Proceedings of the International Conference on Inventive Computation Technologies (ICICT 2023)
IEEE Xplore Part Number: CFP23F70-ART; ISBN: 979-8-3503-9849-6

Emotion based Music Recommendation System using Deep Learning Model

Jaladi Sam Joel
Department of ECE
Karunya Institute of Technology and Sciences
Coimbatore, India
jaladisam@karunya.edu.in

B. Ernest Thompson
Department of ECE
Karunya Institute of Technology and Sciences
Coimbatore, India
ernestthompson@karunya.edu.in

Steve Renny Thomas
Department of ECE
Karunya Institute of Technology and Sciences
Coimbatore, India
steverenny@karunya.edu.in

T. Revanth Kumar
Department of ECE
Karunya Institute of Technology and Sciences
Coimbatore, India
tagaramvara@karunya.edu.in

Shajin Prince
Department of ECE
Karunya Institute of Technology and Sciences
Coimbatore, India
shajin@karunya.edu

Bini D
Department of Robotics Engineering
Karunya Institute of Technology and Sciences
Coimbatore, India
bini@karunya.edu

2023 International Conference on Inventive Computation Technologies (ICICT) | 979-8-3503-9849-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICICT57646.2023.10134389

***Abstract*— The ability to read a person's emotions from their face is crucial. Using a camera, the necessary input is immediately taken from the subject's face. This input may be used, among other things, to extract data that can be used to infer a person's mood. The musical selections fit the "mood" determined from the source supplied before may then be obtained using this data. This reduces the time-consuming and laborious effort of manually classifying songs into various categories and assists in creating a playlist that is suitable for a particular person depending on their emotional characteristics. The Music Thespian depends on Facial Expressions and aims to scout out and comprehend the data before building a playlist according to the specified parameters. To construct an emotion-based music system, the suggested system is designed to identifying human emotions. The article explains the techniques currently in use by music players to identify emotions, the approach that is used with our music player, and the most effective manner to make use of our emotion detection technology. A quick explanation of how the systems operate, create playlists, and classify emotions is provided.**

Keywords— Emotion, Music Therapy, Deep Neural Network, Facial Expression Recognition

I. INTRODUCTION

The computerized evaluation and interpretation of tunes by computers is a novel potential in the music industry data recovery. A wide range of study issues in this area are studied by scholars due to the diversity and depth of music material, including Musicology makes use of computer science, digital signal processing, mathematics, and statistics. Music similarity analysis, audio artist identification, audio to score alignment, inquiry by singing or humming, automatic audio genre/mood categorization, and other recent developments in music information retrieval are only a few examples.

One of the practical applications that may be offered is music suggestion based on content. More advanced context-based music recommendations are developed by using the context information. A content-based music recommendation system requires multidisciplinary work in the areas of emotion description, emotion detection/recognition, feature-based classification, and inference-based recommendation. An emotion descriptor has been successfully used to characterize music taxonomy. Emotions may be translated as a series of actual numbers set of continuous values, which is a presumption for emotion representation. In their circumflex model, investigators depicted each affect more than two bipolar dimensions., as a groundbreaking approach to describing human emotions. Arousal and sleep are the two aspects in question. Later, Russel's model was put to music by another researcher. In Thayer's approach, "arousal" and "valence" are the two key dimensions. Along the arousal level, the phrases for emotions ranged from silent to lively, while along the valence dimension, they ranged from negative to positive. The emotional 2-D plane may be split into quadrants, four into quadrants, four using Thayer's approach, and each quadrant can have eleven emotion adjectives put over it [1]. In contrast, Xiang et al. proposed a "mental state transition network" to describe how people change via their emotions. The mental states of the network include joyful, sad, angry, disgusted, afraid, surprised, and calm. However, other feelings like anxiety and excitement aren't taken into account. Automatic feeling identification and acceptance in the music is developing quickly because to advancements in digital signal processing and many efficient component extraction techniques.

Emotion finding and recognition have numerous more potential uses, including systems for interacting with computers and music pleasure [2], [3]. The initial study on emotion recognition using music was presented by Feng. By assessing two tempo and enunciation of variables that are translated into four different moods—joy, wrath, melancholy, and terror—they applied the Computational Media Aesthetics (CMA) viewpoint.

However, other feelings like anxiety and excitement aren't taken into account. By assessing two tempo and articulation variables that are translated into four different moods—joy, wrath, melancholy, and terror—they applied the Computational Media Aesthetics (CMA) viewpoint.

There are numerous additional possible uses, such as audio entertainment and HCI systems., can benefit from emotion detection and recognition. The first study on

emotion recognition using music was presented by Feng. By studying two tempo and articulation factors that are related to one other, they used the Computational Media Aesthetics (CMA) viewpoint mapped into four types of moods: joy, rage, sadness, and terror.

II. Related Work

The Automatic Face Detection and Facial Expression Recognition System to identify facial expressions was proposed initially. This system consists of three phases. 1. Face recognition 2. Expression recognition, followed by 3. feature extraction. The RGB Color Model is employed in the initial stage of face detection., lighting adjustment to obtain the face, and morphological processes to keep the necessary facial features, such as the lips and eyes. For the purpose of extracting facial features, this system also employs the AAM, or Active Appearance Model Method. This method uses a statistical model of the shape and appearance of the face to extract facial features. In this approach, the points on the face, such as the eyes, eyebrows, and mouth, as well as a data file, are created that contains information about the model points that were detected. The method also detects the face and uses the input of an expression to determine how the AAM Model should change.

Bezier Curve Fitting for the Analysis of Facial Expressions was used to recognize Emotion. A system based on Bezier curve fitting was proposed. This system uses a two-step process to detect and analyse the facial area in the input original image, and then it verifies the facial emotion of the distinguishing features in the region of interest [4-6]. To gauge the location of the face and the placement of the mouth and eyes on the face, feature maps were employed after the initial step of face identification, which used colour still images according to skin-color pixels and initiated spatial filtering. In order to apply a Bezier bend to the eye and mouth, this approach first extracts the area of prior to deriving points from the feature map of interest. This technique employs training and testing to comprehend emotion of the Hausdorff distance using a Bezier bend between the supplied face image and the database image.

Utilizing a library of photos, the user of animated music recommendation system can obtain music recommendations based on the genre of each image. The Nokia Research Center created this technique for making music recommendations. Audio signal processing and textual meta tags are used in this system to describe the genre. Utilising emotion identification from facial expressions in human-computer interaction. A completely automated face countenance and identification system built on a facial recognition in three steps, facial characteristic facial expression and extraction categorization procedure was proposed later.

Discovery of Emotion based Music through association with movie music [7-9]. With the expansion of electronic music, users can benefit from the development of music recommendations. The users' preferences for music are the basis for the current recommendation methods. However, there are occasions when selecting music based on the mood is necessary. In this research, a unique approach for association finding from film music-based emotion-based music recommendation. In order to uncover associations between emotions and musical features, the methodology looked into the extraction of musical features and adjusted the affinity graph. According to experimental findings, the suggested approach averages 85% accuracy. Interactive mood-based games Finding and recommending music. A considerable portion of research in recommender systems focuses on enhancing prediction and ranking. But recent studies have highlighted the importance of other components of the recommendations, such transparency, control, and overall user experience. On the basis of these components, we provide MoodPlay, a hybrid music endorsement engine with a comprehensible interface that combines content and mood-based filters. We show users how to utilise MoodPlay to browse music collections by hidden emotional dimensions as we move through how to combine user input at the moment of suggestion with calculations based on a previous user profile [10]. The findings of a user study (N=240) that examined four circumstances with different levels of visualisation, control and interaction are reviewed.

III. Deep Convolutional Neural Network

A. Convolutional Neural Network

Unlike traditional neural networks, CNN layers include neurons organised in three dimensions: width, height, and depth. A layer's neurons will only be partially linked to the layer before it, interactivity and window-size control rather than being completely interconnected. The main constructing layers utilized in convolution neural networks are the Convolution stage, pooling stage, and fully connected layer.

The final output layer would also contain dimensions as we would compress combining the entire image into one trajectory of class grades with the premise that the CNN architecture (number of classes).

A Convolutional Neural Network is a Deep Learning technique which has a picture-taking capability, assign different components and objects in the image importance (learnable weights and biases), and know how to distinguish between them [11]. When compared to straightforward methods filters, a Conv Net requires far as compared to other classifications, less pre-processing systems. CNNs can be trained to detect and localize faces in images. They work by using convolutional layers to extract features from the input image and a fully connected layer to classify the presence of a face. CNNs have shown state-of-the-art performance in various computer vision tasks, including image classification and object detection. They have achieved high accuracy on benchmark datasets such as ImageNet, demonstrating their ability to learn complex features from images.

Because they are hand-engineered, ConvNets may pick up on these filters and traits with the right training. The organisation of the Visual Cortex and the brain's neuronal interconnection system both have an impact on the design of a ConvNet [12-14] specific neurons only in reaction to stimuli in this little area of the visual field, known as the Receptive Field. There are several overlapping areas like this that make up the total visual field. CNNs can be adapted to different types of input data and can be used for a wide range of computer vision tasks [15–17]. They can be trained on large datasets and can learn from examples, making them suitable for various applications. The choice of CNN as an algorithm for a specific task can be justified based on its performance, adaptability, interpretability, availability, and scalability.

B. *Convolutional Layer*

The example above duplicates our 5x5x1 input picture with a green region. The element that executes the Kernel/Filter, K, convolution technique is used in the first convolutional layer, which is symbolized by the color yellow. K is represented as a 3x3x1 matrix. CNNs are capable of automatically extracting relevant features from images, without requiring manual feature engineering. The use of convolutional layers in CNNs allows the network to learn hierarchical representations of the image, where lower layers extract basic features (e.g. edges, corners), and higher layers extract more complex features (e.g. shapes, textures). CNNs have achieved state-of-the-art performance on a wide range of computer vision tasks, including image classification, object detection, and image segmentation. This is due to their ability to learn complex features from images and generalize well to new, unseen data.

C. *Pooling Layer*

The Pooling layer is similar to the Convolutional Layer in that it in responsibility of lowering the Convolved Feature's spatial size. The amount of processing resources required for processing the data will be decreased by Diminishing the dimensions. Additionally, by enabling the extraction of dominant traits that are rotational and positional invariant, it contributes in correctly training the model. The two main forms of pooling are Average and Maximum Pooling. Max Pooling brings back the highest worth from the region of the picture that a kernel has covered. The result of typical pooling is the mean of each number from the region of the picture that a Kernel covered.

Average Pooling: The pooling technique known as average pooling, which establishes the feature map is down sampled (pooled) based on the average value for patches. It usually follows a convolutional layer in application. As a result of the small addition of translation invariance, the values of the vast majority of pooled outputs are not significantly affected when the size of the picture is changed. While it extracts features more smoothly, Max Pooling extracts traits that are more visible, such edges.

Max Pooling: The "Max Pooling" convolution technique involves the Kernel extracting the most value possible from the region it convolves. The Convolutional Neural Network is only informed by Max Pooling that we will only carry forward that information if it is the greatest information available in terms of amplitude. On a 4*4 channel, we can perform the most pooling with a 2*2 kernel and a 2 stride. If we examine the initial 2*2 set, which is what a kernel is concentrating on, the channel comprises four values: 8,3,4,7. By using Max-Pooling, the highest value in that set, "8," is selected.

D. *Fully Connected Layer*

Batch size, number of inputs, and number of outputs are the three factors that characterize a fully-connected layer. Forward propagation, computation of the activation gradient, and computation of the weight gradient are all directly stated as matrix multiplications. Different frameworks have different mappings of the three parameters to the GEMM dimensions (General Matrix Multiplication, background in the Matrix Multiplication Background User's Guide), but the fundamental ideas are the same. The architecture of the basic CNN is represented in Fig. 1.

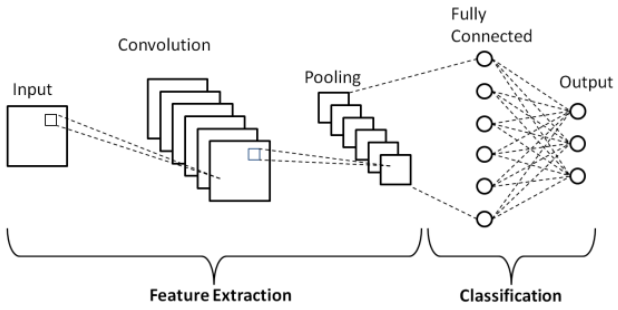

Fig. 1. Smple CNN Architecture

IV. IMPLEMENTED METHODOLOGY

The overall structure of the implemented work is explained in Fig. 2. CNN is the classifier used in this research.

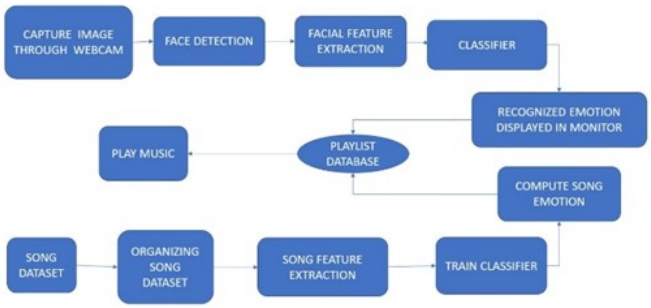

Fig. 2. Methodology

A. *Image Acquisition*

Since the Streamlit was used to build this project, which is used to build online apps, the moment the project is executed, a web page is opened and the OpenCV video stream is transformed to Grayscale to better detect faces. From the stream, the frames are taken out and processed. as

50x50-pixel images that are grayscale. The project is consistent with this dimension.

B. Face Detection

For faces, the collected photos are screened. A rectangular face is then created around each face found in each frame of the Realtime video using the Haar cascade and CNN algorithm. The technique involves training a classifier on positive and negative samples of the desired feature. Before the image is supplied to the model to get the prediction, this is a step in the preprocessing.

C. Emotion Detection

The Keras CNN model is fed the processed images. We've already trained our model. Once the image has been cropped, a DEEP LEARNING model has been trained to use it to predict the mood of the image. This will occur 30 to 40 times in 2 to 3 seconds. Once we obtain a list of emotions (which may contain duplicate components), it will first sort the list according to frequency and then eliminate the duplicates. After completing the aforementioned processes, a list of user emotions in sorted order will be available.

D. Music Recommendation

Nearly 90,000 songs in the database we utilised are arranged according to how pleasant and emotionally significant they are. The database is iterated over with the emotions sorted out, and songs are suggested depending on the emotions found in the list. Furthermore, the recognition two or more emotions can be performed in the same frame. In certain situations, songs are suggested in line with it.

V. COLLECTION OF DATA

Flicker-Faces-HQ Dataset (FFHQ): The Flickr-Faces-HQ Dataset (FFHQ) a collection of human faces that offers far higher protection for accessories like in terms of age, ethnicity, and choice of eyeglasses, sunglasses, and hats than CELEBA-HQ picture backdrop. After being retrieved from Flickr, the images were automatically cropped and aligned. There are 70,000 high-quality PNG pictures in the dataset with 1024 X 1024 and resolution is varied in terms of age, race, and picture backdrop.

Google Facial Expression comparison dataset: The triplet face images in this dataset, which was produced by Google, feature human annotations identifying the two faces with the most similar facial expressions. The dataset's purpose is to assist academics operating on problems connected to face analysis of expression, including expression-based picture picture album with expression-based retrieval summarization, feeling categorization, synthesis of expression, etc. 500K triplet pictures and 156K face shots make up the 200MB dataset.

Labelled Faces in The Wild Home (LFW) Dataset: A collection of face images called the dataset for Labelled Faces in the Wild (LFW) was assembled to study the issue of unrestricted face identification It is titled "Faces in the Wild." a publicly available standard for pair matching, commonly referred to as face verification. The collection may deploy facial recognition technology and other kinds of face identification. Over 13,000 face images totaling 173MB are included in the collection, which was compiled from the internet.

The triplet face images in this dataset, which was produced by Google, feature human annotations identifying the two faces with the most similar facial expressions. The goal of the dataset is to assist academics operating on problems connected to face analysis of expression, including based on expression picture album with expression-based retrieval summarization, feeling categorization, synthesis of expression, etc. 500K triplet pictures and 156K face shots make up the 200MB dataset.

African music dataset*:* The Royal Museum of Central Africa (RMCA) in Belgium is the source of the African music dataset. Songs can be divided into four categories: country, purpose, ethnic group, and instrumentation. When compared to the normal western music dataset, the characteristics of this collection of songs are very different.

GTZAN Dataset: 10 genres, including hip-hop, rock, classical, blues, country, disco, jazz, reggae, and pop, are represented in the GTZAN dataset, also known as the Genre Collection dataset, from MARSYAS (Music Analysis Retrieval and Synthesis for Audio Signals). The most often used dataset for machine learning research on music genre identification (MGR) is one that is publicly available. The experimental results outperformed other well-known methods in terms of classification performance. The 10 classifiers used majority voting to decide, and on the GTZAN dataset, they had an average accuracy of 94%.

VI. RESULTS AND DISCUSSIONS

The process of providing suggestions requires careful consideration of many different aspects, including the specific situation, individual preferences, sentiments, and emotions. The personalization, human emotions, contextual desires, and emotional variables gaps in today's music recommendation algorithms are challenges. The proposed work employed CNN from deep learning for the classification of emotions. A multilayered neural network termed a Convolutional Neural Network can recognize intricate details in the data. The performance metrics such as Accuracy, Recall, Precision are used to analyse the network model. The confusion matrix was used to calculate the true positive (TP), false positive (FP), true negative (TN), false negative (FN) to evaluate the model.

In order to grasp the pattern of the pictures provided for the CNN model, it applies a variety of filters on the image. With the use of CNN layers, the model in this system is trained by passing a certain number of photos for each of the four emotions. Once trained, the model can then predict the test data's emotions. By integrating non-linearity

through MaxPooling, which increases the model's performance in CNN, the model becomes more capable of handling a variety of data. The music was recommended based on the mood of the individual's recognized emotion. Figure 3 (a) depicts the developed page for music recommendation system based on the recognized emotion. (b) shows the recommended music based on the recognized emotion. (c) exhibits the accuracy plot using CNN model, (d) illustrates the recognized emotion, and Table 1 presents the performance analysis of CNN model. The experiments were tested using the FER database. The accuracy achieved by using the neural network model is 96%, 97%, 93%, 94% for recognising Happy, Fear, Sad and Surprise. Increasing the depth of a CNN can improve its performance by allowing the network to learn more complex features. This can be achieved by adding more convolutional and pooling layers to the network.

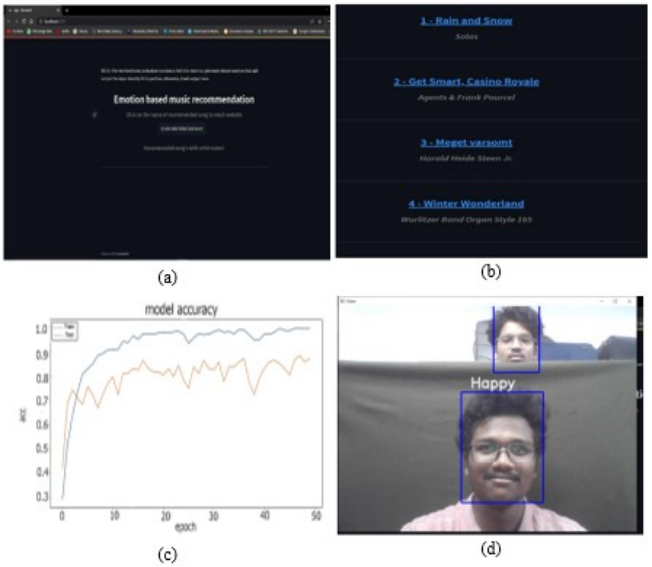

Fig. 3: (a) Home Page (b) Recommended Music based on Emotion recognized (c) Accuracy plot using CNN model (d) Emotion Detected

TABLE I. PERFORMANCE ANALYSIS USING CNN MODEL

	Specificity	Sensitivity	Precision	Accuracy
Happy	0.95	0.97	0.96	96 %
Fear	0.96	0.96	0.94	97 %
Sad	0.95	0.94	0.93	93 %
Surprise	0.97	0.92	0.95	94 %

CONCLUSION

In this research paper, an emotion-driven recommendation model that effectively addresses personal preferences and particular life and activity conditions. The prime objective of the study's approach is to maximize the benefits that listening to music may provide for individuals. Giving the algorithm feedback on the results of the recommendations will enable it to improve the music selections over time. In order to find the best suitable musical choices for each user, the system is built to listen to each unique user and understand their listening goals, moods, and contextual preferences. For this, the system is given data from a variety of sources.

In the context of this study, the key data processing techniques are specified, and the experimental prototype has been developed. However, machine learning model need a lot of data to train the models in order to create predictions that are as accurate as possible and more or less relevant. The procedure of collecting data is now underway. In order to fine-tune and test the model for accurate suggestions and minimize any side effects, this type of system need extensive clinical study and collaboration with psychologists. The accuracy achieved by using the neural network model is 96%, 97%, 93%, 94% for recognising Happy, Fear, Sad and Surprise.

The further development of this work might be viewed in this sense as the production of music by artificially intelligent systems with specific musical characteristics to affect emotional states of people.

ACKNOWLEDGMENT

We thank the signal processing laboratory, Department of Electronics and Communication, Karunya Institute of Technology and Sciences for the support extended for this research.

REFERENCES

[1] C. Loconsole, C. R. Miranda, G. Augusto, A. Frisoli, and V. Orvalho, "Real-time emotion recognition: Novel method for geometrical facial features extraction," in VISAPP 2014 - Proceedings of the 9th International Conference on Computer Vision Theory and Applications, 2014, vol. 1, pp. 378–385, doi: 10.5220/0004738903780385.

[2] S. Bhutada and T. Iv, "EMOTION BASED MUSIC," J. Emerg. Technol. Innov. Res., vol. 7, no. 4, pp. 2170–2175, 2020.

[3] B. T. Nguyen, H. Chi Minh city, V. H. Minh Trinh, T. V Phan, and H. D. Nguyen, "An Efficient Real-Time Emotion Detection Using Camera and Facial Landmarks," in Seventh International Conference on Information Science and Technology, 2017, pp. 251–255.

[4] N. Chouhan, A. Khan, J. Zeb, and M. Hussnain, "Deep convolutional neural network and emotional learning based breast cancer detection using digital mammography," Comput. Biol. Med., vol. 132, no. March, p. 104318, 2021, doi: 10.1016/j.compbiomed.2021.104318.

[5] W. Deng, J. Hu, S. Zhang, and J. Guo, "DeepEmo : Real-world Facial Expression Analysis via Deep Learning," IEEE IEEE VCIP, 2015.

[6] E. Reinertsen and G. D. Clifford, "Emotional Detection and Music Recommendation System based on User Facial Expression Emotional Detection and Music Recommendation System based on User Facial Expression," 2020, doi: 10.1088/1757-899X/912/6/062007.

[7] Z. Liu, W. Xu, W. Zhang, and Q. Jiang, "An emotion-based personalized music recommendation framework for emotion improvement," Inf. Process. Manag., vol. 60, no. 3, p. 103256, 2023, doi: 10.1016/j.ipm.2022.103256.

[8] R. Saranya, "EMOTION BASED MUSIC RECOMMENDATION SYSTEM," Int. Res. J. Eng. Technol., vol. 6, no. 3, 2019.

[9] M. Athavle, D. Mudale, and U. Shrivastav, "Music Recommendation Based on Face Emotion Recognition," vol. 02, no. 018, pp. 1–11, 2021.

[10] A. Mahadik, "Mood based music recommendation system," no. March, 2022.

[11] L. Nwosu, H. Wang, J. Lu, I. Unwala, X. Yang, and T. Zhang, "Deep Convolutional Neural Network for Facial Expression Recognition using Facial Parts," IEEE 15th Intl Conf Dependable, Auton. Secur. Comput., pp. 1318–1321, 2017, doi: 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.213.

[12] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, "Deep learning-based facial emotion recognition for human–computer interaction applications," Neural Comput. Appl., vol. 8, 2021, doi: 10.1007/s00521-021-06012-8.

[13] I. More, V. Shirpurkar, Y. Gautam, and N. Singh, "Melomaniac-Emotion Based Music Recommendation System," IJARIIE, no. 3, pp. 1323–1329, 2021.

[14] R. De Prisco, A. Guarino, and D. Malandrino, "applied sciences Induced Emotion-Based Music Recommendation through Reinforcement Learning," 2022.

[15] I. More, V. Shirpurkar, Y. Gautam, and N. Singh, "Melomaniac-Emotion Based Music Recommendation System," IJARIIE, no. 3, pp. 1323–1329, 2021.

[16] R. K. G. A, R. K. Kumar, and G. Sanyal, "Facial Emotion Analysis using Deep Convolution Neural Network," pp. 1–6.

[17] M. Athavle, D. Mudale, and U. Shrivastav, "Music Recommendation Based on Face Emotion Recognition," vol. 02, no. 018, pp. 1–11, 2021.