Received: March 31, 2024. Revised: June 9, 2024.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A Model for Song Recommendation Based on Facial Emotion Analysis and Musical Emotion

Hung Nguyen[1] Nha Tran[1]* Y Tran[1] Dat Ly[1] Anh Tran[1] An Nguyen[1] Hau Vo[2]

[1]*University of Education, Ho Chi Minh city, Viet Nam*
[2]*Western Australian International School System (Middle School & High School), Vietnam*
* Corresponding author's Email: nhatt@hcmue.edu.vn

Abstract: Enhancing user experience in music listening through facial emotion recognition presents a novel avenue for curating personalized song playlists, obviating the need for traditional search methodologies like keyword inputs or manual playlist exploration. This research pivots towards leveraging advanced pre-trained models, specifically ResNet18 and VGG19, for the accurate detection and analysis of listeners' emotions via facial expressions. By integrating these models with the Facial Emotion Recognition (FER) 2013, and The Extended Cohn-Kanade Dataset (CK+) dataset, our study aims to precisely identify a range of emotions including happiness, anger, sadness, surprise, disgust, and neutrality. In parallel, we employ Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA) techniques to discern musical emotions from Spotify's dataset, encompassing moods such as happiness, sadness, energy, and calmness. A unique set of rules is formulated to amalgamate the insights from facial emotion recognition with musical emotion data, facilitating the automated suggestion of playlists that resonate with the user's current emotional state. Our experiments demonstrate notable results with the ResNet18 model achieved 72.67% accuracy and a 72.30% F1 score on the FER2013 dataset, with significant improvements on the CK+ dataset, reaching 96.97% accuracy and a 96.60% F1 score. The VGG19 model similarly excelled, particularly on the CK+ dataset with 95.96% accuracy and a 96.60% F1 score. On the FER2013 dataset, the VGG19 model achieved an accuracy of 72.64%, and an F1 score of 72.60%. Our music emotion recognition approach, using SVM on Spotify's dataset, achieved an 86.96% accuracy and an 84.54% F1 score. When using the LDA model on the same dataset, the accuracy was 79.00% and an F1 score of 76.72%. To bring this concept to fruition, we have developed an intuitive web application that allows users to capture their facial expressions via their device's camera, subsequently offering song recommendations tailored to their detected mood.

Keywords: Musical emotions, Facial expressions, Spotify API, Support vector machine, Linear discriminant analysis.

1. Introduction

Music has always been present in life, fulfilling various communal and personal reasons, dating back to the beginning of humanity. One of the primary reasons individuals seek out and interact with music is to express and regulate their feelings. It can make individuals happy, sad, cry, or recall certain memories. Recently, the study direction of Music Emotion Recognition (MER) has gained considerable attention in academia and industry, as Tianyue Jiang et al. [1] have developed a system that integrates emotion analysis with conventional online music streaming services. This system creates an interactive interface between humans and music, based on the analysis of multi-dimensional emotions and using deep learning techniques. Therefore, it has spurred researchers from many different fields over the past decades, especially machine learning researchers.

Many of us often feel overwhelmed and spend a lot of time thinking when it comes to choosing music from the vast array of options available in the world's music collection. The diversity of music makes the musical experience of listeners varied, also contributing to expressing the current psychological

state of the listener. Over 60% of listeners report that their music collections become so expansive at times that locating a specific track they wish to play becomes a challenge [2]. Very few music listening applications previously have been able to solve the problem for users about choosing a random song suitable for their current mood.

With the growth and broader application of music emotion recognition, the area of facial emotion recognition has similarly drawn the interest of researchers and has achieved notable accuracy [3-4]. Khang et al. [5] proposed using advanced technologies like CNN and Long Short-Term Memory (LSTM) to recognize emotions from facial expressions and voice, suggesting combined models capable of utilizing both image and sound data. Similarly, the research team of Khaireddin et al. [6] presented a model based on the VGGnet architecture achieving the highest single-network accuracy of 73.28% on the FER2013 dataset without using additional training data, proving the effectiveness of deep learning models for FER. Shan Li et al. [7] provided a comprehensive review of deep facial expression recognition (FER) systems, including datasets, algorithms, and training strategies, addressing issues like overfitting and expression-irrelevant variations, while also discussing modern deep neural networks for FER, their advantages and limitations, and providing detailed information on related issues, application scenarios, challenges, and future directions.

In the research direction aiming to create a music recommendation system, some researchers have considered it, but they did not generate song recommendations based on facial emotions. Specifically, Hung-Chen et al. [8] designed a Music Recommendation System (MRS) to provide music recommendation services based on music data groups and user preferences. Similarly, in the paper by Han-Saem Park et al. [9], Using the fuzzy system, Bayesian networks, and utility theory, a situation-Aware Music Recommendation System (CA-MRS) was created to suggest appropriate music based on the user's situation.

Recognizing that potential, we have integrated both fields of facial emotion recognition and music emotion recognition to support users in choosing songs based on their current emotions. In this study, we develop a model to recommend a playlist to users based on emotion analysis from facial expressions combined with musical features from songs. Contributions in this study include:

• Utilizing optimized versions of pre-trained models, VGG19 and ResNet18, for the accurate recognition of user emotions via facial expressions, ensuring streamlined deployment in applications.

• Analyzing musical features from the Spotify music database. From there, classifying musical emotions based on SVM and LDA models.

• Integrating facial expression recognition and music emotion to recommend a playlist to users based on a set of rules that represent the relationship between facial emotions and musical emotions.

• Creating a web music listening application integrated with a music recommendation function through facial emotions.

In the following section, we will present the subsequent parts: Section 2 will provide an overview of related works on facial emotion detection and musical emotion, Section 3 will detail the experimental database, Section 4 will elaborate on the proposed methodology, Section 5 will discuss the results and their implications, and finally, Section 6 will conclude our work and discuss future research directions.

2. Related work

A person's current emotional state is often conveyed through their facial expressions. In our interactions with others, we typically communicate emotions using non-verbal cues, including hand gestures, facial expressions, and the tone of our voice [10-13]. Various approaches have been proposed to recognize human emotions, such as one approach involves extracting features from electroencephalogram (EEG) signals to identify emotions while listening to music [14]. However, this method faces significant drawbacks such as the complexity in data collection and the need for specialized equipment. In another study [15], a novel approach for automated music selection based on facial emotion recognition is proposed, departing from manual sorting and wearable devices. Utilizing a Convolutional Neural Network for emotion detection and Pygame & Tkinter for music recommendations, the system streamlines computational processes and enhances accuracy, tested on the FER2013 dataset. By capturing facial expressions via an inbuilt camera and generating music playlists accordingly, the system offers improved computational efficiency compared to existing methods.

A more practical approach, in the field of music recommendation research, camera images are often used to detect emotions on the face. To recognize human emotions, research [16] focuses on facial expressions and speech through microphones and web cameras. A. Gupte et al. [17] created an Android

app called X-Beats, designed as a personalized music player that selects songs based on the user's mood. They utilized Eclipse and OpenCV for implementing facial recognition through skin color analysis. However, the accuracy of facial recognition can be affected by lighting conditions and the position of the camera. Furthermore, their research includes a comparative analysis of various algorithms employed in facial recognition technology.

Preema et al. [18] argued that creating and managing large playlists as well as selecting songs from these playlists can be very time-consuming and difficult. Therefore, it would be very useful if the music player could automatically select a song based on the current mood of the user. The paper also suggests using the Viola-Jones algorithm and multi-class SVM for face detection and corresponding emotion recognition. Nevertheless, the computational cost and processing time can be high. H. Immanuel James et al. [19] introduced a novel emotion-based music recommendation system, drawing insights from film music to find correlations between emotions and music features. While achieving an impressive average accuracy rate of 85%, their system did not account for emotions such as disgust, fear, and neutrality, limiting its applicability. They focused on music feature extraction and refining the affinity graph for better identifying links between music and emotions like happiness, anger, sadness, and surprise, while also addressing the absence of emotions such as disgust, fear, and neutrality in their system. Their approach resulted in the model achieving an impressive average accuracy rate of 85%.

G. Dhand et al. [20] presented a model for song recommendation based on facial emotion detection and suggesting appropriate music to users. For facial recognition, they used a JavaScript facial recognition framework called face-api.js. The 68 points face landmark detection mode was used to analyze the user's expression. Despite its effectiveness, the framework's performance may vary with different facial expressions and lighting conditions. To detect music mood, two methods were used: Logistic Regression and Neural Networks, with 5 emotions identified as aggressive, dark, energetic, happy, relaxing.

Currently, there are a number of modern music recommendation systems such as EMO Player is an application designed to play music in alignment with the user's emotional state. It achieves this by capturing the user's emotions through the camera and selecting songs that match those emotions. The emotion detection process is facilitated by the Support Vector Machine (SVM) algorithm [21]. MusicAI is an end-user app installed on a smartphone device for providing users with online music listening. Once the user selects their current emotion, MusicAI displays a list of songs for the user to listen to [22]. Reel Time AI: This system operates by requiring users to register their accounts. Then, users can upload images of large gatherings such as shopping centers, cinemas, and restaurants. The system then identifies happy and sad emotions. It recognizes which faces express happiness and which express sadness and makes assessments about the situation based on the faces of those present [23].

It's important to note that recommending the same playlist solely based on emotions might not be ideal due to different preferences and needs among individuals, such as children and adults experiencing the same emotion but choosing different music genres. Furthermore, recommendations can vary based on cultural background or nationality. Therefore, it is crucial to consider additional factors beyond emotions to make more suitable and personalized music recommendations [24].

Therefore, our work aims to address these limitations by incorporating additional factors beyond emotions to make more suitable and personalized music recommendations. This includes considering individual preferences, cultural backgrounds, and situational contexts to enhance the accuracy and relevance of the recommendations.

3. Datasets

3.1 Datasets FER

FER-2013 [25] a dataset created using Google's search API, comprising 06 basic and neutral emotions by matching 184 emotion-related keywords. Detailed information about the ethnicity or race of individuals is not provided in this dataset. It contains over 30,000 grayscale images, each labeled with one of 07 emotion labels: angry, disgusted, fearful, happy, sad, surprised, and neutral, as detailed in Table 1. Each sample in the dataset shows significant variation in age, facial orientation, or other aspects, closely resembling real-life situations. One of the challenges of this dataset is the imbalance of data, low resolution, and image quality. Additionally, this dataset includes some invalid samples, including images without faces, inaccurately cropped faces, and labels not matching the facial expressions. The details are described in Table 1.

Dataset CK+ [26]: Released in 2010 as an expansion of the original Cohn-Kanade (CK) database, this version is set in a laboratory environment, showcasing 593 sequences with

Table 1. The number of each emotion in the FER-2013 dataset

Expression	Training	Validation	Testing	Total
Angry	3196	799	958	4953
Disgust	349	87	111	547
Fear	3278	819	1024	5121
Happy	5772	1443	1774	8999
Sad	3972	993	1233	6198
Surprise	3864	966	1247	6077
Neutral	2537	634	831	4002
Total	22968	5741	7178	35897

Figure. 1 The relationship between facial emotions and musical emotions

Table 2. The number of each emotion in the CK+ dataset

Expression	Training	Validation	Testing	Total
Angry	86	22	27	135
Disgust	114	28	35	177
Fear	48	12	15	75
Happy	133	33	41	207
Sadness	54	13	17	84
Surprise	159	40	50	249
Contempt	34	9	11	54
Total	628	157	196	981

Table 3. Song identification information attributes

Attribute	Description
Name	Name of the song
Album	Album name of song
Artist	Name of artist/group performing the song
Id	Song identifier in Spotify
Release_date	Song release date
Popularity	Song popularity on Spotify (from 0 to 100)
Length	Song duration (measured in seconds)

resolutions of 640x490 and 640x480. It catalogues eight expressions, including the seven basic emotions: Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise, with Contempt as an additional category. In this study, we used 981 images from the dataset with 7 types of expressions: Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Contempt. The details are described in Table 2.

3.2 Spotify data to identify musical emotions

In this paper, we use the data_moods dataset from "Spotify Music data to identify the moods" on Kaggle [25] to determine the listener's emotions based on the audio features of the song. This dataset was created by combining data from Spotify music data through the Spotify API to collect song information and mood data to label the songs. After extracting the information about the music features, we store it in a CSV file named "data_moods.csv". This file contains 686 rows corresponding to 686 songs and 19 columns corresponding to 19 attributes of each song. The attributes include name, album, artist, id, release_date, popularity, length, danceability, acousticness, energy, instrumentalness, liveness, valence, loudness, speechiness, tempo, key, time_signature, mood. The data is divided into 03 types of information, including song identification information, music feature information, and classified into four moods for each song. The song identification information is presented in Table 3. With the data obtained, we divided it into sets consisting of 80% for training and 20% for testing. After division based on each label.

For the audio feature information of each song, there are 11 columns presented in Table 4. The mood is categorized into types of moods, including 4 types: Happy, Sad, Calm, Energetic.

4. Method

We propose a method to solve the problem of recommending a playlist to users based on the analysis of facial expressions and musical emotions as illustrated in Fig. 2. Initially, we use a camera to scan the user's face. Then, we employ a facial emotion recognition model to determine their current emotion. For musical emotions, we randomly select a list of songs and then use a music mood recognition model to analyze the emotions of each song. Finally, we integrate both facial and musical aspects using a set of proposed rules to provide users with a playlist that matches the emotions of their faces. The output of the model is to recommend the most suitable playlist for the user's mood after scanning with the camera.

4.1 Facial emotion recognition

4.1.1. VGG19

In our study, the architecture of the VGG19 [26] model is specifically described in Fig. 3. The network input consists of preprocessed images including conversion to grayscale and resizing to 48x48 pixels. Next, the input data passes through a series of processing blocks comprising convolutional layers, Batch Normalization operations to normalize the data, and ReLU activation functions to extract image features at various scales and levels. Between these blocks are MaxPooling layers that reduce the spatial dimension of features, enhancing computational efficiency. Finally, the output from the feature extraction blocks is fed into a fully-connected layer with a Softmax activation function to perform the classification task, assigning emotion labels to the image inputs. The deep and complex architecture of VGG19 has been proven to be highly suitable for the problem of emotion recognition, a field that requires the ability to extract diverse and subtle features from facial expressions.

4.1.2. ResNet18

ResNet [27] was introduced to tackle one of the challenges encountered in neural networks: the instability arising from increasing network depth, leading to issues like Vanishing Gradients or Exploding Gradients. Its main idea comes from the shortcut connection technique, which helps connect across one or more layers. Such a block is called a Residual Block, illustrated as Fig. 4. In this study, we use the ResNet18 model for the task of facial emotion recognition. The architecture of the model is described as in Fig. 4.

For each input, which is a 48x48 image that has gone through preprocessing steps, it first passes through an initial convolutional layer to extract feature maps. Then, these feature maps sequentially go through multiple residual blocks built based on the residual architecture. In each block, higher-level features are learned through convolutions, ReLU activation, and Batch Normalization. Shortcut connections allow information to pass from the input to the output of the block, helping to avoid the vanishing gradient problem. Finally, the feature maps from the last block are passed through a pooling layer to reduce their size, and then they are flattened into a feature vector. This vector goes through the fully connected layer to make a classification prediction for the input image.

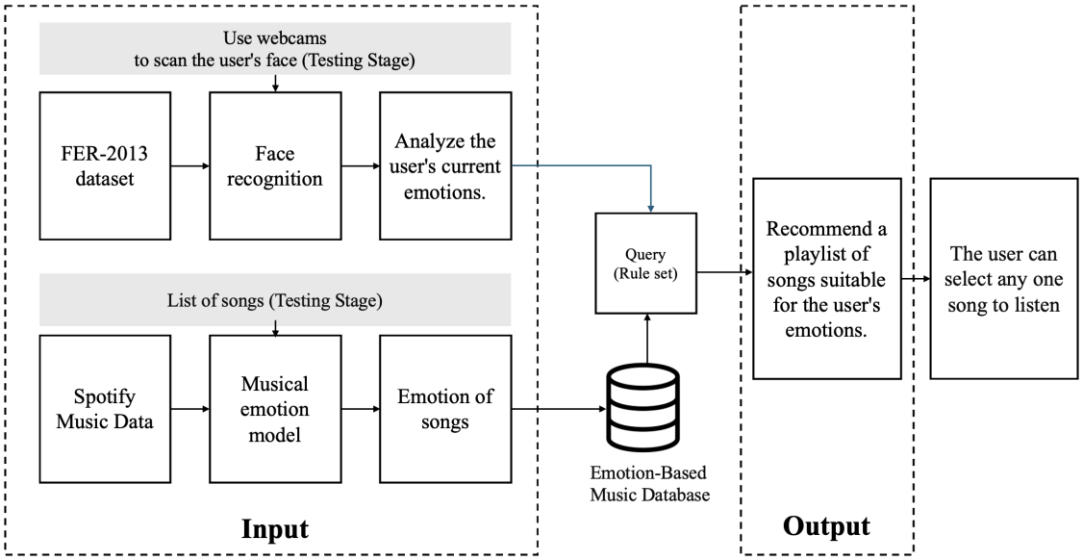


Figure. 2 Overview diagram of model implementation steps

Table 4. Music characteristic information attributes

Attribute	Description
Danceability	Measures the extent to which a song is acoustic, ranging from 0.0 to 1.0.
Acousticness	Reflects the intensity and energy of the music. For example, genres like death metal tend to have high energy.
Energy	Reflecting the power and vibrancy of music. For example, genres like death metal often exude high energy.
Instrumentalness	Predicts whether a song contains vocals.
Liveness	Detects the presence of a live audience in a song.
Valence	Describes the overall positivity or negativity of a song.
Loudness	Measures the loudness of a song in decibels (dB).
Speechiness	Identifies the presence of speech in a song.
Tempo	Indicates the tempo in beats per minute (BPM) of a song.
Key	Represents the song's musical key using standard notation.
Time_signature	Estimates the overall time signature of a song.

4.2 Model for recognizing musical emotions

A cross-disciplinary study field that includes audio signal processing, natural language processing (NLP), and music psychology is called Music Emotion Recognition (MER) [28]. In order to comprehend the relationship between emotions and music, it seeks to extract and analyse musical elements. In order to analyse emotions based on music information, we use two models in this study: Support Vector Machine (SVM) [29] and Linear Discriminant Analysis (LDA) [30].

4.2.1. Support vector machine

Support Vector Machine (SVM) is known as an effective algorithm for binary classification and supervised learning prediction tasks. SVM is used to determine a set of linear separating hyperplanes, which are linear functions in a high-dimensional feature space. Suppose we have a training dataset D=$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where x_i is the feature vector and y_i belongs to {-1, 1} is the corresponding class label.

The SVM algorithm can be described through the following steps: First, we need to find the optimal hyperplane expressed by the equation:

$$W^T x + b = 0 \tag{1}$$

To find this hyperplane means finding w and b so that the margin is maximized (equation x), where the margin is the nearest distance from any point to the plane.

$$margin = \frac{2}{\|w\|} \tag{2}$$

Where: *x* is the feature vector of a data point; *n* is the number of data points in the dataset; *w* is the normal vector of the separating hyperplane; *b* is the bias term of the hyperplane.

After finding the hyperplane, we classify new data points using the sign function sgn. The sgn function determines the sign, taking the value of 1 if the argument is non-negative and -1 otherwise. In the SVM, we use the sign function applied to the decision function $f(x) = W^T x + b$:

$$sgn(W^T x + b) = \begin{cases} 1 \; if \; W^T x + b \geq 0 \\ -1 \; if \; W^T x + b < 0 \end{cases} \tag{3}$$

These two values represent the two classes.

4.2.2. Linear discriminant analysis

Linear Discriminant Analysis (LDA) is effective for classification problems, finding a linear combination of features that best separates classes by maximizing separation and minimizing within-class variance. In music emotion classification, LDA classifies emotions into Calm, Happy, Sad, and Energetic by projecting data onto a lower-dimensional space where these classes are well-separated. Given a training dataset D=$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where x_i is the feature vector and y_i is the class label.

The LDA algorithm follows these steps:

1. Compute the within-class scatter matrix S_W:

$$S_W = \sum_{c=1}^{C} S_W^{(c)} = \sum_{c=1}^{C} \sum_{x \in c} (x - \mu_c)(x - \mu_c)^T \quad (4)$$

where $S_W^{(c)}$ is the scatter matrix for class c, and μ_c is the mean vector of class c.

2. Compute the between-class scatter matrix S_B:

$$S_B = \sum_{c=1}^{C} N_c (\mu_c - \mu)(\mu_c - \mu)^T \quad (5)$$

where N_c is the number of samples in class c, μ_c is the mean vector of class c, and μ is the overall mean vector of the data.

3. Solve the generalized eigenvalue problem:

$$S_W^{-1} S_B w = \lambda w \quad (6)$$

Here, w are the eigenvectors (discriminant directions), and λ are the eigenvalues.

4. Sort the eigenvectors by eigenvalues, selecting the top eigenvectors to form matrix W.

5. Project the data onto the new subspace:

$$y = W^T x \quad (7)$$

where W is the matrix of the top eigenvectors.

LDA enhances class separability, ensuring high classification accuracy and reducing computational complexity, making it effective for emotion recognition tasks in music.

4.3 Set of rules to represent the relationship between facial emotions and music

To explore the relationship between emotions analyzed from facial expressions and musical emotions, we have developed a rule set to represent this relationship, helping the overarching model to suggest playlists based on this rule set. The proposed rule set is presented in Table 5 and Fig. 1.

Algorithm 1 presents a set of rules that aim to establish a relationship between facial emotions detected through computer vision techniques and musical emotions. This relationship can be leveraged to suggest personalized playlists tailored to an individual's current emotional state, as reflected by their facial expressions.

Algorithm 1

E_f : Set of facial expressions

E_m : Set of music emotions

N: Number of songs in the database

$S \in R^{6\times4}$: Display ratio matrix, where S_{ij} is the di splayratio of music emotion j when facial emotion is i

Matrix S:

$$S = \begin{pmatrix} 0.5 & 0.25 & 0.25 & 0 \\ 0.25 & 0.5 & 0.25 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 0.5 & 0.5 \\ 0.25 & 0 & 0.25 & 0.5 \\ 0..25 & 0.25 & 0 & 0.5 \end{pmatrix}$$

In this matrix:

- Rows correspond to E_f.
- Columns correspond to E_m.

INPUT:

$E_f \in$ {"Happy", "Sad", "Fear", "Angry", "Disgust", "Surprise"}

$E_m \in$ { "Happy", "Sad", "Calm", "Energetic"}

OUTPUT:

Array L, a list containing the recommended songs

ASSUMPTIONS:

1. E_f (Disgust) = {Energetic, Happy, Calm}
2. E_f (Angry) = {Energetic}
3. E_f (Fear) = {Calm}
4. E_f (Happy) = {Happy, Calm, Sad}
5. E_f (Sad) = {Sad, Calm}
6. E_f (Surprise) = {Energetic, Happy}

Step 1: Identify the facial emotion $e_f \in E_f$

Step 2: Find the row i in the matrix S corresponding to e_f

Step 3: If $S_{ij} > 0$, add [$S_{ij} \times N$] elements of $e_m \in E_m$ corresponding to column j to the list L

Figure. 3 VGG19 model architecture

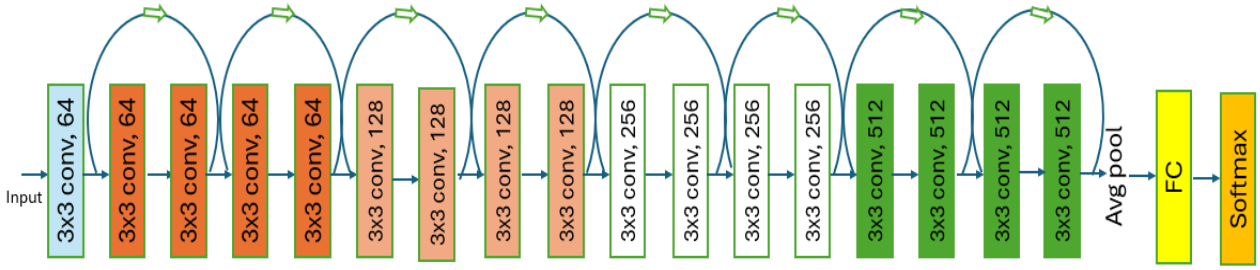


Figure. 4 ResNet18 model architecture

Table 5. Diagram showing the relationship between facial emotions and musical emotions

Facial emotions	Musical Emotions	Function
Disgust	Energetic / Happy	These songs bring energy and joy, helping to reduce feelings of disgust and improve mood.
	Calm	Soft music helps create a peaceful space, supporting the process of processing and reducing negative emotions.
Angry	Energetic	Provides a channel to release negative energy through positive activity, helping to reduce feelings of anger
	Calm	Brings relaxation and soothes the mood, helping to reduce stress and anger.
Fear	Happy	Cheerful melodies provide positive distraction, helping to lessen feelings of fear.
	Calm	Gentle music creates a feeling of safety and helps reduce stress, diminishing fear.
Happy or Sad (Happy)	Reflecting mood / Calm	Choose music that reflects this mood or soft music (Calm) to maintain a feeling of happiness and comfort.
Happy or Sad (Sad)	Sad / Calm	Sad music can provide empathy and help process sorrow. Soft music (Calm) also aids in returning to a state of mental balance and relaxation.
Surprise	Energetic / Happy	Enhance positive reactions to surprises through lively and cheerful melodies.
	Sad	Suitable when surprise brings about negative emotions, helping the listener to accept and process those feelings.

5. Results

We use a GPU P100 device and 13 GB RAM to train the model. The model employs the AdamW [31] optimizer with parameters adjusted to a learning rate of 0.001, weight decay of 0.00004, and uses a cross-entropy loss function.

Furthermore, to assess the quality of the model, in this study, we utilize measures such as Accuracy, Precision, Recall, and F1-scores, described by formulas (8), (9), (10), and (11). The Accuracy measure often has drawbacks with imbalanced data. When it fails to reflect the actual effectiveness of the model, the accuracy result can make the model appear to perform well while it may simply predict the class with more data, and vice versa. Therefore, employing additional measures like recall and precision is crucial for an objective evaluation of the model.

$$\text{Accuracy (acc)} = \frac{\text{TP} + \text{TN}}{\text{TP+FN} + \text{TN} + \text{FP}} \quad (8)$$

$$\text{Precision (P)} = \frac{\text{TP}}{\text{TP+FP}} \quad (9)$$

$$\text{Recall (R)} = \frac{\text{TP}}{\text{TP+FN}} \quad (10)$$

$$\text{F1} - \text{score} = \frac{2 \text{ x } (\text{P x R})}{(\text{P} + \text{R})} \quad (11)$$

TP (True Positive): The count of correctly predicted positive cases.

TN (True Negative): The count of correctly predicted negative cases.

FP (False Positive): The count of instances where observations labeled as negative are incorrectly predicted as positive.

FN (False Negative): The count of instances where observations labeled as positive are incorrectly predicted as negative.

Table 6. Our performance on the FER-2013 testing set

Models	Acc	R	P	F1 score
ResNet18	72.67%	72.30%	72.30%	72.30%
VGG19	72.64%	72.60%	72.70%	72.60%

Table 7. Our performance on the CK+ testing set

Models	Acc	R	P	F1 score
ResNet18	96.97%	96.7%	96.50%	96.60%
VGG19	95.96%	96.60%	96.60%	96.60%

5.1 Results of facial emotion recognition

We have conducted experiments on facial emotion recognition using two different datasets: FER-2013 and CK+. The results of these experiments are presented in Tables 6 and Table 7. For the FER-2013 dataset, the results indicated that the ResNet18 model achieved an accuracy of 72.67%, with Recall, Precision, and F1-score all at 72.30%. Meanwhile, the VGG19 model also produced similar outcomes with an accuracy of 72.64%, Recall at 72.60%, Precision at 72.70%, and an F1-score of 72.60%.

Subsequently, we extended our experiments to the CK+ dataset, known for its accuracy and diversity in emotional expression. On this dataset, both the ResNet18 and VGG19 models showed significant improvements in performance. Specifically, ResNet18 reached an accuracy of up to 96.97%, with Recall at 96.7%, Precision at 96.50%, and an F1-score of 96.60%. The VGG19 model was similarly impressive, with an accuracy of 95.96%, Recall at 96.60%, Precision, and F1-score all at 96.60%.

These results demonstrate the superior ability of the ResNet18 and VGG19 models in recognizing facial emotions across both datasets. The difference in performance between the two datasets also reflects their distinct characteristics and the level of challenge they pose for emotion recognition tasks. While the FER-2013 dataset provides a comprehensive challenge with its diversity of emotional expressions, CK+ with its better-structured data and higher quality images has allowed these models to achieve higher accuracy in emotion recognition, emphasizing the importance of quality data in training deep learning models.

Notably, the emotion Happiness was the most accurately predicted, a trend consistent with previous findings and likely due to the abundance of Happiness emotion data in the datasets. This abundance helps the models to better predict this specific emotion. Otherwise, the emotion Disgust, which was the least accurately predicted in the experiments, remains challenging to predict accurately, possibly due to the limited amount of data available for this emotion, making it difficult for the models to learn its distinctive features. This particular insight is further detailed in Fig. 8, which visualizes the predictive performance of each emotion, highlighting the disparities in model accuracy across different emotional expressions.

Table 8. Experimental results of SVM and LDA models with 11 features

Models	Acc	P	R	F1 score
SVM	82.61%	81.16%	80.66%	80.57%
LDA	78.98%	77.09%	76.91%	76.72%

Table 9. Experimental results of SVM and LDA models with 9 features

Models	Acc	P	R	F1 score
SVM	86.96%	85.72%	84.63%	84.54%
LDA	79.00%	77.09%	76.91%	76.72%

Table 10. Performance evaluation table of SVM model

Mood	Acc	P	R	F1 score
Calm	99.27%	97.50%	100.00%	98.73%
Energetic	88.41%	70.27%	83.87%	76.47%
Happy	88.41%	80.00%	57.14%	66.67%
Sad	97.82%	95.12%	97.50%	96.29%
Average	93.48	85.72%	84.63%	84.54%

Table 11. Performance evaluation table of LDA model

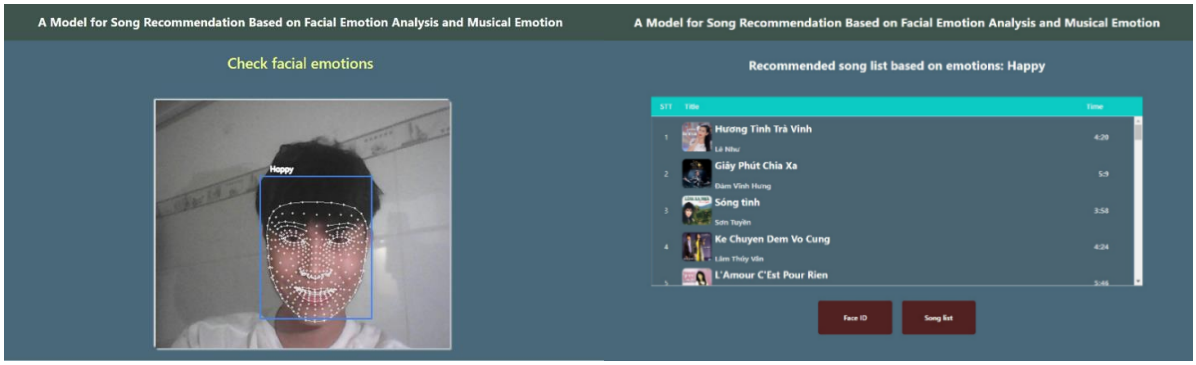
Mood	Acc	P	R	F1 score
Calm	97.83%	95.00%	97.44%	96.20%
Energetic	83.34%	65.38%	54.84%	59.65%
Happy	82.61%	55.88%	67.86%	61.29%
Sad	94.20%	92.11%	87.50%	89.74%
Average	89.49%	77.09%	76.91%	76.72%

5.2 Results of music emotion recognition

In our research, we use 11 music features to train the music mood recognition model using SVM and LDA techniques, with the results presented in Table 8, and Fig. 6. After carefully reviewing the dataset and experimental outcomes, we noticed that two of the features “Key” and “Time_signature” didn't really help us figure out the emotions in the music. So, we decided to try again without these two features and see if the model got better at predicting the mood. Therefore, we continued to experiment with

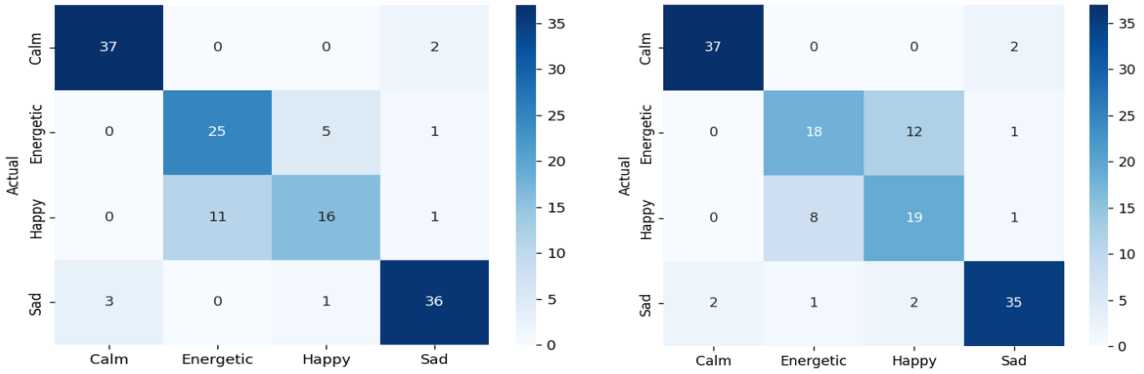
the models by removing the two features “Key” and “Time_signature”. The results shown in Table 9 indicate that the model predicts more accurately compared to using 11 features (including “Key” and “Time_signature”). Moreover, when we compared the SVM and LDA techniques, we found that the SVM method was more accurate than LDA in both the original and the new tests. These comparisons are shown in Table 10, Table 11, Fig. 6, and Fig. 7.

Based on the results from the SVM and LDA models, a visual overview is provided through the metrics of Accuracy, Precision, Recall, and F1 score, used to evaluate the overall performance of the model. Notably, the SVM model shows particularly impressive performance in recognizing the Calm mood with an Accuracy of 99.27% and an F1 score of 98.73%, indicating the model's exceptionally high ability to accurately classify music with a calm mood. This may be due to the clear representation of the Calm mood's musical features, facilitating the SVM model's distinction. Both models face difficulties in recognizing Energetic and Happy moods, with the SVM model having an F1 score for Energetic of 76.47% and for Happy of 66.67%.



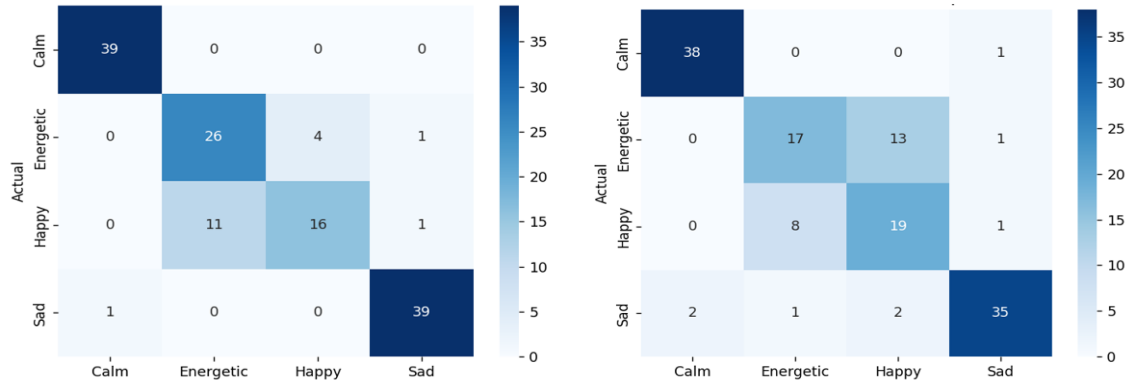
(a) (b)

Figure. 5 A music listening website system integrated with emotion-based suggestions



(a) (b)

Figure. 6 Confusion matrix of 11 feature: (a) SVM model and (b) LDA model



(a) (b)

Figure. 7 Confusion matrix of 9 feature: (a) SVM model and (b) LDA model

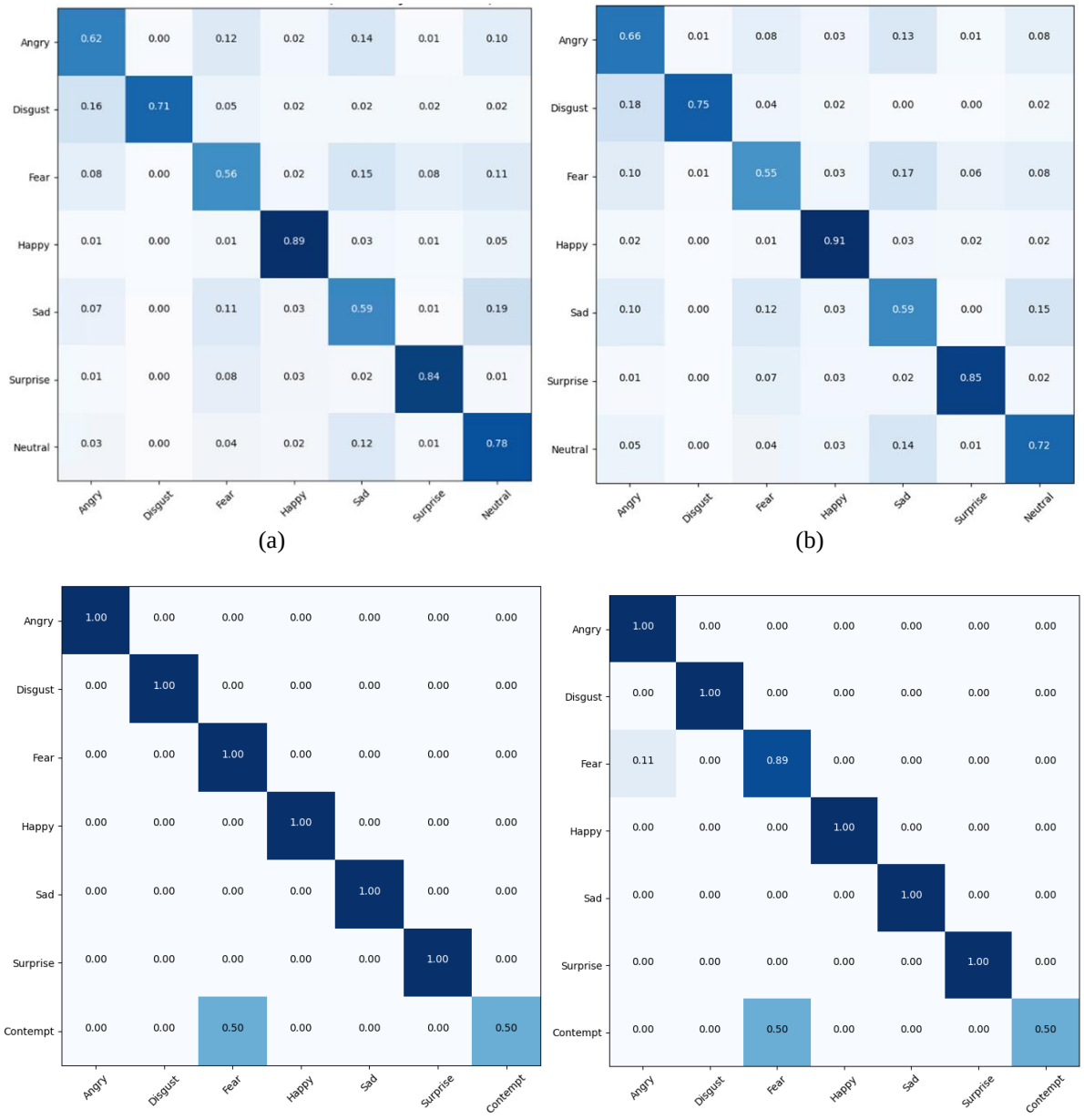


Figure. 8 Confusion matrix: (a) ResNet18 on FER2013, (b) VGG19 on FER2013, (c) ResNet18 on CK+ and (d) VGG19 on CK+

This may reflect the complexity and diversity of these moods in music, as well as their musical overlap with other moods. While the LDA model also shows good recognition capabilities for the Calm mood with an Accuracy of 97.83% and an F1 score of 96.20%, its performance is lower for the Energetic and Happy moods compared to the SVM model. This indicates that the LDA model might not be as effective as the SVM in handling the complexity and overlap of musical features between moods. Both models demonstrate stability and high performance in recognizing the Sad mood, with SVM having an F1 score of 96.29% and LDA 89.74%, indicating that the distinct characteristics of this mood in music can be effectively learned and recognized by the models.

5.3 Build a song recommendation application

To evaluate the proposed model, we have developed a web application that suggests music based on the user's facial emotions. Fig. 5 visually illustrates the application testing results step by step:

In Fig. 5 (a), at the application's interface screen, users will see a camera activated, and their task is to position their face within the camera frame so that the system can recognize the face. Afterward, the system will scan the user's face, and the model will analyze

and display their current emotion within a specific time frame of 10 seconds. The system will display the final emotion result screen after scanning the user's face. At this point, they have two options: one is to scan their face again; the other is to select "SongList". In Fig. 5 (b), when the user selects the SongList button, the system queries the emotion-based music database using the proposed set of rules to find songs that match the user's detected facial emotions. The user can then choose their preferred song from the suggested list to enjoy. The system displays the music player for the selected song. Here, users can listen to the music.

5.4 Discussion

In this paper, we achieved some notable results. Specifically, the facial emotion recognition models using VGG19 and ResNet18 demonstrated good performance when evaluated on the FER-2013 and CK+ datasets. Additionally, the music emotion recognition model on the Spotify dataset also achieved high accuracy when experimented with the SVM model.

A music recommendation application based on facial emotion and music emotion has been developed and tested with a user-friendly interface. This application allows users to use their camera to scan their face and receive a playlist that matches their current mood from the Spotify music database.

Compared to previous studies, this research has contributed to enhancing performance by applying modern deep learning models and proposing a playlist based on rules derived from psychological factors. The system directly analyzes data from the Spotify music platform, extracting musical features to be classified by the SVM model. Therefore, the suggested songs are directly taken from the Spotify database, ensuring they are up-to-date and diverse.

Previous studies, such as Athavle et al. [15], have shown facial emotion recognition performance reaching 71% when using deep learning models on the FER-2013 dataset. In contrast, by utilizing advanced CNN models like VGG19 and ResNet18, we achieved higher accuracy on the same FER-2013 dataset with an accuracy of 72.67%. Their study also implemented a music-playing model based on facial emotion, but it did not integrate as many types of music emotions and as rich a dataset as the current study. Similarly, in the research by A. Mahadik [23], a music recommendation system based on emotion was proposed. However, the playlist recommendation method suggested by the author was solely based on facial emotion. Meanwhile, the method proposed in this study has built a set of rules to establish the relationship between facial emotion and music emotion to recommend playlists that suit the user's emotions.

In this study, there are some limitations, including uneven data in terms of quantity and quality, which can affect the model's performance. Environmental factors such as lighting conditions and camera quality can significantly impact the accuracy of the facial emotion recognition model. Experimental results have also shown that the performance of our facial recognition model is lower compared to the study by Y. Khaireddin [6], with an accuracy of 73.28% on the FER-2013 dataset. Additionally, our system lacks personalization as the current system does not consider individual factors such as age, gender, and the user's listening history, which could help improve the accuracy and personalization of music recommendations.

6. Conclusion

In this paper, we proposed a model to recommend playlists by analyzing the listener's current emotions through facial expressions. We utilize advanced pre-trained models, including ResNet18 and VGG19, to accurately detect and analyze emotions such as happiness, anger, sadness, surprise, disgust, and neutrality. By integrating these models with the Facial Emotion Recognition (FER) 2013 and The Extended Cohn-Kanade Dataset (CK+), we achieve precise emotion detection. Additionally, we employ SVM and LDA models to analyze musical emotions in the Spotify Music dataset and propose a set of rules linking facial expressions to musical emotions for playlist recommendations. Furthermore, we develop a user-friendly web application that allows users to capture facial expressions through their device's camera and receive personalized song recommendations based on their mood. Our analysis of musical emotions using SVM on the Spotify Music dataset yields promising results. The ResNet18 model achieves 72.67% accuracy on FER2013 and 96.97% accuracy on CK+, while VGG19 achieves 95.96% accuracy on CK+. Additionally, our music emotion recognition approach with SVM achieves an 86.96% accuracy rate. These findings demonstrate the effectiveness of our model in accurately recognizing both facial and musical emotions. In the future, we will explore models like Vision - LLM, Graph Attention Networks, and knowledge graphs to improve the accuracy of the proposed model while also ensuring speed in real-world deployment. Combining user behavior and emotions to create a system that better understands users' desires when experiencing the application.

Conflicts of Interest

The authors declare no competing interests relevant to this content article.

Author Contributions

H.N., N.T initiated the research topic, proposed the main ideas, and designed the model. Y.T, A.T, A.H, HV developed the mathematical model, and contributed to the results analysis. D.L builds model music emotion and builds a website to suggest song lists. All authors collaborated on writing and reviewing the manuscript.

References

[1] T. Jiang, S. Deng, P. Wu, and H. Jiang, “Real-Time Human-Music Emotional Interaction Based on Deep Learning and Multimodal Sentiment Analysis”, *Wireless Communications and Mobile Computing*, 2023.

[2] S. M. Florence và M. Uma, “Emotional Detection and Music Recommendation System Based on User Facial Expression”, *IOP Conference Series: Materials Science and Engineering*, Vol. 912, No. 6, p. 062007, 2020.

[3] F. Z. Canal, T. R. Müller, J. C. Matias, G. G. Scotton, A. R. de Sa Junior, E. Pozzebon, and A. C. Sobieranski, “A survey on facial emotion recognition techniques: A state-of-the-art literature review”, *Information Sciences*, Vol. 582, pp. 593-617, 2022.

[4] N. Ahmed, Z. Al Aghbari, and S. Girija, “A systematic survey on multimodal emotion recognition using learning algorithms”, *Intelligent Systems with Applications*, Vol. 17, 200171, 2023.

[5] K. Ngo, T. Nguyen, N. Tran, T. Lam, T. Nguyen, H. Nguyen, “Emotion Recognition Method based on Multimodal Fusion using Sequence Images and Audio”, *Int. J. Adv. Engr*, Vol. 5, No. 1, pp. 33-41, 2022.

[6] Y. Khaireddin and Z. Chen, “Facial emotion recognition: State of the art performance on FER2013”, *arXiv preprint arXiv:2105.03588*, 2021.

[7] S. Li and W. Deng, “Deep facial expression recognition: A survey”, *IEEE Transactions on Affective Computing*, Vol. 13, No. 3, pp. 1195-1215, 2020.

[8] H. C. Chen and A. L. Chen, "A music recommendation system based on music data grouping and user interests”, In: *Proc. of the Tenth International Conference on Information and Knowledge Management*, pp. 231-238, 2001.

[9] H. S. Park, J. O. Yoo, and S. B. Cho, “A context-aware music recommendation system using fuzzy Bayesian networks with utility theory”, In: *Proc. of the Third International Conference on Fuzzy Systems and Knowledge Discovery*, Vol. 3, pp. 970-979, 2006.

[10] M. Athavle, D. Mudale, U. Shrivastav, and M. Gupta, “Music recommendation based on face emotion recognition”, *Journal of Informatics Electrical and Electronics Engineering (JIEEE)*, Vol. 2, No. 2, pp. 1-11, 2021.

[11] H. Nguyen, K. Kotani, F. Chen, and B. Le, “A thermal facial emotion database and its analysis”, *Image and Video Technology*, pp. 397-408, 2014.

[12] M. J. A. Dujaili, “Survey on facial expressions recognition: databases, features and classification schemes”, *Multimedia Tools and Applications*, Vol. 83, No. 3, pp. 7457-7478, 2024.

[13] N. Ahmed, Z. Al Aghbari, and S. Girija “A systematic survey on multimodal emotion recognition using learning algorithms”, *Intelligent Systems with Applications*, Vol. 17, No. 200171, 2023.

[14] Y. P. Lin, C. H. Wang, T. P. Jung, T. L. Wu, S. K. Jeng, J. R. Duann, and J. H. Chen, “EEG-based emotion recognition in music listening”, *IEEE Transactions on Biomedical Engineering*, Vol. 57, No. 7, pp. 1798-1806, 2010.

[15] M. Athavle, D. Mudale, U. Shrivastav, and M. Gupta, "Music recommendation based on face emotion recognition", *Journal of Informatics Electrical and Electronics Engineering (JIEEE)*, Vol. 2, No. 2, pp. 1-11, 2021, doi: 10.54060/JIEEE/002.02.018.

[16] L. Singh, S. Singh, N. Aggarwal, R. Singh, and G. Singla, “An efficient temporal feature aggregation of audio-video signals for human emotion

recognition", In: *Proc. of the International Conference on Signal Processing*, Computing and Control, pp. 660-668, 2021.

[17] A. Gupte, A. Naganarayanan, and M. Krishnan, "Emotion based music player-xbeats", *International Journal of Advanced Engineering Research and Science*, Vol. 3, No. 9, pp. 236854, 2016.

[18] J. S. Preema, S. M. Rajashree, H. Savitri, and S. J. Shruthi, "Review on facial expression-based music player", *International Journal of Engineering Research & Technology (IJERT) ICRTT*, Vol. 6, No. 15, 2018.

[19] H. I. James, J. J. A. Arnold, J. M. M. Ruban, M. Tamilarasan, and R. Saranya, "Emotion based music recommendation system", *Emotion*, Vol. 6, No. 03, 2019.

[20] G. Dhand, T. Beri, T. Sobti, and V. Angrish, "Music Recommendation Using Sentiment Analysis from Facial Recognition", In: *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*, 2022.

[21] P. Hemanth, A. C. Adarsh, P. Ajith, and V. A. Kumar, "EMO player: emotion based music player", *International Research Journal of Engineering and Technology (IRJET)*, Vol. 5, No. 4, pp. 4822-4827, 2018.

[22] A. Abdul, J. Chen, H. Y. Liao, and S. H. Chang, "An emotion-aware personalized music recommendation system using a convolutional neural networks approach", *Applied Sciences*, Vol. 8, No. 7, 1103, 2018.

[23] A. Mahadik, S. Milgir, J. Patel, V. B. Jagan, and V. Kavathekar, "Mood based music recommendation system", *International Jounal of Engineering Research and Technology (IJERT)*, Vol. 10, 2021.

[24] D. S. Chakrapani, S. Iram, S. R. Bhat, Supritha Agni, L. Supritha, and S. Leelavathi, "Music recommendation based on facial emotion recognition", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 12, No. 4, 2023.

[25] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al., "Challenges in representation learning: A report on three machine learning contests," *International conference on neural information processing*, pp. 117–124, Springer, 2013.

[26] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression", In: *Proc. of 2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, pp. 94–101, IEEE, 2010.

[27] MusicBlogger, "Spotify music data to identify the moods [Data set]", *Kaggle*, 2023. [Online]. Available: https://www.kaggle.com/datasets/musicblogger/spotify-music-data-to-identify-the-moods. [Accessed: Mar. 13, 2024].

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", *arXiv preprint arXiv:1409.1556*, 2014.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.

[30] D. Han, Y. Kong, J. Han, and G. Wang, "A survey of music emotion recognition", *Frontiers of Computer Science*, Vol. 16, No. 6, pp. 166335, 2022.

[31] C. Cortes and V. Vapnik, "Support-vector network", *Machine Learning*, Vol. 20, No. 3, pp. 273-297, 1995.

[32] P. Xanthopoulos, P. M. Pardalos, and T. B. Trafalis, "Linear discriminant analysis", In: *Robust Data Mining, Springer*, pp. 27–33, 2013.

[33] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization", *arXiv preprint arXiv:1711.05101*, 2017.