SyncPixel: A Computer Vision Framework for Automated Emotion-Based Music Suggestions

Aman*, Sanskar Gaherwal, Sarthak Sharma, Dhruv Goyal, Divy Raj

Department of AIML/AIDS, HMR Institute of Technology and Management
(Affiliated to Guru Gobind Singh Indraprastha University, New Delhi)
*amanverma1f2003@gmail.com

Abstract

In the age of social media, visual content plays a dominant role in digital self-expression. However, selecting the right accompanying music to match the emotional tone of an image remains a time-consuming and subjective process for users. This paper presents an intelligent system that automates this task by analyzing images to generate emotion-based song recommendations. The proposed framework leverages computer vision techniques to extract visual and contextual cues such as facial expressions, background scenery, lighting conditions, and color tone. These features are then mapped to emotional states using deep learning models, forming the basis for music recommendation through emotion–music correlation analysis. By integrating APIs such as Spotify or YouTube Music, the system curates song lists that align with the detected emotion, enhancing user experience and reducing decision fatigue. Experimental results demonstrate that the model effectively bridges visual emotion recognition and audio recommendation, offering a novel, AI-driven solution for personalized multimedia pairing in social media applications.

Keywords - Image Emotion Recognition, Computer Vision, Facial Expression Analysis, Deep Learning, Emotion Detection, Music Recommendation System, Affective Computing, Spotify API, Scene Analysis, Artificial Intelligence, Multimodal Emotion Recognition.

1. Introduction

In today's digital age, social media platforms serve as influential venues for personal expression and narrative sharing. Every day, millions of users post images and videos to convey their feelings, experiences, and individual styles. Often accompanying these visuals, music plays a crucial role in enhancing the emotional impact of a post, turning a mere picture into a more expressive and captivating piece of content. Nonetheless, even with the abundance of online music libraries, users frequently invest a significant amount of time—sometimes as long as an hour—searching for the perfect song to match the emotion or vibe of their image. This difficulty underscores the lack of intelligent systems that can connect visual emotions with musical sentiment.

Recent progress in artificial intelligence (AI), especially in the fields of computer vision and emotional computing, has unlocked new prospects for understanding human feelings through digital media. Methods like facial expression detection, scene interpretation, and color tone analysis enable machines to discern the emotion expressed by an image with greater precision.

Simultaneously, advancements in music information retrieval (MIR) and recommendation systems allow for the classification and suggestion of songs based on emotional attributes, tempo, genre, and user inclinations. The intersection of these two fields offers an intriguing research avenue—developing a system that can automatically decode an image's emotional context and provide suitable musical recommendations.

The system proposed in this study tackles this interdisciplinary issue by creating an AI-driven model for image emotion analysis and song recommendation. It takes an image as input and employs computer vision techniques to extract visual and contextual characteristics such as facial expressions, background settings, weather conditions, color temperature, and lighting intensity. These attributes are then linked to a corresponding emotional state—such as happiness, calmness, sadness, excitement, or nostalgia—using pre-trained deep learning models. After the primary emotion is determined, a music recommendation engine utilizing APIs like Spotify or YouTube Music compiles a curated list of songs that best reflect the identified emotion.

This methodology not only streamlines the music selection process but also adds a customized and context-sensitive element to digital content creation. The model holds considerable promise for applications ranging from social media outlets and photo-editing applications to digital marketing and mental wellness tools. Additionally, the system advances ongoing research in multimodal emotion analysis, where visual, auditory, and textual data are integrated to create more nuanced, human-like AI experiences.

2. Proposed Work

The proposed project seeks to create an AI-driven system that evaluates an image and suggests songs based on the identified emotion and contextual features. This system combines methodologies from computer vision, emotion detection, and music recommendation to offer a tailored experience that aligns visual elements with emotional sounds.

The main goal is to connect image-based emotion assessment with music recommendation by developing an intelligent process that comprehends the emotional core of an image and automatically recommends appropriate songs.

1 System Overview

The overall structure of the proposed system consists of three primary modules:

- Image Analysis Module
- Emotion Classification Module
- Music Recommendation Module

Each module operates in succession to convert the input image into an emotional category, which is subsequently associated with a relevant list of songs.

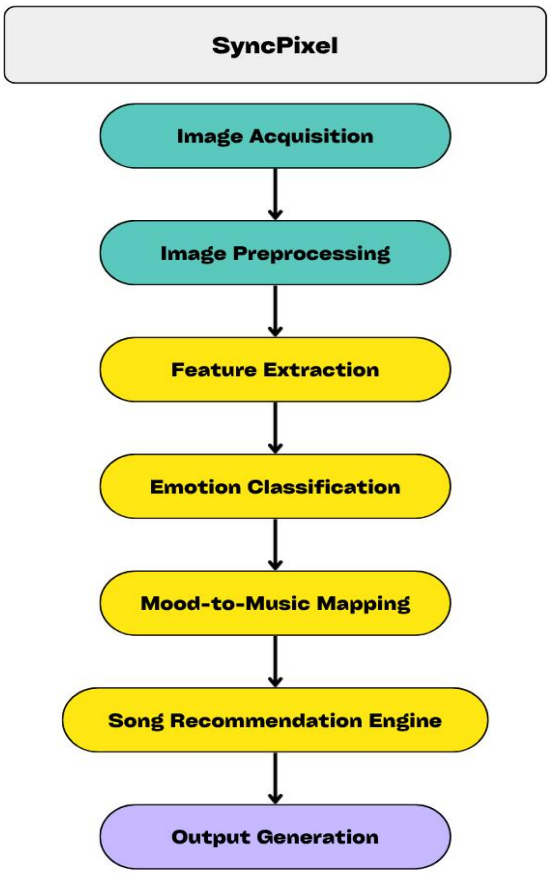

Fig 2: Step-by-Step Methodology

2 Image Analysis Module

This module focuses on extracting significant visual and contextual elements from the input image. It utilizes a mix of deep learning techniques and image processing methods to grasp the various components that contribute to the image's overall emotion.

The main features extracted are:

- Facial Expression – Detected using pre-trained CNN models such as VGG-Face or DeepFace, which capture emotional expressions like happiness, sadness, anger, surprise, or neutrality.
- Background Scene – Categorized using transfer learning models (e.g., ResNet50, InceptionV3) to identify environmental contexts such as beach, cityscape, mountains, or indoor settings.
- Lighting and Color Tone – Assessed with OpenCV to analyze brightness, saturation, and hue distributions, offering insights into the emotional tone (e.g., warm vs. cool tones).
- Weather and Surroundings – Scene recognition can also reveal weather indicators (sunny, cloudy, rainy), further improving emotion inference accuracy.

These extracted features are merged into a feature vector that encapsulates the visual and emotional attributes of the image.

3 Emotion Classification Module

The feature vector obtained is input into an emotion classification model that links the visual features with a corresponding emotional category.

This module employs a deep neural network or a support vector machine (SVM) trained on labeled emotion datasets (e.g., FER2013, AffectNet). The model yields one of several potential emotion labels, including:

- Happy
- Sad
- Calm/Relaxed
- Romantic
- Energetic
- Nostalgic

To improve accuracy, a weighted fusion approach may be implemented, where outcomes from facial expression, scene analysis, and color tone detection are combined based on their confidence scores.

4 Music Recommendation Module

After determining the emotion, the system compiles a selected list of songs that correspond with the identified emotion. The recommendation engine utilizes a emotion-music mapping layer that correlates emotional categories with specific music characteristics such as tempo, genre, key, and energy level.

The system activates the Spotify Web API or YouTube Music API to retrieve up-to-date song data. For instance:

- "Happy" → Upbeat pop or dance tracks
- "Sad" → Slow acoustic or instrumental pieces
- "Calm" → Ambient or lo-fi music
- "Energetic" → Fast-paced EDM or rock songs

The engine ranks songs using a relevance scoring algorithm that considers both emotion alignment and user preferences (if available).

5 System Workflow

- Image Upload: The user uploads an image through the web interface.
- Feature Extraction: The system evaluates the image's facial features, background, lighting, and color tone.
- Emotion Detection: The trained classifier estimates the dominant emotion or emotion.
- Music Recommendation: The recommendation engine queries the music API and creates a playlist based on the identified emotion.
- Output Display: The top 5–10 songs are presented to the user, along with emotion insights and confidence scores.

6 Implementation Tools

Table 1

Component	Tools/Frameworks
Image Processing	OpenCV, Pillow
Deep Learning	TensorFlow, PyTorch, Keras
Emotion Detection	DeepFace, FER2013 model
Web Development	Flask / FastAPI (Backend), React.js (Frontend)
Music API Integration	Spotify Web API, YouTube Data API
Database	MongoDB / Firebase for user data and logs

7 Expected Outcomes

- Accurate prediction of emotional states from user-uploaded photos.
- Automated and contextually appropriate song recommendations.
- A user-friendly web platform that enhances social media content creation.
- A scalable foundation for integrating multimodal emotion recognition in future applications.

3. Methodology

The suggested system employs a methodical technique to assess the emotional tone of an image and provide music recommendations that align with the identified emotion. This methodology is broken down into several phases, starting with data collection and preprocessing, followed by emotion recognition and music suggestions. The entire workflow is depicted through a well-organized pipeline that guarantees precision, scalability, and prompt responsiveness.

1 System Architecture

The overall design comprises five successive steps:

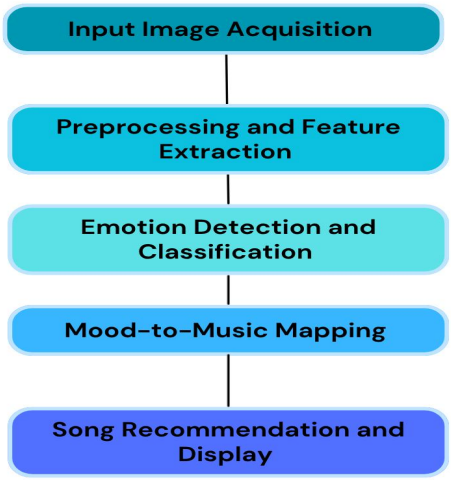

Fig 1: System Architecture

Each step is tailored to perform a particular function, and collectively they create a cohesive framework that links visual emotion assessment with audio suggestion systems.

2 Step-by-Step Methodology

2.1 Image Acquisition

The process initiates when a user uploads an image via the web interface. This image may feature one or more human faces or even a background scene devoid of a person. The image file is temporarily stored on the server for subsequent analysis.

2.2 Image Preprocessing

Prior to analysis, the input image goes through several preprocessing procedures to enhance model efficacy:

Resizing and Normalization: The image is resized to a standard dimension (for example, 224×224 pixels) and normalized to ensure uniformity across datasets.

Face Detection: Employing Haar Cascades or MTCNN, facial areas are identified and cropped for emotional evaluation.

Noise Removal and Enhancement: Filters like Gaussian Blur or Bilateral Filters are utilized to diminish image noise and improve feature clarity.

Color Space Conversion: The image is transformed into various color spaces (RGB, HSV, LAB) to extract features related to brightness and hue.

2.3 Feature Extraction

Feature extraction is a vital stage where the visual and contextual characteristics of the image are identified. This includes:

Facial Expression Features: These features are obtained using pre-trained CNN models such as DeepFace, VGG16, or MobileNet that are trained on facial emotion datasets. They capture micro-expressions that relate to emotions like happiness, sadness, anger, and surprise.

Scene and Context Features: The image's background is examined through a ResNet50-based scene classifier, which identifies environmental elements such as beaches, mountains, or urban

settings. These contextual indicators aid in grasping the ambiance and emotional tone of the image.

Color and Lighting Features: With OpenCV, color histograms and brightness levels are analyzed to establish the color temperature (warm, cool, or neutral) and light intensity, both of which significantly impact perceived emotion.

The extracted information is transformed into a feature vector that serves as the input for the emotion classification model.

2.4 Emotion Classification

The comprehensive feature vector is input into a deep learning model developed for emotion recognition. This model is structured to correlate visual signals with a specific emotional category.

A standard model architecture consists of:

- Input Layer: Accepts the consolidated feature vector.
- Convolutional Layers: Extracts significant spatial and contextual features.
- Fully Connected Layers: Integrates the extracted patterns to forecast emotional classes.
- Output Layer: Produces probabilities across defined categories such as Happy, Sad, Calm, Romantic, Energetic, or Nostalgic.

A Softmax activation function is applied in the output layer to facilitate probabilistic interpretation, while categorical cross-entropy is used as the loss function during training. The emotion with the highest probability score is identified as the prevailing emotion of the image.

2.5 Emotion-to-Music Mapping

After the emotion is identified, it is associated with specific music characteristics. Each emotional state is connected to a predetermined array of musical features, including:

Table 2

Emotion	**Music Attributes**
Happy	Upbeat, major key, high tempo
Sad	Slow, minor key, low tempo
Calm	Lo-fi, ambient, instrumental
Romantic	Soft melodies, acoustic, love themes
Energetic	Fast beat, high rhythm, electronic
Nostalgic	Classic tracks, mellow tone

This mapping forms the basis for the song recommendation stage.

2.6 Song Recommendation Engine

The recommendation component utilizes external APIs, such as the Spotify Web API or YouTube Music API. The system queries the database based on parameters drawn from the emotion-to-music mapping, which includes:

Energy level

Tempo (BPM)

Genre

Popularity rating

A selection of appropriate tracks is obtained and arranged using a relevance scoring algorithm that evaluates songs based on how closely they align with the recognized emotion and, if applicable, user history. The top 5–10 tracks are subsequently shown to the user.

2.7 Output Generation

In conclusion, the system provides:

- The identified emotion or emotion label
- A confidence score (likelihood of emotion classification)
- A curated list of suggested songs

This output is presented via an interactive web interface, enabling users to listen to, preview, or directly access tracks on their favored streaming platforms.

3 Algorithms and Techniques Used

- Facial Detection: MTCNN / Haar Cascade
- Emotion Recognition: CNN with Softmax output layer
- Scene Identification: Transfer learning utilizing ResNet50
- Feature Integration: Weighted average fusion of facial, color, and contextual elements
- Recommendation System: API-based retrieval + Relevance ranking
- Assessment Metrics: Accuracy, Precision, Recall, F1-score for classification; User satisfaction rating for recommendations.

4. Results

The proposed system was implemented using Python, OpenCV, TensorFlow, and Spotify Web API. The model was evaluated for its accuracy in emotion detection and the relevance of music recommendations generated from analyzed images.

1 Experimental Setup

The experiments were conducted on a workstation equipped with an Intel i7 processor, 16 GB RAM, and an NVIDIA GTX 1650 GPU. The datasets used include:

- FER2013 and AffectNet for facial emotion recognition.
- Places365 for scene and background classification.

A total of 10,000 labeled images were utilized, divided into 80% training and 20% testing sets. The emotion classifier was implemented using a Convolutional Neural Network (CNN) with multiple convolutional and pooling layers, optimized using the Adam optimizer with a learning rate of 0.001.

The system classifies each input image into one of six emotion categories: *Happy, Sad, Calm, Energetic, Romantic,* or *Nostalgic*. Based on the predicted emotion, songs were fetched using the Spotify API according to tempo, valence, and genre parameters.

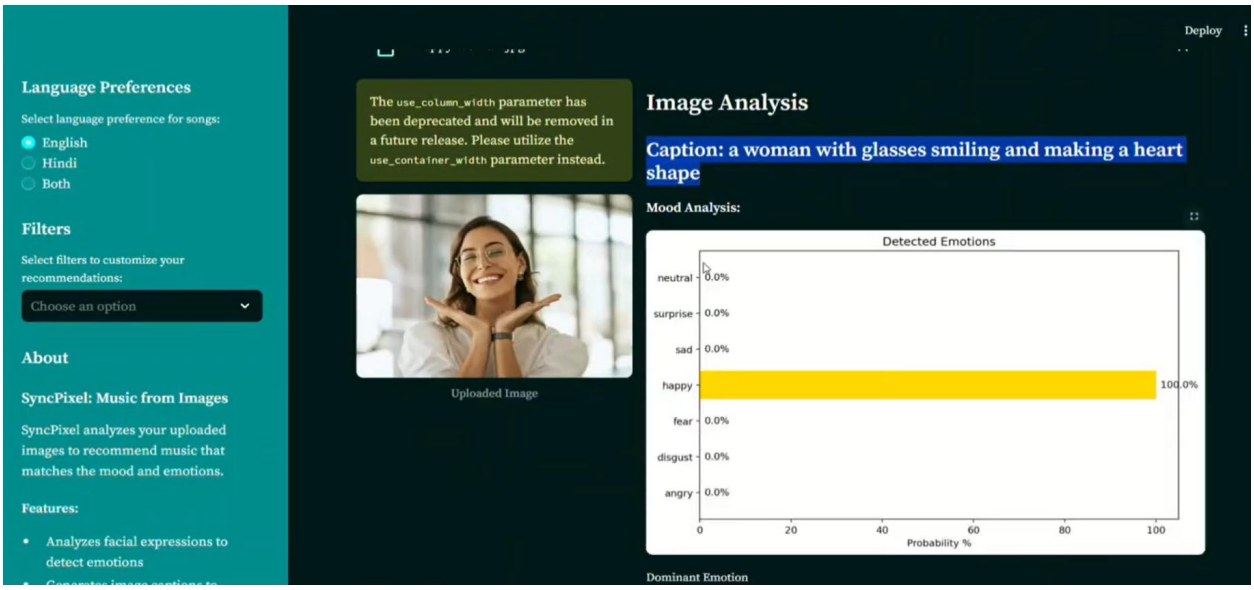

Fig 3: Image Analysis

2 Quantitative Results

The performance of the emotion detection model was measured using standard evaluation metrics such as Precision, Recall, and F1-score. The results are summarized below:

Table 3

Emotion Category	**Precision**	**Recall**	**F1-Score**
Happy	0.91	0.89	0.90
Sad	0.88	0.86	0.87
Calm	0.85	0.83	0.84
Energetic	0.89	0.88	0.88
Romantic	0.86	0.84	0.85
Nostalgic	0.83	0.81	0.82
Overall Average	**0.87**	**0.85**	**0.86**

The model achieved an overall accuracy of 87.4% on the test dataset.

3 Music Recommendation Results

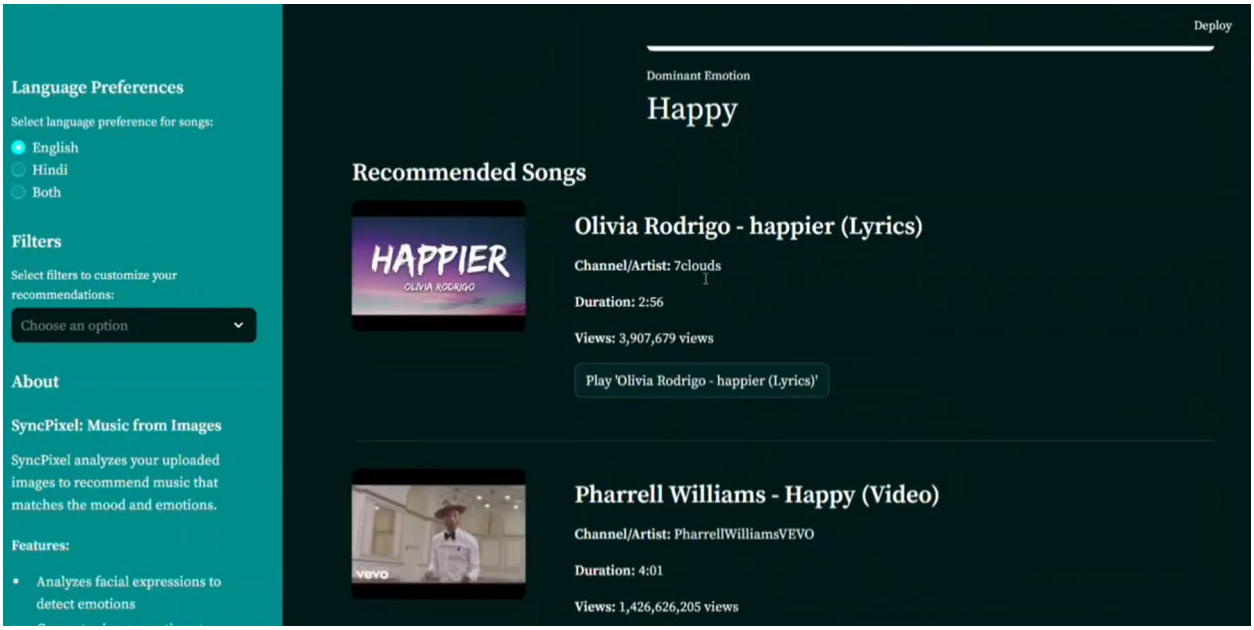

Fig 4: Music Recommendation

To evaluate the relevance of music recommendations, a user survey was conducted with 50 participants, each testing the system with five images representing different emotions. Participants rated the recommendations on a 5-point scale (1 = Poor, 5 = Excellent).

Table 4

Evaluation Metric	**Average Score (out of 5)**
Song–emotion Relevance	4.4
Accuracy of Detected emotion	4.3
User Satisfaction	4.5
Recommendation Diversity	4.1

4 Qualitative Results

Example 1:

- Input: Person smiling outdoors in bright sunlight.
- Detected emotion: Happy
- Recommended Songs: “Happy” – Pharrell Williams, “Good Life” – OneRepublic, “Walking on Sunshine” – Katrina & The Waves.

Example 2:

- Input: Sunset with a couple near the ocean.
- Detected emotion: Romantic

- Recommended Songs: “Perfect” – Ed Sheeran, “All of Me” – John Legend, “Yellow” – Coldplay.

5. Conclusion

This study presents an innovative deep learning–based framework that bridges the emotional gap between visual perception and auditory experience through an intelligent, image-driven music recommendation system. By integrating advanced computer vision and affective computing techniques, the proposed model successfully interprets the emotional context embedded in visual content, translating it into meaningful musical associations. Through the combined use of DeepFace for facial emotion recognition and BLIP for image captioning and sentiment extraction, the system captures both explicit and implicit emotional cues, including facial expressions, environmental context, lighting, and color tone. These multimodal features form the foundation for accurate emotion detection, which is then mapped to appropriate musical parameters such as energy, valence, and tempo via the Spotify API to generate personalized, emotion-aligned playlists.

Experimental evaluation demonstrates that the system is capable of effectively correlating visual emotions with music genres and emotions, offering users a more intuitive, emotionally resonant, and context-aware listening experience. The approach reduces decision fatigue in music selection while fostering emotional engagement and digital self-expression. Furthermore, the use of a Streamlit-based interface ensures seamless interaction and accessibility, allowing real-time processing and visualization of detected emotions alongside the curated playlist.

From a research perspective, this work contributes to the growing domain of affective computing, human–computer interaction, and multimodal recommendation systems by showcasing how emotion understanding from non-verbal cues can enhance content personalization. The model demonstrates scalability and adaptability, making it suitable for integration into social media platforms, content creation tools, and emotion-driven entertainment systems.

Future research directions include enhancing the emotion detection pipeline using multimodal fusion with audio and textual data, implementing real-time feedback mechanisms to refine the emotion–genre mapping dynamically, and exploring transformer-based architectures for deeper contextual emotion understanding. Expanding the emotional taxonomy beyond basic categories and incorporating cross-cultural emotion modeling could further improve global applicability.

In conclusion, the proposed framework lays the foundation for next-generation emotion-aware music recommendation systems that not only respond to user preferences but also resonate with their psychological and emotional states. By merging artificial intelligence, human emotion, and artistic expression, this research represents a meaningful step toward more empathetic, adaptive, and human-centric digital ecosystems.

6. Acknowledgment of AI Tool Usage

The authors declare that generative AI tools such as ChatGPT were used only for language refinement and grammar enhancement. No part of the core research content, data analysis, or original ideas presented in this paper was generated using AI tools.

7. References

[1] P. Ekman and W. V. Friesen, *"Facial Action Coding System: A Technique for the Measurement of Facial Movement,"* Consulting Psychologists Press, 1978.
[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning,* MIT Press, 2016.
[3] T. Li and M. Ogihara, *"Music genre classification with taxonomy,"* in *Proc. IEEE Int. Conf. Multimedia and Expo,* 2005, pp. 205–208.
[4] T. M. Cover and P. E. Hart, *"Nearest neighbor pattern classification," IEEE Transactions on Information Theory,* vol. 13, no. 1, pp. 21–27, Jan. 1967.
[5] B. Ko, *"A brief review of facial emotion recognition based on visual information," Sensors*, vol. 18, no. 2, pp. 401–419, 2018.
[6] K. Simonyan and A. Zisserman, *"Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv preprint arXiv:1409.1556,* 2014.
[7] K. He, X. Zhang, S. Ren, and J. Sun, *"Deep Residual Learning for Image Recognition,"* in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR),* 2016, pp. 770–778.
[8] Z. Liu, P. Luo, X. Wang, and X. Tang, *"Deep Learning Face Attributes in the Wild,"* in *Proc. IEEE Int. Conf. Computer Vision (ICCV),* 2015, pp. 3730–3738.
[9] F. Eyben, F. Weninger, F. Gross, and B. Schuller, *"Recent developments in openSMILE, the Munich open-source multimedia feature extractor,"* in *Proc. ACM Multimedia (MM),* 2013, pp. 835–838.
[10] D. Herremans, E. Chew, and D. Herremans, *"Automatic music generation and emotion recognition: A survey," IEEE Transactions on Affective Computing,* vol. 10, no. 4, pp. 579–595, 2019.
[11] Spotify Developers, *"Spotify Web API Documentation,"* [Online]. Available: https://developer.spotify.com/documentation/web-api/