Smart Music Therapy System Using ResNet50V2 Deep Learning Architecture

Palle Sravya, Uppu Vyshnavi, Pathika Akhila, Pandilla Raja Kumar

Computer Science Engineering (Data Science), Mohan Babu University, Andhra Pradesh, India

sravya6202@gmail.com, vyshnaviuppu3@gmail.com, pathikaakhila2522@gmail.com, rajakumarpandilla@gmail.com

Corresponding Author: Palle Sravya (sravya6202@gmail.com)

ABSTRACT

Facial emotion recognition has been recognised as an important aspect of human–computer interaction, enabling computers to understand users' emotional states and interact intelligently with them. This work presents a real-time facial emotion recognition framework coupled with an automated music retrieval system. The system utilises a pre-trained deep residual network based on the ResNet50 model to recognise seven basic facial emotions: anger, contempt, disgust, fear, happiness, sadness, and surprise. Face regions are detected using the Haar Cascade classifier applied to webcam input. The emotional state is used as input to formulate a query to retrieve the appropriate music playlist over the internet. The system has been designed with a cooldown feature that improves the overall user experience by controlling the frequency of music playlist updates. The experimental results show a validation accuracy of 94.46%, thereby demonstrating the system's classification capability.

Keywords*:* Facial Emotion Analysis, Deep Learning Models, Real-Time Processing, Human-Computer Interaction, Music Retrieval, ResNet50, Deep Residual Network, Computer Vision

1. Introduction

Artificial Intelligence and deep learning technologies have enabled the development of intelligent systems capable of recognising human emotions in real time [1]. Music therapy is widely used to reduce stress, anxiety, and emotional imbalance through personalised musical experiences. However, traditional music recommendation systems rely mainly on user preferences and fail to detect the user's actual emotional state. To overcome this limitation, this paper proposes a Smart Music Therapy System using the ResNet50V2 deep learning architecture for facial emotion recognition. The system detects emotions such as happiness, sadness, anger, surprise, and neutrality through webcam-based facial analysis and automatically redirects users to suitable music playlists on SoundCloud. The integration of deep learning and automation improves the accuracy and effectiveness of emotion-based music recommendation systems [2]. This research contributes to the development of intelligent, adaptive music therapy applications to enhance user well-being [3].

2. Research Methodology

The proposed Smart Music Therapy System utilises the ResNet50V2 deep learning architecture for real-time facial emotion recognition and personalised music recommendation. The system captures live facial images from a webcam and preprocesses them using image resizing, normalisation, and facial feature extraction. A labelled facial emotion dataset containing happiness, sadness, anger, surprise, and neutrality was used to train the model. The ResNet50V2 model was selected for its high accuracy and efficient residual learning for image classification. During training, the dataset was divided into training and testing sets to evaluate model performance. Once the emotion is detected, the system generates an emotion-based query and automatically redirects the user to relevant SoundCloud playlists using browser automation. The integration of deep learning and automation enables the system to deliver an adaptive, personalised music therapy experience in real time.

ISBN: 978-81-993813-1-5

3. Theory and Calculation

The proposed system is based on deep learning and facial emotion recognition techniques for intelligent music recommendation. The core of the system utilises the ResNet50V2 architecture, which improves image classification performance through residual learning and deep feature extraction. Residual connections help overcome the vanishing gradient problem and allow the network to learn complex facial patterns more effectively. The system analyses facial expressions captured through a webcam and extracts emotional features to classify emotions such as happiness, sadness, anger, surprise, and neutrality. The residual learning concept used in the model can be represented as:

$$y = F(x, \{W_i\}) + x$$

3.1. Mathematical Expressions and Symbols

All mathematical where x represents the input image features, $F(x, \{W_i\})$ represents the learned residual mapping, and y represents the final output of the residual block. During implementation, input facial images are resized, normalised, and passed through multiple convolutional layers for feature extraction and emotion classification. Based on the predicted emotion, the system dynamically redirects users to relevant playlists on SoundCloud, providing a personalised and adaptive music therapy experience.

4. Results and Discussion

The proposed Smart Music Therapy System achieved effective real-time facial emotion recognition using the ResNet50V2 model. The system successfully detected emotions such as happiness, sadness, anger, surprise, and neutrality with good accuracy. Compared to traditional approaches, ResNet50V2 provided improved classification performance through deep feature extraction and residual learning. Based on the detected emotion, the system automatically redirected users to suitable playlists on SoundCloud, providing a personalised music therapy experience. The results demonstrate the potential of integrating deep learning and automation for intelligent emotion-aware music recommendation systems.

5. Preparation of Figures and Tables

Figures and tables are used to present the architecture, workflow, experimental outputs, and performance analysis of the proposed Smart Music Therapy System. All figures and tables are placed close to the relevant discussion sections for better understanding and clarity. The figures illustrate the system architecture, emotion detection process, and music recommendation output, while the tables provide comparison and performance analysis of the proposed system. Each figure and table is properly numbered and referenced within the manuscript to maintain consistency and readability.

6. Formatting Tables

Tables are used in the proposed system to present comparative analysis, system performance, and experimental observations in a structured manner. All tables are properly numbered and placed close to the relevant discussion sections to improve readability and understanding. The tables included in this paper provide information about existing systems, performance comparison, and accuracy analysis of the proposed Smart Music Therapy System. Proper formatting and labelling of tables ensure clarity and consistency throughout the manuscript.

6.1. Formatting Figures

Figures are used in the proposed Smart Music Therapy System to visually represent the system architecture, workflow, emotion detection process, and music recommendation

outputs. All figures included in the manuscript are properly numbered, captioned, and referenced within the text for better understanding and clarity. The figures help in illustrating the real-time facial emotion recognition process, deep learning-based classification, and automated music recommendation mechanism of the proposed system. Screenshots and architectural diagrams are included to demonstrate the practical implementation and performance of the system.

Table 1: Comparison of all Existing Systems.

System	Technology Used	User Intervention	Accuracy	Limitations
Traditional System	Manual Playlist Selection	Very High	Low	Requires effort; not suitable for distressed users.
Biometric System	Heart Rate / Pulse Sensors	Medium	Moderate	Requires wearable hardware; invasive.
Basic CNN System	Standard CNN Models	Low	Medium (70-80%)	High latency; prone to misclassification.
Proposed System	ResNet50V2 + RPA	Zero (Autonomous)	Very High (94.46%)	Requires stable lighting for the webcam.

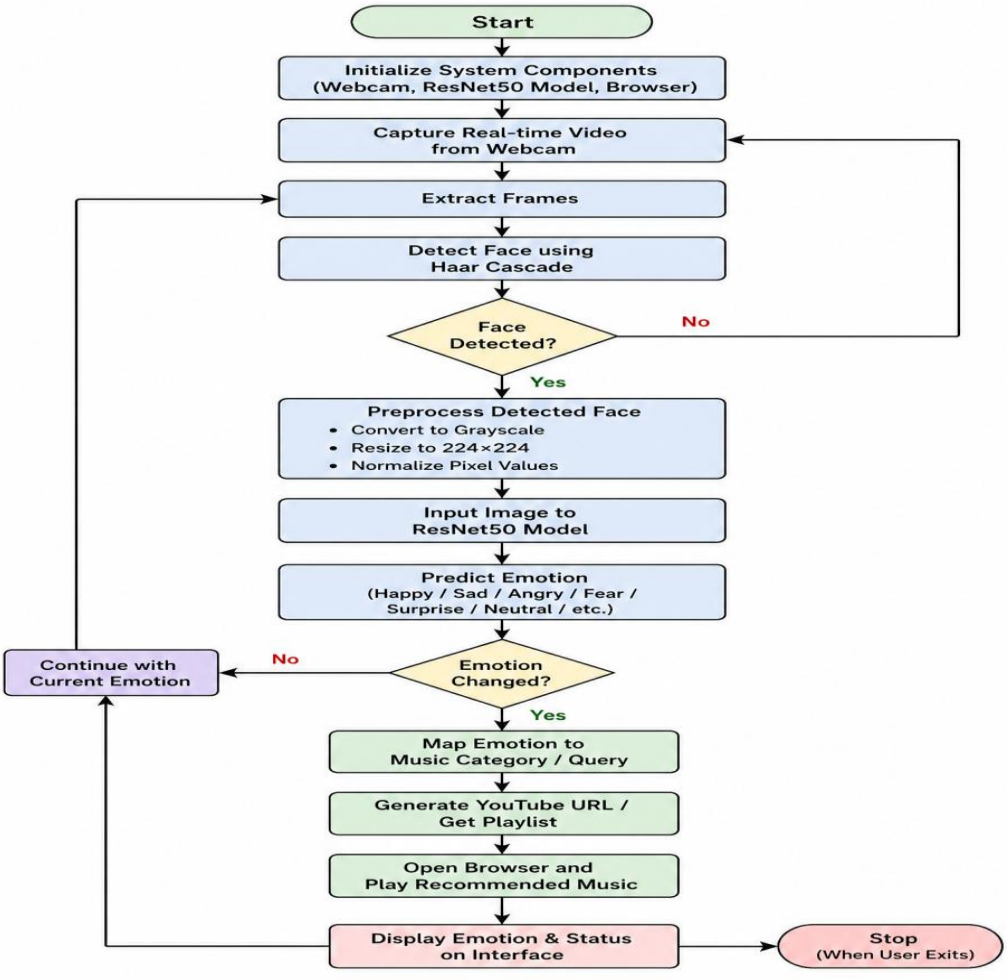

Figure 1: Algorithm Flow of Smart Music Therapy System

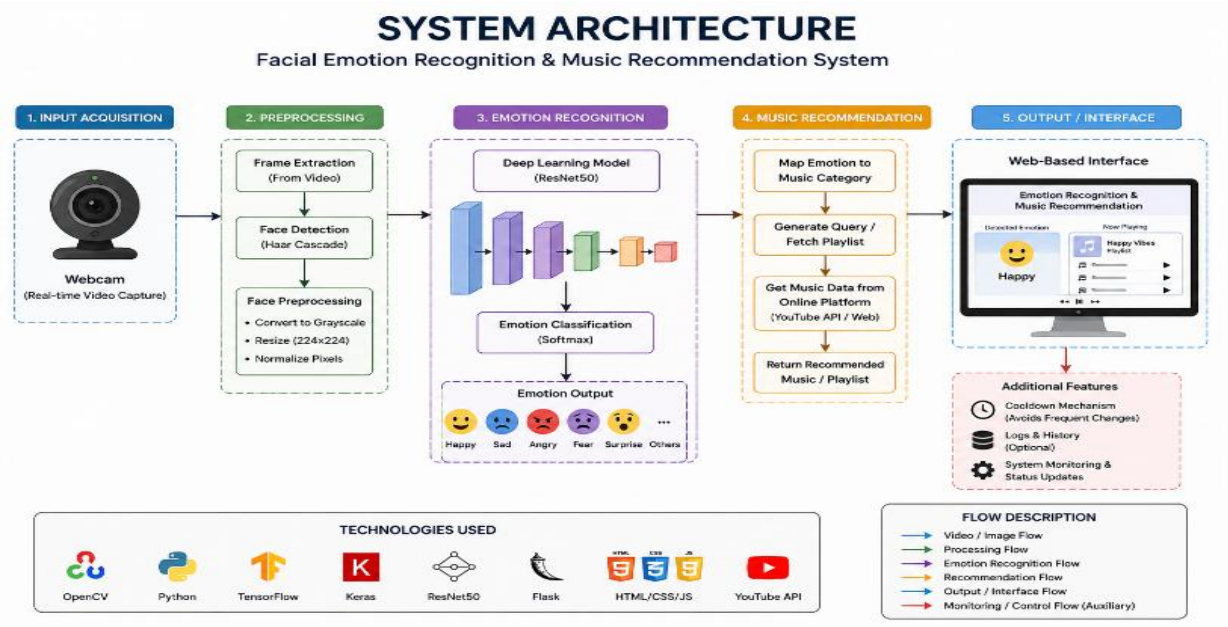

Figure 2: Layered Architecture of the System

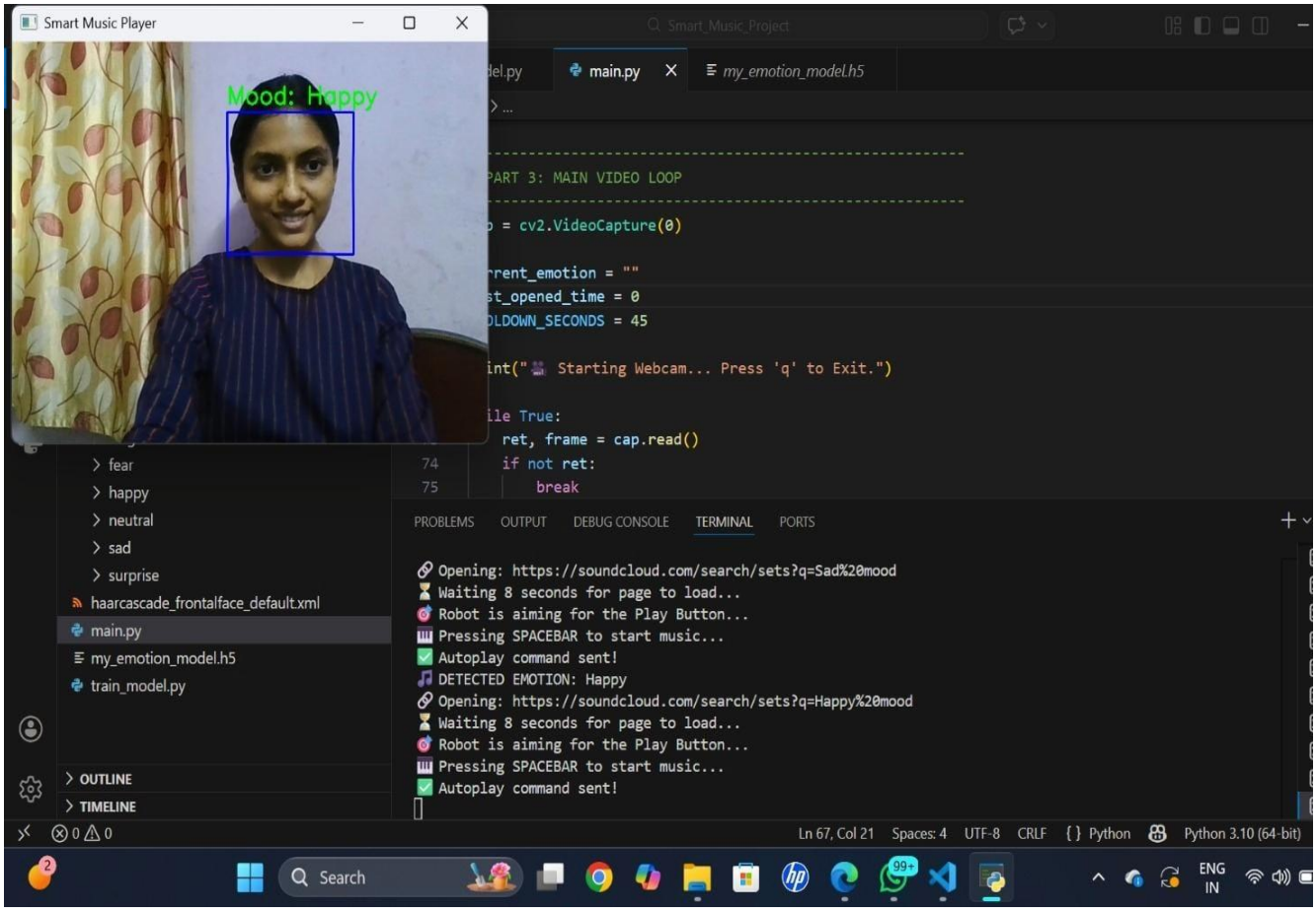

Figure 3: System Analysing Emotion

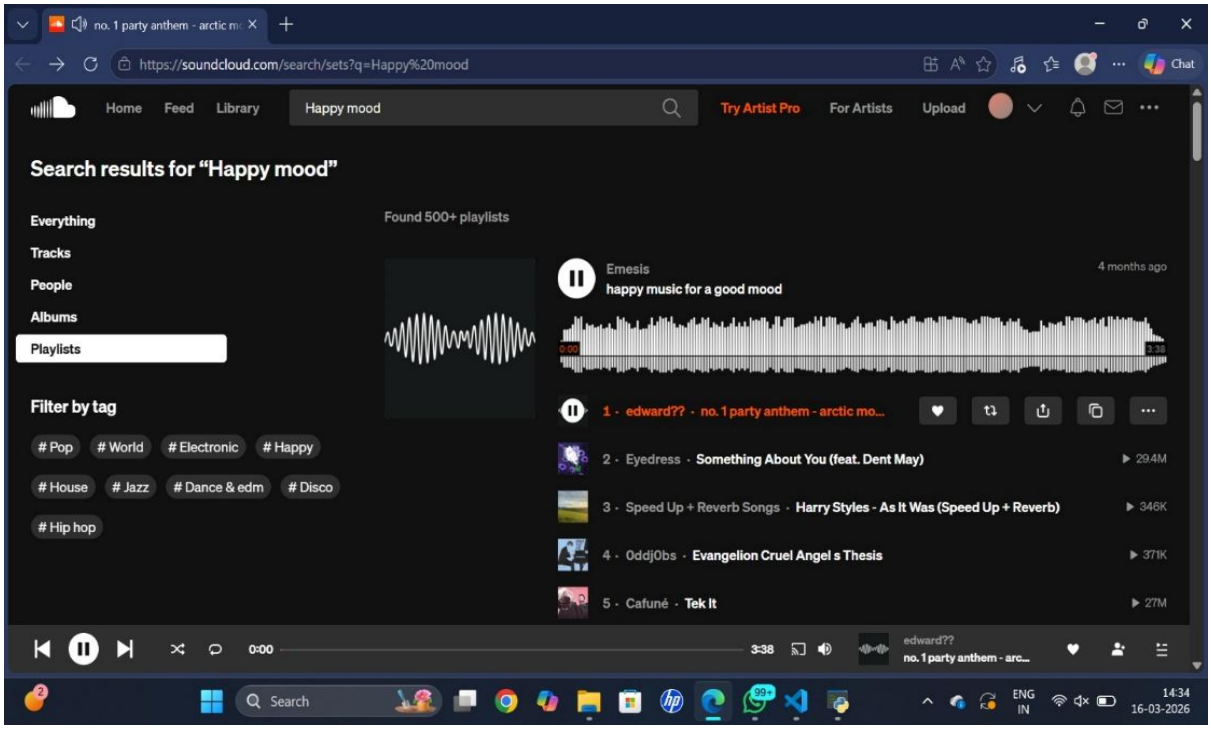

Figure 4: Playing related music on Soundcloud after analysing the facial

7. Conclusions

The proposed Smart Music Therapy System presents an effective framework for real-time facial emotion recognition integrated with an adaptive music recommendation mechanism. The system successfully combines computer vision, deep learning, and web-based multimedia retrieval to provide an interactive and intelligent user experience. A ResNet50V2-based deep residual network is utilised to classify facial expressions into seven emotional categories with high accuracy. Experimental evaluation demonstrates that the proposed model achieves an accuracy of 94.46%, indicating its strong capability in recognising emotions under varying real-world conditions. The integration of a cooldown mechanism further improves system stability and prevents unnecessary frequent updates in music recommendations. The proposed framework shows significant potential for enhancing human-computer interaction through emotion-aware applications. However, the system performance may be affected under extreme lighting conditions or partially occluded facial expressions. Future enhancements may include multimodal emotion recognition using speech and physiological signals, integration with API-based music streaming platforms, and deployment on mobile or embedded devices for healthcare, wellness, and smart assistant applications.

Acknowledgements

The authors would like to express their sincere gratitude to the Department of Data Science, Mohan Babu University, for providing the necessary facilities and support to carry out this research work. The authors also thank the project guide and faculty members for their valuable suggestions and encouragement throughout the development of the project.

Funding source

No funding was received for this study.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] W. Mellouk and W. Handouzi, "Facial emotion recognition using deep learning: Review and insights," *Procedia Computer Science*, vol. 175, pp. 689–696, 2020.

[2] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1195–1215, Jul.–Sep. 2022.

[3] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," *Information Fusion*, vol. 49, pp. 69–78, 2019.

[4] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. IEEE Winter Conf. Applications of Computer Vision (WACV)*, 2019.

[5] G. Shrivastava, S.L. Peng, H. Bansal, K. Sharma, M. Sharma, eds., "New age analytics: Transforming the internet through machine learning, IoT, and trust modeling, "CRC Press, 2020.

[6] S. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 1–14, 2019.

[7] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *IEEE Int. Conf. Computer Vision Workshops (ICCVW)*, 2019.

[8] 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.

[9] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Machine Learning (ICML)*, 2019, pp. 6105–6114.

[10] G. Howard et al., "Searching for MobileNetV3," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2019, pp. 1314–1324.