Improving Emotion Detection and Music Recommendation Through Advanced Facial Recognition and Optimized Hyper-parameters Tuning

Vipin Sharma
x22207406
MSc in Data Analytics
National College of Ireland

Abstract

Recent years have seen a revolution in face detection all thanks to convolutional Neural Network (CNN) and deep learning techniques. As a result, many Android applications and web portals utilizing machine learning, convolutional neural networks (CNNs), and pre-trained deep learning models like VGG and Alexnet have been developed. The study's main goal is to find out how optimized hyper-parameter tuning affects model accuracy using advanced pre-trained deep learning models like VGG-16, ResNet50, and Xception. The research methodology involves training and fine-tuning these deep learning models on the facial expression FER dataset to accurately categorize different types of emotions such as happiness, sadness, anger, and surprise. For the music GTZAN dataset use the traditional machine learning algorithm Support Vector Machine (SVM) and random forest to classify Music. After this, the trained models are Combined into a system that can detect facial expressions and recommend music selections that correspond with the user's facial expressions.

***Keywords*— Deep Learning, Face Detection, Music Recommendation, Hyper-parameter Tuning, CNN.**

1 Introduction

In recent years, there has been growing interest in the field of deep learning techniques to enhance many aspects of user experience and human-machine interaction. The relationship between the music recommendation system and the facial expressions of human beings is a very interesting field for study. Priyanka et al. (2023) describe that Emotion is influenced by music and they both have close relationships and Facial expression is a common way for human to communicate their emotions. However, the human emotional state can be changed by a particular kind of music. Listening to Music helps improve the blood flow of the brain and there was a study conducted on cancer patients that found that listening or playing music helps reduce the stress level and anxiety levels of the patients significantly Lahoti et al. (2022). People enjoy listening to music but the traditional music systems or music players take more time because users first open any music player and then search for the songs that they like and suit their mood. This is a very time-consuming task and selecting good music depending on their mood is very challenging. There were many studies conducted to recognize facial expressions and recommend music according to facial expressions. According to JS et al. (2024) nearly 50% of human emotions are delivered by facial expression Moreover In face-to-face conversation body language of a human being conveys 55% of the message while words only provide 7%. Tripathi et al. (2021) describe that seven basic emotions that can be observed in the human face are Fear, anger, disgust, happiness, sadness, surprise, and neutral. In the field of psychology to accurately identifying facial expressions and human emotions is very important to understanding a person's behavior and emotions. Vayadande et al. (2023) explain that the Person's emotions and facial expressions play a very important role in communication. Both the methods Verbal and nonverbal can be used to recognize emotions. Non-verbal communication includes facial expressions and gestures and verbal communication includes sounds like a voice. There are many models developed by using the Python Keras library and traditional machine learning models like classification, support vector machine (SVM), Haar Cascade algorithm, etc to detect facial expression but Today everybody

knows how deep learning, machine learning, and Artificial intelligence affects technology and creates a very good model which can do the work very easily, fastly, and accurately also it saves the human times and provide flexibility. Taking that idea in this research deploying the advanced Deep learning models performing an optimized hyperparameters tuning and checking how the fine-tuning hyperparameters will affect the model accuracy. Hyper-parameters include various things Learning Rate, batch size, Number of Epochs, Dropout layers, Activation Function, Optimizers, kernel size, paddings, etc. These all are used to optimize and improve the accuracy of the model and also increase user satisfaction by recommending the music according to their facial expression.

1.1 Research Question:

What impact can combining advanced facial recognition with traditional machine learning for song recommendation improve the accuracy and flexibility of deep learning-based facial expression recognition systems with optimized hyper-parameters?

1.2 Research Objective

The main objective of this research is how the optimized hyperparameter tuning affects the accuracy of the model. Also, Using advanced deep learning models VGG16, Resnet50, and Xception create a good model that can better align the user's emotional states and also increase the user's music experience and consumption. Creating a model using the Keras library will be very easy Much research has already been carried out but good hyperparameters play a very important role in deep learning by determining the architecture and configuration of Convolutional Neural Network (CNN). Especially in the field of Facial expression well-tuned parameters gives guarantee that the CNN architecture effectively detects complex expressions and features of the human face and gives a more accurate emotion classification that would help to enhance the performance of the music recommendation system by suggesting the songs for the user on their current emotional states. The benefits of this research by combining facial expressions with music can be found in various domains such as mental health, entertainment, and user experience even this has therapeutic benefits in some cases particularly when it relates to mood control.

The research that has been already conducted on this topic has been summarized in section 2 of this report and I also discuss what are the methodology and methods used in this report discussed in section 3. I will also discuss the configuration and proposed tools required to create the models described in the same section. In the last section 4, the project flow and ethics were discussed, and moreover in section 5 conclusion and references.

2 Literature Review

Music is a part of our daily routine and it has a very significant role in our life. The relationship between the music and the facial expression is the very intersecting aspect of our lives. This literature review finds how other researchers explore this area using machine learning, Convolutional neural networks (CNN), and some pre-trained deep learning models. There are a total of 3 subsections created, The First section focuses on the Utilization of Convolutional Neural Networks (CNN) and Machine learning algorithms. This section talks about how the previous research uses CNN architecture and machine learning to detect facial expressions and make a music recommendation. The second section talks about how pre-trained deep learning models like VGG-16, DCNN, etc are used to classify the songs and facial images and what problems and future scope comes while implementing deep learning models. The last section talks about what are web applications and Android applications are developed to detect live facial expressions and recommend songs according to user's moods.

2.1 Utilization of Convolutional Neural Networks (CNN) and Machine learning

Priyanka et al. (2023) uses the Convolutional neural network (CNN) model which detects the live expression of the facial and recommends personalized songs on Spotify. In this paper author uses the 2-convolutional layers, 2-max pooling layer, and 1- fully connected layer to detect the expression of the image and according to the expression or moods, the list of songs played. Now the users have lots of songs to make a playlist manually or to keep track all the songs according to the mood of the user is

very difficult. The accuracy of the CNN model is good around 96% but the problem with this model is that it does not manage the head rotation and there is a need to improve the recognition rate.

Another research by Joel et al. (2023) was used the same CNN for an emotion-based Music recommendation system but this time the author tested the model on the FER dataset which is the collection of image dataset and the GTZEN dataset which is the collection of the music dataset. The important layers that are used in the constructing layers of the CNN are polling layers, convolution layers, and fully connected layers. They give accuracy only on the four types of facial expressions Happy, Fear, Sad, and Surprise. The accuracy achieved on the FER dataset by using the neural network model is 96%, 97%, 93%, and 94% of recognizing the Happy, Fear, Sad, and Surprise expressions on the face.

The Multitask cascaded convolutional neural network (MTCNN) and Face net architecture-based music recommendation system using emotion detection is also developed by Ghosh et al. (2022). The CNN model is used to predict the emotions and The classification is used to recommend the songs based on their emotion from one of the popular Spotify datasets. The methodology used includes three steps First Face detection recognition, second Mood prediction and third Music recommendation. For the music recommendation author used the Spotify API and divided the songs into different clusters and clustering completed by the Elbow method. The main goal of the cluster in this paper is to group similar songs and connect them with facial expressions.

Another paper by Qayyum et al. (2021) used the CNN architecture to build Android-based emotion detection. The CNN architecture includes the four convolutional layers four max pooling layers and two fully connected layers. The author used both CNN and RNN to check which of the techniques gives the best result on the FER2013 facial dataset. Both the networks are trained on the same dataset with seven types of different emotion classes CNN gets 65% accuracy on the trained model and RNN gets 41% accuracy which shows the RNN is behind the CNN. Similarly, Yu et al. (2020) used the CNN architecture which includes the 4-Conv layers, 2-pooling layer, and 2-fully connected layers and this architecture is the improvement of the Alexnet pre-trained deep learning model. In this paper author detects the facial micro-feature expression using the CNN classifies the emotion using the SVM model and then recommends music according to the mood. For the music use the Python crawler to tale music and store the song file and information into the Excel sheet based on the facial expression. The training accuracy is 77% on the FER image dataset and 62.1% on the Recognition rate, the author also suggested in future work to try to improve the accuracy of the model.

2.2 Utilization of Pre-trained Deep Learning Models

Deep learning models play a very important role in image detection and recognition tasks. Using the Same type of pre-trained deep learning models VGG-16 the two authors Bodhe (2024) and Lamba et al. (2023) build a Recommendations system based on the facial expressions. One author recommends Music as well as Movies and another one recommends the sea of data. The first author created a model called Moodlift it is a web-based software that captures the image from the inbuilt camera webcam and with the help of image segmentation and image processing techniques finds the important features from the users face and detects the emotion and the music, recommendation uses the Spotify API data and for the movie recommended the IMDB website movies where as the other author used FER dataset and apply the harness 3d technology to capture and decode the image expression more accurately. They use the two approaches for facial detection the first one is the Geometric feature-based approach and the second one is the holistic Feature-based approach.

Creating the new Unique architecture in CNN is a very interesting thing. JS et al. (2024) proposes the unique CNN architecture model which includes an 8-Convolutional layer, pooling layer, and dropout layer and it performs better than previous models. the author trained the model using the facial photos and Action units (AU). The AU units help to capture the movement of facial muscles. The author uses the JAFFE and FER-2013 datasets for testing the model.

The main current issue of songs music recommendation system is to fail to give the personalized and best listening music experience for the person. Using the DCNN model Shrestha et al. (2023) solves this problem and captures human emotion through this model and makes an emotion-based music recommendation system. The facial expression is detected by using the DCNN model also the feature is extracted using the MFCC and Music is classified by using the CNN model. The accuracy of the model comes to 79% and the two popular datasets used for the training and testing are FER-2013 and GTZAN. There are some recommendations in this paper for the future to improve the accuracy of using more datasets and more training. The another research was conducted by Sahare et al. (2023) in the field of Emotion-based music players. The main goal of this paper to generate a good music playlist based

on the current mood of the people and help to calm their stress and give relaxation. Using the Cohn Kanada database and HAAR cascade classification for facial expression detection but the problem with the proposed model it is comparatively slow as compared with the other model and less accurate. The unique part of this model is it could detect the background features and suggest the music according to the background nature if the person detected neutral facial emotion.

2.3 Web Portal and Mobile Apps for Music Recommendation Using Facial Expression

Vayadande et al. (2023) make a music player system that uses computer vision techniques to analyze facial expressions and according to the facial expressions recommend the song. The author uses the inbuilt camera which captures the image data and uses the image preprocessing technique and SVM algorithm to extract the facial expressions recommended in the songs. The accuracy of the system using SVM is 0.83 in all the emotion categories, the author also suggests that using Deep learning, artificial intelligence, and other machine learning algorithms improves the accuracy of the model.

One more Real-time web portal has to be designed by Lahoti et al. (2022) for detecting facial mood detection and recommending the music. In the paper, they detect the seven types of expressions. Also, using the OpenCV library detects the customer face and the different expressions are analyzed by the CNN classifier network and when the expressions are detected at the end the songs are recommended according to their mood and emotion. For creating this portal author used the FER dataset and the three main components or objectives of this paper is Face Detection, Analyzing the Mood, and Recommending the Song. HAAR cascade method is also used which use to extract every feature and PCA (Principal component Analysis) is used to detect the facial structure and use the data of more than 1-dimensional.

Mariappan et al. (2012) presents the FaceFetch which is the Multimedia Content-based recommendation system that works on both mobile and desktop. The model trained on the CK+ dataset which is the collection of image datasets with different poses. The FaceFetch system understands the emotional state and suggests different multimedia content such as video, music, and other types of videos according to user interest from the cloud with real-time performance. Using the ProASM techniques to extract the important features from the images and Using the Support vector machine (SVM) we classify the images or features into different types of facial expressions.

2.4 Conclusion of Literature Review

In all the above research found that only one or two authors use advanced pre-trained deep learning methods like VGG-16, SVM, and CNN but many pre-trained deep learning models are developed now which give good accuracy on image datasets like ResNet50, Xception, and GoogleNet. The main goal of this paper is to create a model using the pre-trained latest deep learning models like ResNet, VGG-16, and Xception, and perform optimized hyperparameter tuning. There are various things that come under hyperparameter tuning like the Activation function, Kernal Size, padding, learning rate, and number of epochs. After that compare the results and accuracy of the model with the previous results and the important things to check how the hyperparameter tuning affects the accuracy and improves the results.

3 Proposed Methodology and Specification

The methodology used for the research is KDD Methodology (Knowledge Discovery in Databases). KDD is a systematic approach suitable for extracting patterns from datasets. As shown in Figure 1 the kdd includes five steps that are Selection, Preprocessing, Transformation, Deployment, and Interpretation evaluation.

3.1 Proposed Methodology

3.1.1 Data Selection

Two datasets FER2013 and GTZAN are used in this research. The FER2013 consists of 32198 images, each is greyscale 48 X 48 pixels. face images are categorized into seven categories The GTZAN dataset has a collection of 10 genres with 100 audio files and the songs will be collected from different sources like CD, radio, and microphone, and each song is 30 seconds long.

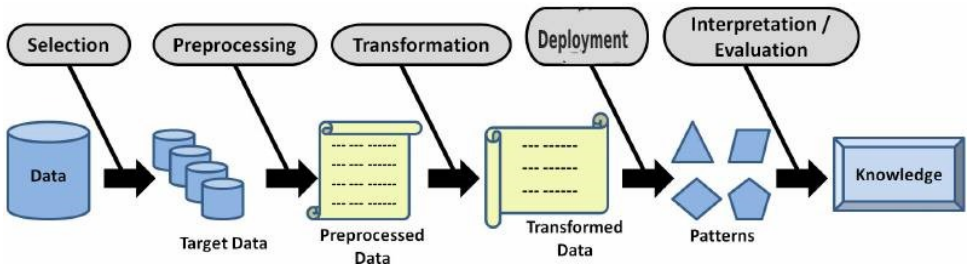

Figure 1: Project Plan for Research Work (Original Illustration)

3.1.2 Data processing

Data cleaning is performed by handling missing or corrupted images or songs, also Image resizing, Normalization, data augmentation, and feature extraction. It is important to do the feature extraction to extract relevant features from the images so that our model gives good results. Depending on the model Image resizing is also required and while working with the image data make sure all the images have the equal size.

3.1.3 Data Transformation

The data transformation step is important when talking about the image dataset because normalizing the pixel value is very important So in our case, I have to normalize the pixels of the image to a range between 0 and 1 by dividing by 255. By doing this the convergence is faster when performing the model training. I would also divide the dataset into 70% training, 20% testing, and 10% validation.

3.1.4 Model Deployment

In this step, I used the latest pre-trained deep learning model ResNet50, Xception, and VGG16.

- **Resnet50**
  Resnet50 stands for Residual Network with 50 layers. It is a pre-trained deep-learning model that solves the vanishing gradient problem. ResNet architectures have the better capacity to train very a very deep networks which has allowed them to produce advanced performance in image detection, image classification, and other computer vision tasks.

- **Xception**
  Xception is a deep convolutional neural network (CNN). It stands for Extreme Inception means working very deeply in deep learning. It is also used for performing fine-tuning and transfer learning tasks also it helps to reduce the complexity and give the string performance.

- **VGG16**
  VGG16 stands for Visual Geometry Group 16. It consists the 16 layers with 3*3 convolutional filters including max polling layers. In the image classification task, the VGG16 models perform very well because It was trained on the Imagenet dataset which includes approximately 20,000 categories and contains 14 million labeled data.

As per my knowledge from reading previous research papers no one can use all these models. After that, performed the optimized hyperparameter tuning parameters to create a good facial expression recognition model which is the main goal of our research, and for the music classification used the traditional machine learning algorithm Support vector machine and Random forest to classify the songs.

3.1.5 Model Evaluation

In the evaluation step, I have to compare which deep learning model gives the best results and accuracy and how much accuracy will be changed by performing the hyper-parameter tuning. Also evaluate the appropriate metrics such as precision, recall, F1-score, etc. This step also helps in assessing how well the models perform in detecting the facial expressions and classifying the songs using machine learning models.

3.1.6 Integration with Music Recommendation System

When the facial expression detection model is created by using deep learning models and hyper-tuning and the classify the songs using a machine learning algorithm then integrates both things to complete our research.

To achieve the goal of improving facial expression recognition with optimized hyper-parameters tuning in the deep learning models and recommending the music according to the facial expression. To making a face detection model using deep learning is very easy but doing the optimized hyperparameter tuning will affect the accuracy of the model because it will detect complete pixels of the image then using the traditional machine learning model recommend the songs according to facial expression.

3.2 Design Specification

The Below Figure 2 shows the complete workflow of my model design.

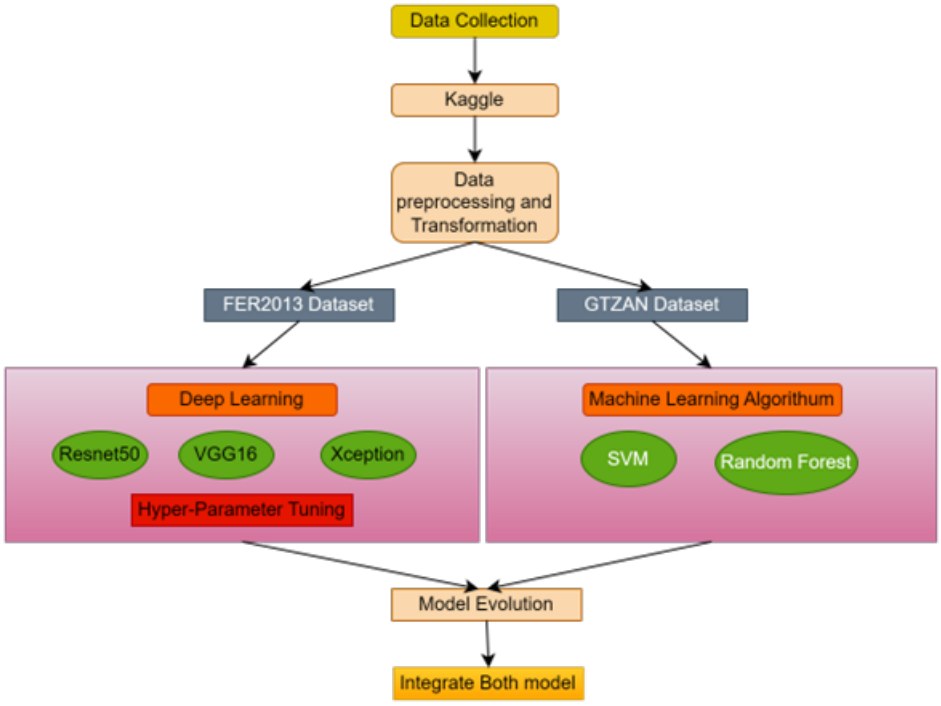

Figure 2: Model Design (Original Illustration)

Using the Python Keras library is it very easy to create a model that can easily detect facial expressions and classify or recommend the songs according to facial expressions but this research will go in-depth and work with the optimized hyper-parameter tuning. How it will affect the model accuracy and evaluation matrices. Also, I can use the advanced pre-trained deep learning models like Resnet50, VGG16, Xception to classify the Images into different categories and compare the accuracy of the models. For the classification of the music I can use the traditional machine learning algorithms Support vector machine(SVM) and Random Forest. At last, Perform the Model Evolution and then Integrate both the models and the complete system that recommended the songs according to facial expression.

3.3 Setup and Configuration

The research will be conducted using the following system configuration setup:

1. Device: Asus VivoBook 15
2. Operating System: Microsoft Windows 11 64-bit
3. RAM: 8.00 GB

4. CPU: 12th Gen Intel(R) Core(TM) i5-1235U CPU @1300Mhz
5. GPU: Intel(R) Iris Graphics
6. Version: 10.0.22631 Build 22631

3.4 Proposed Tools

The Proposed tools that will be used for this research work consist of a collection of Python packages and Python programming languages. Implementing Python packages such as NumPy, Pandas, Matplotlib, TensorFlow, Keras, Scikit-learn, and Seaborn provides help to perform data manipulation, visualization, and machine learning model development. These python packages give different features like finding the insightful data visualization using matplotlib and seaborn library, to build the effective facial expression deep learning model (Tensor Flow, Keras, Scikit-learn) will be used. For efficient numerical computation, the two basic and important libraries (Numpy and Pandas). The Python language was selected for performing the coding of this project because it will give flexibility also python language is user-friendly and effectively supports in the machine learning, deep learning, and data analytics domains. Furthermore for the interactive coding environment using Jupyter Notebook helped to improve the code readability for the other users and improve the documentation of my project. To created the Flow diagram and system architecture using the visualization tools draw.io and lastly Using excel I created the Gantt chart which shows the complete plan of our research project.

4 Ethics and Project Plain

4.1 Ethical Considerations of the Research

When performing the project Ethical consideration is very important. Especially when there are Social perspectives, the environment, human involvement, resources, and private datasets involved. In this project using deep learning concepts and detecting facial expressions, there are many ethical concerns. Keeping this in mind collecting the Face data from the open source library kaggle. It is ensured that there is no personal data of humans are involved in this research project.

4.2 Project Plan

Figure 3: Project Plan for Research Work

The project management timeline for this research project is clearly shown in Figure 3. A complete plan of how I execute my research. The research starts with the Literature review phase which runs from May 20 2024 to June 10 2024 taking a total of 21 days. After this move to the next phase which is data collection and preprocessing which is scheduled from May 25 2024 to June 3 2024 taking up a period of 9 days. Following that the model building and Training phase from June 5 2024 to June 20 2024 takes 15 days and tuning hyperparameters from June 20 2024 to June 30 2024 takes 10 days. Doing all the above steps I completed 50% of the research work and after this, the Model Evaluation phase is set for July 1 to 10th which is taking 9 days. After performing the model evaluation conduct a meeting with the supervisor in July then start report writing a report which will take 23 days. When the writing report is completed there is a final meeting conducted between 05-08-2024 to 08-08-2024. Finally the research project with finished by submission of the final report on 12-08-2024. By following this planned schedule the research project will be completed successfully.

5 Conclusion

This study explores how deep learning has changed emotion-based technology applications by combining facial expression recognition, with music suggestions. Using advanced pre-trained deep learning models such as VGG 16, ResNet50, and Xception the research successfully identifies various emotional states based on facial expressions with precision. Significant gains in model accuracy and reliability can be achieved by carefully applying optimization techniques and performing hyperparameter tuning, highlighting the importance of parameter optimization in performance optimization. Also, the benefit of applying hyper-parameter tuning is that it will detect the small pixels of the face image which can help to improve the model. This research also helps the mental health support industries, entertainment, and human customer experience.

References

Bodhe, H. V. (2024), 'Movie & music recommendation system based on facial expressions'.

Ghosh, O., Sonkusare, R., Kulkarni, S. & Laddha, S. (2022), Music recommendation system based on emotion detection using image processing and deep networks, *in* '2022 2nd International Conference on Intelligent Technologies (CONIT)', IEEE, pp. 1–5.

Joel, J. S., Thompson, B. E., Thomas, S. R., Kumar, T. R., Prince, S. & Bini, D. (2023), Emotion based music recommendation system using deep learning model, *in* '2023 International Conference on Inventive Computation Technologies (ICICT)', IEEE, pp. 227–232.

JS, G., Gleran Lobo, D., Aman, M. et al. (2024), 'Reading faces, recommending choices: A systematic review of facial emotion recognition and recommendation sy', *International Journal of Computing and Digital Systems* **15**(1), 1–12.

Lahoti, M., Gajam, S., Kasat, A. & Raul, N. (2022), Music recommendation system based on facial mood detection, *in* '2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT)', IEEE, pp. 284–289.

Lamba, S., Singh, D. & Yadav, A. (2023), 'Recommendation system based on facial expression'.

Mariappan, M. B., Suk, M. & Prabhakaran, B. (2012), Facefetch: A user emotion driven multimedia content recommendation system based on facial expression recognition, *in* '2012 IEEE International Symposium on Multimedia', IEEE, pp. 84–87.

Priyanka, V. T., Reddy, Y. R., Vajja, D., Ramesh, G. & Gomathy, S. (2023), A novel emotion based music recommendation system using cnn, *in* '2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)', IEEE, pp. 592–596.

Qayyum, R., Akre, V., Hafeez, T., Khattak, H. A., Nawaz, A., Ahmed, S., Mohindru, P., Khan, D. & ur Rahman, K. (2021), Android based emotion detection using convolutions neural networks, *in* '2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)', IEEE, pp. 360–365.

Sahare, R., Bhoyar, I., Borkar, D., Shedame, A., Deotale, A. & Mistry, S. (2023), 'Emotion based music player'.

Shrestha, A., Prajapati, B., Maka, E. & Shrestha, P. (2023), Music recommendation system with emotion detection (using dcnn model), Technical report.

Tripathi, P., Ankit, K., Sharma, R. & Kumar, T. (2021), Facial expression recognition through cnn, *in* '2021 6th International Conference on Communication and Electronics Systems (ICCES)', pp. 1–5.

Vayadande, K., Narkhede, P., Nikam, S., Punde, N., Hukare, S. & Thakur, R. (2023), Facial emotion based song recommendation system, *in* '2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)', IEEE, pp. 240–248.

Yu, Z., Zhao, M., Wu, Y., Liu, P. & Chen, H. (2020), Research on automatic music recommendation algorithm based on facial micro-expression recognition, *in* '2020 39th Chinese Control Conference (CCC)', pp. 7257–7263.