25, no. 6, pp. 5296-5304, 2021.

[6] MadhuriAthavle,etal."MusicRecommendationBased on Face Emotion Recognition." Journal of unformatics Electrical and Electronics Engineering, Vol 2. No.2, 2021

[7] Akrati Gupta, Saurabh Kumar, Rachit Kumar, Vikash Kumar Mishra "Emojify –Create your own emoji with DeepLearning"IRJMETS,Volume:04/Issue:07/July-2022

[8] Chaturvedi, V., Kaur, A.B., Varshney, V. et al.Music mood and human emotion recognition based on physiological signals: a systematic review. Multimedia Systems**28**,21–44(2022).https://doi.org/10.1007/s00530-021-00786-6.

[9] Ghosh, Oindrella, et al. "Music Recommendation System based on Emotion Detection using Image Processing and Deep Networks."2022 2nd International Conference on Intelligent Technologies (CONIT). IEEE, 2022.

[10] Sana, S. K., et al. "Facial emotion recognition based musicsystemusingconvolutionalneuralnetworks."MaterialsToday:Proceedings62(2022):4699-4706.

[11] Phaneendra, A., et al. "EMUSE–An emotion based music recommendation system."International Research Journal of Modernization in Engineering Technology and Science 4.5 (2022): 4159-4163.
[12] Shanthakumari, R., et al. "Spotify Genre Recommendation Based On User Emotion Using Deep Learning."2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT). IEEE, 2022.
[13] Bhowmick, Anusha, et al. "Song Recommendation System based on Mood Detection using Spotify's Web API."2022 International Interdisciplinary Humanitarian Conference for Sustainability (IIHC). IEEE, 2022.
[14] Dubey, Arnav, et al. "Digital Content Recommendation System through Facial Emotion Recognition."Int. J. Res. Appl. Sci. Eng. Technol11 (2023): 1272-1276.
[15] Bokhare, Anuja, and Tripti Kothari. "Emotion Detection-Based Video Recommendation System Using Machine Learning and Deep Learning Framework." SN Computer Science 4.3 (2023): 215.
[16] Sharath, P., G. Senthil Kumar, and Boj KS Vishnu. "Music Recommendation System Using FacialEmotions."Advances in Science and Technology 124 (2023): 44-52