International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-7795

Analyzing Emotional Responses to Music

BHARATHVAJ G
Department of Artificial Intelligence and Machine Learning
Bamnari Amman Institute of Technology
Sathyamangalam,Tamilnadu, India

SAKTHI GANISHKA M
Department of Artificial Intelligence and Machine Learning
Bamnari Amman Institute of Technology
Sathyamangalam,Tamilnadu, India

SASHITHRA R
Department of Artificial Intelligence and Machine Learning
Bamnari Amman Institute of Technology
Sathyamangalam,Tamilnadu, India

DHANUSHKA N
Department of Artificial Intelligence and Machine Learning
Bamnari Amman Institute of Technology
Sathyamangalam,Tamilnadu, India

Abstract **— Music plays a vital role in evoking emotional responses, making it a valuable tool in therapeutic settings, particularly for individuals with mental health challenges. This paper presents a system designed to detect and analyze users' emotional responses to music using machine learning and facial expression recognition techniques. The system classifies emotions such as happiness, sadness, anger, and relaxation by integrating audio feature extraction and real-time facial expression analysis. By combining both modalities, the system provides more accurate and objective emotional insights during music therapy sessions. The primary goal is to assist therapists by offering real-time feedback on patients' emotional states, allowing for personalized music therapy interventions. Audio features such as Mel Frequency Cepstral Coefficients (MFCC), chroma, and tempo are extracted from music, while facial expressions are analyzed using Convolutional Neural Networks (CNNs). The system provides therapists with continuous emotional insights, helping to tailor treatment plans based on the emotional impact of various music genres and styles. The system demonstrates high accuracy in emotion detection, with experimental results achieving an overall accuracy of 87%. This approach not only enhances therapeutic outcomes but also offers broader applications in mental health care, stress management, and personalized music recommendations. By objectively analyzing emotional responses to music, this system contributes to the growing field of affective computing and its role in mental health and wellness.**

Keywords— Emotion recognition, facial expressions, music emotion analysis, machine learning, deep learning.

I. INTRODUCTION

Music has long been recognized for its profound ability to influence and evoke emotions. From ancient times to modern-day therapeutic practices, music has been used as a means of emotional expression, emotional regulation, and even emotional healing. This ability of music to evoke emotional responses makes it a powerful tool in therapeutic contexts, particularly for individuals dealing with mental health challenges such as depression, anxiety, post-traumatic stress disorder (PTSD), and other emotional disturbances. However, while the impact of music on emotional states is widely accepted, objectively quantifying and analyzing these emotional responses remains a challenge, especially in real-time therapeutic settings.

A. The Role of Music in Therapy

Music therapy has become an integral part of mental health care, focusing on using music to achieve non-musical goals such as emotional healing, improved self-awareness, enhanced communication skills, and better mental well-being. Studies have shown that music therapy can significantly improve the mood, stress levels, and overall emotional states of individuals suffering from mental health disorders. It is particularly useful for patients who may struggle with verbal communication or those who have difficulty expressing their emotions through traditional forms of therapy.

However, one of the major challenges in music therapy is the subjective nature of measuring emotional responses. Currently, therapists often rely on patient self-reports or observational assessments to gauge the effectiveness of the music on the patient's emotional state. While these methods are useful, they are prone to subjectivity and may not accurately reflect the patient's internal emotional experience. Additionally, patients may not always be able to articulate their emotions clearly, particularly in cases where cognitive or emotional impairments are present. This limitation highlights the need for more objective, data-driven methods to analyze emotional responses in music therapy sessions.

B. Objective Emotion Recognition in Therapy

Advances in artificial intelligence (AI) and machine learning (ML) have opened new possibilities for

developing systems that can objectively detect and analyze human emotions. Emotion recognition, a subfield of affective computing, involves using AI techniques to identify human emotions based on physiological signals, facial expressions, speech, and other behavioral cues. Among these, facial expression analysis has proven to be a highly effective method for emotion detection, as facial expressions are closely linked to emotional states and can be measured in real-time without being intrusive.

In recent years, machine learning algorithms such as Convolutional Neural Networks (CNNs) have demonstrated significant success in classifying emotions based on facial expressions. These models can be trained on large datasets of facial images, labeled with corresponding emotions, to automatically detect emotional states such as happiness, sadness, anger, and fear. When integrated with real-time data processing, these algorithms provide a powerful tool for identifying emotions objectively during therapeutic interventions.

C. Integrating Emotion Recognition with Music Therapy

While emotion recognition techniques have been extensively studied, there has been limited research on their application in music therapy, particularly when integrated with real-time analysis of both facial expressions and the music itself. The emotional response to music is highly complex and involves both external (facial expressions, body language) and internal (cognitive and emotional) processes. Combining these modalities offers a more holistic approach to understanding emotional responses in therapeutic contexts.

This paper presents a system that integrates facial expression analysis with music feature extraction to detect and analyze emotional responses to music. By analyzing both facial expressions and audio features from the music, the system aims to provide real-time, objective feedback to therapists about the patient's emotional state. The music features are extracted using advanced audio processing techniques, including Mel Frequency Cepstral Coefficients (MFCC), chroma features, tempo, and spectral contrast. These features have been shown to correlate with different emotional states triggered by music.

For facial expression analysis, the system uses a pre-trained CNN model to detect emotions from facial landmarks such as eye movement, mouth shape, and eyebrow positioning. These facial features are processed in real-time, and the emotional state is classified into categories such as happiness, sadness, anger, and relaxation. By fusing the results from both the facial expression recognition and music emotion recognition modules, the system provides a more accurate and comprehensive assessment of the patient's emotional state during music therapy.

II. LITERATURE SURVEY

Emotion recognition, particularly in the context of music and facial expression analysis, has seen significant advancements in recent years. However, the integration of both modalities—facial expression analysis and music feature extraction—for real-time emotion detection is still in its early stages. This section reviews the relevant work in facial emotion recognition, music emotion recognition, and the integration of multi-modal systems, highlighting the progress and gaps in each area.

A. Facial Emotion Recognition

Facial expressions serve as an essential medium for understanding emotions, and numerous studies have explored how to automatically detect emotions from facial cues. Early approaches to facial emotion recognition relied on handcrafted features such as the distances between facial landmarks (e.g., eyes, nose, and mouth), combined with traditional classifiers like Support Vector Machines (SVMs) or k-Nearest Neighbors (k-NN). For example, Bartlett et al. [1] developed a system that used Gabor filters to extract facial features and applied SVMs to classify emotions. While these systems showed promise, they were limited by their reliance on predefined features and their sensitivity to variations in lighting and facial orientation.

The emergence of deep learning, especially Convolutional Neural Networks (CNNs), has significantly transformed the domain of facial emotion recognition. CNNs automatically learn hierarchical representations from raw pixel data, eliminating the need for handcrafted features. In a landmark study, Goodfellow et al. [2] introduced the FER2013 dataset, a large dataset of facial expressions annotated with seven emotion categories, which became a standard benchmark for facial emotion recognition. Using CNNs, the researchers were able to achieve significant improvements in accuracy over traditional methods.

Subsequent work has focused on improving the robustness of CNN models for facial emotion recognition. Zeng et al. [3] proposed an attention-based CNN model that focused on critical facial regions (e.g., eyes, mouth) to improve emotion classification in the presence of occlusions or non-ideal lighting conditions. Similarly, Mollahosseini et al. [4] developed a multi-view CNN that aggregated information from different facial angles, further enhancing the system's accuracy in real-world scenarios. These advancements have made CNN-based facial emotion recognition systems highly effective, achieving accuracy rates above 70-80% on challenging datasets like FER2013.

B. Music Emotion Recognition

Music emotion recognition (MER) is another well-established field that seeks to classify emotions based on audio features extracted from music. The emotional impact of music is influenced by various factors, including tempo, pitch, timbre, rhythm, and harmony. Early work in MER focused on extracting low-level audio features such as Mel Frequency Cepstral Coefficients (MFCC), chroma, and spectral contrast, which were then classified using traditional machine learning models like SVMs and Random Forests. For instance, Tzanetakis and Cook [5] proposed an audio-based music genre classification system

that was later adapted for emotion recognition. Their system utilized MFCC features and achieved reasonable accuracy in distinguishing between basic emotions.

More recent studies have explored the use of deep learning models for MER. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, have been employed to model the temporal and spectral aspects of music that are crucial for emotion classification. For example, Kim et al. [6] used a CNN-LSTM hybrid model to capture both spatial and temporal features from music, achieving state-of-the-art results in recognizing emotions such as happiness, calmness, and sadness. The integration of LSTMs allowed the model to capture long-term dependencies in the music, which are essential for recognizing emotional progression over time.

Other researchers have focused on feature fusion techniques to improve MER accuracy. Yang and Chen [7] proposed a multi-modal system that combined acoustic features with lyrics-based features to enhance emotion recognition. Their system achieved higher accuracy in identifying emotions that are more influenced by lyrical content, such as sadness and anger, compared to systems that relied solely on acoustic features. However, the inclusion of lyrics complicates the system, as it requires accurate text extraction and language understanding, limiting its applicability to instrumental music or multilingual contexts.

While deep learning has greatly improved the performance of MER systems, the field still faces challenges. Music is inherently subjective, and different listeners may experience different emotions in response to the same piece of music. This subjectivity complicates the training process, as emotion labels are often inconsistent across listeners. Additionally, the emotional complexity of music, which can evoke multiple emotions simultaneously, is difficult to capture with current models that focus on classifying single emotions.

D. Gaps and Future Directions

While facial emotion recognition and music emotion recognition are well-researched areas, there is a noticeable gap in combining these modalities for real-time therapeutic applications. Most existing systems focus on entertainment or music recommendation but lack a clinical focus on enhancing therapeutic outcomes. The proposed system in this paper addresses this gap by integrating facial emotion recognition and music feature extraction to provide real-time emotional feedback during music therapy sessions.

Future research could explore the integration of additional modalities such as physiological signals (e.g., heart rate, skin conductance) to further improve the accuracy of emotion detection in therapeutic contexts.

III. MATERIALS AND METHODS

The proposed system for analyzing emotion responses to music integrates real-time facial expression analysis with music emotion recognition using advanced machine learning techniques. The system is designed to capture facial data and audio features simultaneously, classify emotional states based on these data, and provide real-time feedback on the user's emotional state. This section describes the architecture of the system, the data collection process, feature extraction techniques, machine learning models employed, and the methodology used to integrate the outputs from the facial and music emotion recognition modules.

A. System Architecture

The system architecture consists of two main modules: one for facial emotion recognition and another for music emotion recognition. Each module operates independently to extract features from facial expressions and music, respectively. The outputs from these modules are then fused to classify the overall emotional state of the user in real-time. The following steps summarize the system flow:

1. Input Data Acquisition: The system captures real-time facial data using a webcam and processes audio signals from the music being played. Both data streams are processed simultaneously.
2. Feature Extraction: The facial emotion recognition module extracts facial landmarks and other key features to classify emotions, while the music emotion recognition module extracts audio features from the music.
3. Emotion Classification: Both modules independently classify emotions into categories such as happiness, sadness, anger, and relaxation.
4. Fusion of Emotion Classifications: A late fusion approach is employed, where the probabilities from both modules are averaged to determine the final emotional state of the user.
5. Real-Time Feedback: The system provides real-time feedback to therapists or users, indicating the current emotional state based on both facial expressions and music-induced emotions.

Figure 1 below illustrates the high-level architecture of the system, showing how the facial and music emotion recognition modules interact and how the outputs are combined to classify emotions in real-time.

B. Facial Emotion Recognition Module

The facial emotion recognition module is responsible for detecting facial expressions and classifying them into predefined emotional categories. This module utilizes a Convolutional Neural Network (CNN) to automatically learn and classify emotions from facial images. The following subsections describe the steps involved in facial emotion recognition.

1. Data Collection and Preprocessing

For facial emotion recognition, the system uses a webcam to capture real-time images of the user's face. These images are then processed in real-time to extract facial features. The system focuses on detecting facial regions such as the eyes, eyebrows, mouth, and nose, which are highly informative for emotion recognition.

The CNN model was trained and validated using the FER2013 dataset. FER2013 is a widely used face recognition model that contains more than 35,000 facial expressions classified into seven different emotions: happy, sad, angry, scared, surprised, dirty and average. The dataset was preprocessed by normalizing the pixel values, resizing the images to 48x48 pixels, and converting them to grayscale to reduce computational complexity.

The facial emotion recognition preprocessing steps include:

- Face Detection: Faces are detected in real-time using the Haar Cascade Classifier, a well-established method for detecting objects (in this case, faces) in images. Once detected, regions of interest (ROIs) corresponding to faces are seeded and passed to CNN for inference.
- Image Resizing: To standardize input dimensions, facial images are resized to 48x48 pixels.
- Grayscale Conversion: Images are converted to grayscale to simplify the data and focus on essential facial features without the complexity of color.
- Normalization: Pixel values are normalized to have zero mean and unit variance, improving the performance of the CNN.

2. CNN Model for Facial Emotion Recognition

The CNN architecture used for facial emotion recognition is designed to automatically extract features from facial images. The architecture is composed of several convolutional layers, pooling layers, and fully connected layers. The CNN architecture is structured as follows:

- Convolutional Layers: These layers apply filters (kernels) to the input images to learn feature representations. Each convolutional layer applies a set of filters to detect edges, textures, and higher-order features such as facial expressions.
- Activation Function (ReLU): The Rectified Linear Unit (ReLU) activation function is applied after each convolutional layer to introduce non-linearity into the model. The ReLU function makes the model robust to noise by replacing all negative pixel values with zero.
- Max Pooling Layers: Max pooling is applied after each convolutional layer to reduce the spatial dimensions of the feature maps while retaining the most important features. This minimizes computational complexity and prevents overfitting.
- Fully Connected Layers: After a series of convolutional and pooling layers, the feature maps are flattened into a one-dimensional vector and passed through fully connected layers. These layers ultimately sort thoughts into predefined categories.
- Softmax Layer: The final output layer of the CNN uses the softmax activation function to produce a probability distribution over the seven emotion classes. The class with the highest probability is selected as the predicted emotion.

3. Model Training and Optimization

The CNN model was trained on the FER2013 dataset, employing the Adam optimizer with a 0.001 learning rate and a batch size of 32. The model was trained for 50 epochs, and cross-entropy loss was used as the loss function. Early stopping was implemented to prevent overfitting, and data augmentation techniques such as random rotation and horizontal flipping were used to increase the variability of the training data.

The trained model achieved an accuracy of 84% on the validation set, which is comparable to state-of-the-art models for facial emotion recognition.

C. Music Emotion Recognition Module

The music emotion recognition module classifies emotions based on the features extracted from audio tracks. Music is processed using the Librosa library, which provides tools for audio signal processing, including feature extraction. The steps involved in music emotion recognition are detailed below.

1. Audio Data Collection and Preprocessing

The music emotion recognition module processes audio signals from music tracks. For the purpose of this study, a curated playlist was used, consisting of songs that were pre-labeled with the expected emotional responses (happiness, sadness, anger, and relaxation). These tracks were selected based on their known emotional impact, as determined by previous studies in music therapy and emotion recognition.

The following preprocessing steps were applied to the audio data:

- Downsampling: Audio tracks were downsampled to 22,050 Hz to reduce computational complexity while maintaining the quality of the audio.
- Windowing and Framing: The audio signal was divided into overlapping frames, typically 2048 samples per frame, with a 50% overlap between consecutive frames. This framing allows for the analysis of time-varying aspects of the audio signal.
- Normalization: Audio signals were normalized to have zero mean and unit variance to improve the stability of the feature extraction process.

2. Feature Extraction for Music Emotion Recognition

Music emotions are influenced by various acoustic properties, such as melody, rhythm, tempo, and harmony. The following features were extracted from the audio signals to capture these properties:

- Mel Frequency Cepstral Coefficients (MFCC): MFCCs are a representation of the short-term power spectrum of the sound and are commonly used in speech and music recognition. In this study, 13 MFCCs were extracted from each frame of the audio signal. These coefficients capture important timbral information and are particularly effective at distinguishing between different emotional states.
- Chroma Features: Chroma features represent the 12 different pitch classes in music (e.g., C, C#, D, etc.) and are useful for identifying harmonic structures. These features are important for distinguishing emotions like happiness and relaxation, which are often associated with specific harmonic patterns.
- Tempo: The speed of a song, often measured in beats per minute (BPM), plays an important role in shaping how we emotionally perceive the music. Faster tempos are generally associated with emotions like happiness and anger, while slower tempos are linked to sadness and relaxation.
- Spectral Contrast: Spectral contrast measures the difference in amplitude between peaks and valleys in the sound spectrum. This feature helps differentiate between different textures in the music, which can evoke different emotional responses.
- Zero-Crossing Rate (ZCR): ZCR represents the rate at which the audio signal changes sign from positive to negative or vice versa. This feature is useful for identifying emotions such as anger, which are often characterized by more abrupt changes in the audio signal.

3. Random Forest Classifier for Music Emotion Recognition

After extracting the aforementioned features, the system employs a Random Forest Classifier to classify the music into one of four emotional categories: happiness, sadness, anger, and relaxation. Random Forest was chosen due to its robustness and ability to handle a large number of input features.

a. Training Data

The classifier was trained on a subset of the Million Song Dataset (MSD), a large-scale dataset of music tracks labeled with emotional tags. The training set consisted of approximately 10,000 songs, each labeled with one of the four emotion categories. The selection of tracks was based on a diverse range of genres to ensure the model's generalizability across different musical styles.

b. Model Training

A Random Forest Classifier was trained with 100 decision trees, where each tree was constructed using a randomly selected subset of features. Gini impurity was used as the criterion for splitting nodes, and bootstrap aggregation (bagging) was applied to reduce overfitting. The model was trained on 70% of the dataset, while the remaining 30% was used for validation.

- Feature Selection: Before training, feature importance scores were computed using the trained Random Forest model to determine which features had the most significant impact on classification accuracy. This step allowed for the refinement of the feature set, ensuring that only the most relevant features were used for model training.
- Hyperparameter Tuning: A grid search was performed to optimize hyperparameters such as the number of trees, maximum depth of each tree, and the minimum number of samples required to split a node. The optimal hyperparameters were selected based on the validation set performance, ensuring that the model was neither underfitting nor overfitting.

c. Model Evaluation

The Random Forest Classifier's performance was assessed using various evaluation metrics.

- Accuracy: The ratio of correctly classified instances to the total number of instances.
- Precision: The proportion of true positive predictions out of all predicted positives, showing the accuracy of the predicted emotional states.
- Recall: The ratio of true positive predictions to the total actual positives, indicating the model's ability to identify all relevant instances.
- F1 Score: The harmonic mean of precision and recall offers a unified metric that balances the two measures.

The model achieved an accuracy of 88% on the validation set, demonstrating its effectiveness in classifying musical emotions.

D. Integration of Facial and Music Emotion Recognition

The outputs of both the facial and music emotion recognition modules are integrated using a late fusion approach. This involves combining the probabilities from both classifiers to derive the final emotional state of the user. The fusion process is detailed below.

1. Late Fusion Approach

Late fusion is a strategy where separate models are trained independently, and their outputs are combined at a later stage. In this system, the probabilities outputted by the CNN for facial emotion recognition and the Random Forest Classifier for music emotion recognition are averaged to compute a final emotional probability distribution.

• Probability Averaging: For each emotional category, the probabilities obtained from both classifiers are averaged. For example, if the facial recognition model predicts a probability of 0.7 for happiness and the music recognition model predicts 0.6 for happiness, the final probability for happiness would be $(0.7 + 0.6) / 2 = 0.65$.
• Final Emotion Classification: The emotional category with the highest averaged probability is selected as the overall emotional state of the user. This final classification

is presented to the therapist or user in real-time, allowing for immediate insights into the user's emotional experience.

2. Real-Time Feedback Mechanism

To provide instant feedback, the system constantly monitors the user's emotional state.. The integration of both modules allows for rapid adjustments based on the changes in the user's facial expressions and the emotional content of the music being played.

• User Interface: A user-friendly interface displays the real-time emotional state and visual representations of the confidence levels for each emotional category. This interface is designed to be intuitive, allowing therapists to quickly understand the user's emotional responses.

• Feedback for Therapy: The feedback mechanism is designed to assist therapists in making informed decisions about therapy sessions. By understanding the user's emotional responses to music, therapists can tailor their approaches and choose music that aligns with the user's emotional needs.

IV. RESULTS AND DISCUSSION

This section presents the results obtained from testing the proposed system, which integrates facial emotion recognition and music emotion recognition. The system was evaluated on its ability to classify emotions such as happiness, sadness, anger, and relaxation, based on real-time facial expression data and music feature extraction. Evaluate performance using various metrics such as precision, accuracy, recall, and F1score. In addition to quantitative metrics, the system's applicability in therapeutic settings was assessed through qualitative analysis.

A. Experimental setup

To evaluate the system's performance, a controlled experiment was conducted with 30 participants. The participants were asked to listen to a curated playlist of music designed to evoke specific emotions. The playlist included tracks that were pre-labeled with expected emotional outcomes (happiness, sadness, anger, and relaxation) based on prior research in music emotion recognition. During the experiment, the system continuously recorded participants' facial expressions using a webcam and analyzed the music features in real-time using the Librosa library. Each participant was asked to sit in front of a computer in a well-lit room, and facial data were collected while they listened to the music.

The dataset used for training the facial emotion recognition module was derived from the FER2013 dataset, which contains over 35,000 labeled facial expression images. For the music emotion recognition module, the system was trained on a subset of the Million Song Dataset (MSD), focusing on songs with well-defined emotional labels. The Random Forest Classifier was used for the music-based emotion classification, while the facial emotion recognition was performed using a pre-trained Convolutional Neural Network (CNN). The post-fusion method was used to combine the facial results and emotional music to create the final image of the participant.

B. Quantitative Results

The system's performance was evaluated using the following metrics: accuracy, precision, recall, and F1-score. These metrics were computed separately for the facial emotion recognition module, the music emotion recognition module, and the combined system.

1. Facial Emotion Recognition Results

The facial emotion recognition module was trained on the FER2013 dataset and fine-tuned for real-time processing. The performance of the facial recognition system is presented in Table I below.

The system showed high accuracy for recognizing happiness and sadness, which are commonly recognized emotions in facial expression datasets. The classification of anger and relaxation was somewhat lower, likely due to the subtleties involved in distinguishing these emotional states in facial expressions. Relaxation, in particular, is challenging to detect from facial features alone, as it often lacks distinct facial movements compared to emotions like anger or happiness.

2. Music Emotion Recognition Results

The music emotion recognition module was tested on a subset of the Million Song Dataset, focusing on extracting features such as Mel Frequency Cepstral Coefficients (MFCC), chroma, tempo, and spectral contrast. Table II presents the performance of the music emotion classifier.

The music emotion recognition module performed consistently well across all emotional categories, with the highest performance for the recognition of relaxation and happiness. These emotions are more strongly correlated with specific musical features such as slow tempo (for relaxation) or higher energy and faster beats (for happiness). Anger, while relatively well-classified, presented slightly lower recall due to the variability in musical styles that can evoke this emotion.

3. Combined System Results

The combined system demonstrated strong performance, with an overall accuracy of 88.4%. The integration of facial and music-based features improved the detection of emotions like anger and relaxation, which were less accurately classified when using only facial or music data individually. This suggests that the system benefits from the complementary nature of facial and musical cues, particularly in capturing complex emotional states.

The combined system integrates both the facial emotion recognition and music emotion recognition modules using a late fusion approach. The final emotional state is determined by averaging the probability distributions of the two modules. Table III shows the performance of the combined system.

C. Qualitative Analysis and Discussion

The qualitative results from the experiment showed that the system was effective in real-time settings, providing therapists with immediate feedback on the emotional states of participants. During the testing, the system was able to detect transitions between emotional states as participants listened to different tracks, and it provided a detailed log of emotional responses that could be analyzed after each session.

1. Challenges in Emotion Recognition

Despite the promising results, several challenges were observed during the experiment. First, emotions like anger and relaxation were more difficult to classify using facial expressions alone, as facial expressions for these emotions tend to be less distinct or more variable among individuals. Relaxation, for instance, may not always manifest in overt facial expressions, making it difficult for the facial recognition module to consistently detect. The integration of music-based features helped address this issue, as certain musical patterns strongly correlate with relaxation (e.g., slow tempo and smooth rhythms).

Another challenge was the subjectivity of emotional responses to music. While the system was trained on a labeled dataset with general emotional labels for different tracks, individual participants sometimes reported emotional experiences that differed from the expected labels. For instance, some participants found certain "happy" tracks to be annoying or stressful, leading to emotional states that were not aligned with the system's classification. This underscores the subjective nature of music emotion recognition and highlights the need for more personalized emotion recognition models that can adapt to individual preferences and emotional triggers.

2. System Applicability in Therapeutic Settings

One of the key goals of this system is its application in therapeutic settings, particularly for music therapy. The system's ability to provide real-time emotional feedback offers significant potential for enhancing therapeutic outcomes. Therapists can use the system to monitor patients' emotional responses during music therapy sessions, allowing for a more personalized approach to treatment. For instance, if a patient is not responding well to a particular track, the therapist can quickly adjust the music selection based on the emotional feedback provided by the system.

The system's ability to detect multiple emotions simultaneously is particularly useful in therapeutic contexts, where patients may experience mixed emotions. For example, a patient dealing with anxiety may feel both anger and sadness in response to certain music. The system's late fusion approach allows it to capture these overlapping emotional states, providing therapists with a more nuanced understanding of the patient's emotional experience.

D. Limitations and Future Directions

While the system showed strong performance in detecting and classifying emotions, there are several areas for improvement. First, the system currently relies on pre-defined emotional labels for the music tracks, which may not align with the emotional experiences of all users. Future iterations of the system could incorporate user-specific emotion labels, allowing the system to learn and adapt to individual emotional responses to music over time.

Second, while the system integrates facial and music emotion recognition, it does not account for other important physiological signals that can provide additional insights into emotional states. Future work could explore the integration of heart rate variability (HRV), skin conductance, or electroencephalogram (EEG) signals to further improve the system's accuracy and robustness.

Finally, improving the system's ability to detect subtle and mixed emotions will be critical for its success in therapeutic settings. While the current model captures basic emotions effectively, more sophisticated models that can detect complex emotional states, such as ambivalence or emotional neutrality, will enhance its clinical utility.

V. CONCLUSION

This study presents a novel system for analyzing emotional responses to music through the integration of facial expression recognition and music emotion classification. By utilizing advanced machine learning techniques, the system effectively captures and interprets emotional states in real time. The architecture consists of two independent modules: a Convolutional Neural Network (CNN) for facial emotion recognition, which achieved an accuracy of 84% using the FER2013 dataset, and a Random Forest Classifier for music emotion recognition, which attained an accuracy of 88% with the Million Song Dataset.

The integration of these modules via a late fusion approach enhances the overall classification accuracy, providing a comprehensive view of the user's emotional state by considering both facial expressions and the emotional context of music. This robust classification allows the system to recognize various emotions such as happiness, sadness, anger, and relaxation, offering valuable insights into how music influences emotional experiences.

Real-time feedback mechanisms have been incorporated to assist therapists in tailoring their interventions based on the user's emotional responses. By delivering immediate insights, the system enhances therapeutic practices in music therapy, where understanding emotional nuances is crucial for effective treatment.

Future research will aim to expand the dataset to include diverse musical genres and cultural contexts, further improving the model's generalizability. Additionally, enhancing the user interface will facilitate better interaction for therapists and users alike. Ultimately, this system not only contributes to academic discourse on music and emotion but also holds the potential to improve therapeutic outcomes for individuals facing emotional and psychological challenges, thus bridging the gap between technology and mental health support.

REFERENCES

[1] Chen W, Wang W, Wang K, et al. Lane departure warning systems and lane line detection methods based on image processing and semantic segmentation–a review. Journal of Traffic and Transportation Engineering. 2020; 8(4): 319-330.
[2] Zhang F, Xie Y, Chen X, Li Z. Facial expression recognition using deep learning: A survey. Neurocomputing. 2021; 450: 270-287.
[3] Grudic GBA, Szmit DR, Pasquale PJ. Facial expression recognition with deep learning: A review. International Journal of Computer Science and Information Security. 2017; 15(6): 59-65.
[5] Zhang L, Huang Z, Liu J. Facial emotion recognition based on deep learning and convolutional neural networks. Journal of Ambient Intelligence and Humanized Computing. 2020; 11(6): 2401-2410.
[6] Lee MAA, Tsai SM. A review of music emotion recognition: A survey. Applied Sciences. 2022; 12(2): 633.
Hu GD, Liu N, Zhang GQ, et al. Music emotion recognition based on audio features using deep learning: A survey. IEEE Access. 2019; 8: 170703-170719.
[7] Yang DK, Huang CK, Liu XL, et al. Emotion recognition from music: A review. Multimedia Tools and Applications. 2021; 80(3): 4531-4548.
[8] Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge: MIT Press; 2016.
[9] Bishop CM. Pattern Recognition and Machine Learning. New York: Springer; 2006.
[10] Chen KF, Tsai RC, et al. A survey of machine learning techniques for emotion recognition in speech and text: Challenges and opportunities. IEEE Transactions on Affective Computing. 2018; 11(1): 1-20.
[11] Karras T, Aila T, Laine S, et al. A style-based generator architecture for generative adversarial networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2019; 4401-4410.
[12] Kosslyn SM, Ganis G, Thompson WL. Neural foundations of imagery. Nature Reviews Neuroscience. 2001; 2(9): 635-642.
[13] Wang W, Zhang Y, Zhu C, et al. The effects of emotion on attention: A review. Frontiers in Psychology. 2020; 11: 169.
[14] Zhang H, Zhang X, Zhai W, et al. Emotion recognition from audio using deep learning: A survey. IEEE Transactions on Audio, Speech, and Language Processing. 2020; 28: 2490-2505.
[15] Abdurrahman O, Niazi A, Amad A. Emotion detection from audio signals: A review of techniques and future directions. Journal of Ambient Intelligence and Humanized Computing. 2021; 12(9): 9669-9684.
[16] Zhang Y, Zhan Y, Zhai Y. An overview of datasets for music emotion recognition. Computational Intelligence and Neuroscience. 2022; 2022: 1-17.
17] Yang X, Liang X, Chen Y, et al. Emotion recognition from speech: A review of the current state-of-the-art and future directions. Computer Speech & Language. 2021; 68: 101194.
[18] Fong A, Matusz PJ, Tzeng Y, et al. Real-time emotion recognition in music using deep learning. Journal of Music Technology and Education. 2020; 13(1): 37-57.
[19] Yang Q, Wang K, Yan J, et al. Real-time emotion recognition from facial expressions: A review. International Journal of Pattern Recognition and Artificial Intelligence. 2021; 35(3): 2130004.
[20] Stowell D, Giannakopoulou A, Plumbley MD. An open dataset for music emotion recognition. Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR). 2010; 387-392.
[21] Wu H, Liu Y, Zhao Y, et al. Emotion recognition in music based on the use of the Million Song Dataset. Journal of Intelligent Information Systems. 2020; 54(3): 563-577.
[22] Tzeng Y, Matusz PJ, Fong A, et al. Emotion recognition from music using neural networks. ACM Transactions on Intelligent Systems and Technology. 2021; 12(3): 1-22.
[23] Tschacher W, Meier E. Emotions in therapy: Implications for psychotherapy research. Psychotherapy Research. 2020; 30(1): 1-11.
[24] Kim K, Ko B, Yu J. A survey on emotion recognition in the context of music. Artificial Intelligence Review. 2021; 54(1): 1-32.
[25] Zeng D, Zhang Y, Wang Y, et al. Audio-visual emotion recognition: A review of deep learning methods. IEEE Transactions on Affective Computing. 2022; 13(2): 1-1.
[26] Chen Y, Zhou M, Li Y, et al. Deep learning for emotion recognition: A comprehensive review. Artificial Intelligence Review. 2021; 54(2): 1127-1166.
[27] Karam M, Dufour F, et al. Music emotion recognition using deep neural networks: A comprehensive survey. IEEE Access. 2021; 9: 23658-23676.
[28] Arora A, Ghosh S, Singh P, et al. Emotion recognition in music using hybrid feature extraction techniques. Applied Acoustics. 2020; 166: 107338.
[29] Ghaleb A, Al-Naji A, Al-Khaldi M, et al. A survey on music emotion recognition techniques. Expert Systems with Applications. 2021; 164: 113703.
[30] Pons J, Dufour J, et al. A comparison of music emotion recognition approaches: A survey. Journal of New Music Research. 2020; 49(4): 334-353.
[31] Jatobá M, d'Avila C, da Silva R, et al. Music emotion recognition in the wild: A survey. IEEE Transactions on Multimedia. 2020; 22(4): 920-934.