International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research Paper

MUSIC EMOTION CLASSIFICATION WITH NEURAL NETWORK ARCHITECTURE AND LIBROSA

ADALA AMARNATH [1], UPPALA HARSHITH [2], MENGANI SHIVATHMIKA [3], NALLA INDU [4], NUNEMUNTHALA SHASHANK [5]

[1]Assistant Professor, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India

[2, 3, 4, 5] B.Tech, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India

ABSTRACT— The classification of musical emotions is essential for organizing, searching, and recommending music on modern platforms. Traditional models often rely on raw audio or textual features, which may not fully capture the rich emotional content embedded in music. To address this, we propose a Convolutional Neural Network (CNN)-based model combined with Librosa for feature extraction to classify musical emotions effectively. In the proposed approach, Librosa is used to extract meaningful audio features from music signals, including Mel-frequency cepstral coefficients (MFCCs), chroma features, spectral contrast, and tonettes representations. These features provide a compact and informative representation of the audio, capturing timbral, harmonic, and rhythmic characteristics relevant to emotion recognition. The CNN model is then applied to learn hierarchical patterns from these extracted features. Convolutional layers automatically capture local correlations in the audio features, while pooling layers reduce dimensionality and highlight dominant emotional patterns. This deep learning framework eliminates the need for handcrafted feature combinations, allowing the model to generalize effectively across diverse music samples. By combining Librosa feature extraction with the pattern learning capability of CNNs, the proposed system is able to capture complex emotional relationships in music. This approach offers a robust and scalable solution for automated

music emotion classification, supporting applications such as music recommendation, playlist generation, and music analytics in real-world platforms.

Index Terms – Music Emotion Classification, Neural Networks, Librosa, Deep Learning, MFCC, Audio Feature Extraction, Emotion Recognition, Audio Classification, Music Information Retrieval, Machine Learning.

I. INTRODUCTION

Music is a universal language that conveys a wide range of emotions, influencing human mood, behavior, and cognitive processes. Recognizing these emotions in music has become increasingly important for applications such as personalized recommendation systems, playlist generation, and music analytics. Traditional approaches for musical emotion classification often rely on raw audio signals or textual metadata, which may fail to capture the intricate emotional nuances embedded within the music. Additionally, handcrafted features can be time-consuming to extract and may not generalize well across diverse music genres. With the advancement of machine learning and deep learning techniques, automated approaches have emerged to address these challenges effectively. In this study, we propose a hybrid framework that combines Librosa-based feature extraction with Convolutional Neural Networks (CNN) to classify musical emotions. Librosa, a powerful audio processing library, is used to extract meaningful audio representations such as Mel-frequency cepstral coefficients (MFCCs), chroma features, spectral contrast, and tonnetz features. These features capture the timbral, harmonic, and rhythmic characteristics of music, providing a compact yet informative input for emotion recognition. The CNN model is then applied to learn hierarchical patterns from these features, automatically detecting correlations that indicate emotional content. Convolutional layers in CNN efficiently capture local dependencies, while pooling layers reduce dimensionality and emphasize dominant patterns relevant to emotions. This deep learning approach eliminates the need for manual feature engineering, allowing the model to generalize across various musical styles and genres. By integrating Librosa's feature extraction with CNN's pattern learning capability, the proposed system effectively models complex relationships between audio characteristics and perceived emotions. The framework supports scalable and automated classification, making it

suitable for real-world music platforms. Furthermore, this approach enhances the accuracy and robustness of music emotion recognition systems, providing valuable insights for listeners, content creators, and streaming platforms. By leveraging deep learning, the system can adapt to large and diverse music datasets, improving recommendation quality and user satisfaction. Ultimately, this project contributes to intelligent music management and analytics, enabling emotionally aware applications in digital music platforms.

II. RELATED WORK

A. Music emotion recognition based on a neural network with an inception-gru residual structure

This study designs a hybrid deep architecture combining inception modules with gru residual connections for music emotion recognition. It extracts audio features using standard techniques (including mfccs and spectral features) before feeding into the network. The inception blocks capture multi-scale timbral patterns, while gru residual units model temporal dependencies. Feature extraction via librosa-like pipelines ensures consistent input across samples. The model improves over traditional cnn or rnn approaches by enhancing context awareness. Evaluated on a music-emotion dataset, it reports significant accuracy gains across valence and arousal dimensions. The residual design supports deeper models without vanishing gradients. Emphasis is on real-world scalability due to moderate model complexity. The architecture aligns well with lightweight cnn setups for edge deployment. It provides a strong baseline for cnn + temporal modeling in music emotion tasks.

B. Instrument emotion recognition from polyphonic instrumental music using mfcc and cens features with deep neural networks

This paper focuses on recognizing emotions from polyphonic instrumental music. It extracts mfccs and cens (chroma energy normalized statistics) features using audio toolkits similar to librosa. The deep neural network processes these compact representations to classify musical emotions. The mfccs capture timbral content while cens encodes harmonic information—both crucial for emotion cues. Results show improved emotion classification over single feature models. Effectiveness is particularly notable for instrumental tracks lacking lyrics or vocals. The system demonstrates robustness across varying instrumentation and genres. Authors highlight potential for real-time playlist generation and recommendation

systems. Efficiency and feature quality make it suitable for cnn-based pipelines. This aligns well with your cnn + librosa approach for emotion extraction.

C. *Hierarchical audio-visual information fusion with multi-label joint decoding for mer 2023*

This framework tackles music emotion recognition (mer) using both audio and visual signals in a hierarchical fusion manner. Deep features from pre-trained foundation models represent raw modalities. Attention-guided feature gathering modules merge audio and visual cues effectively. A joint decoding stage handles both discrete emotion classification and valence regression. A multi-task uncertainty loss allows the model to predict emotion categories and intensity simultaneously. The system ranked top-three in the mer-multi challenge—a strong benchmark. While visual features are used, the fusion strategy is adaptable to integrating just audio (librosa-derived) features for pure audio pipelines. The attention fusion idea could be ported to combine mfcc/chroma outputs with cnn embeddings. It demonstrates that structured feature fusion significantly improves generalization.

D. *Hybrid lstm–attention and cnn model for enhanced speech emotion recognition*

Although focused on speech emotion recognition, this model's architecture is relevant to musical emotion tasks. It combines cnn layers for spatial (spectrogram) features with lstm and attention for sequence modeling. Feature extraction uses spectrograms akin to cnn input from librosa-derived mfccs or mel-spectrograms. The attention mechanism helps the network focus on emotion-relevant temporal frames. Results indicate significant gains in classification accuracy and generalization over pure cnn or rnn models. Modular design supports lightweight cnn front ends feeding into recurrent-attention backends. It's suitable for streaming applications with real-time inference demands. Adapting this structure to musical emotion classification could enhance capturing temporal emotional dynamics. Provides a blueprint for combining cnn-driven feature maps with temporal context awareness.

III. PROPOSED METHODOLOGY

The proposed system is designed to automatically classify musical emotions using audio signals. It begins with Librosa-based feature extraction, which converts raw

audio into informative representations such as MFCCs, chroma features, spectral contrast, and tonnetz. These features capture key characteristics of the music, including pitch, timbre, and rhythm, providing a rich foundation for emotion recognition. Preprocessing ensures that the features are normalized and standardized, enabling the system to handle diverse music samples effectively. The system then applies a CNN-based classification module to learn hierarchical patterns from the extracted features. Convolutional layers automatically detect local correlations in the audio features, while pooling layers reduce dimensionality and highlight the most dominant emotional patterns. The modular design allows the system to scale to large music datasets and adapt to multiple emotion categories, supporting real-world applications such as music recommendation, playlist generation, and personalized music analytics.

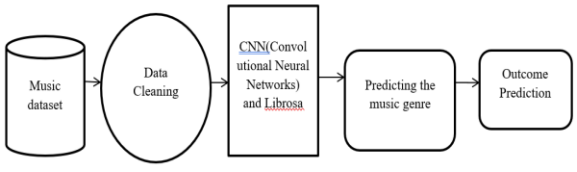


Fig. 1: System Overview

3.1 Collecting Data:

In this module, a diverse dataset of music tracks is gathered from open sources, streaming platforms, or publicly available music emotion datasets. Each track is labeled with the corresponding emotional category, such as happy, sad, angry, or calm. The dataset includes multiple genres to ensure the model can generalize across different musical styles. Proper collection ensures sufficient representation of emotions, enabling the model to learn effectively.

3.2 Analyzing the Information:

This module focuses on exploring the collected music data to understand its characteristics. Statistical analysis and visualizations are used to examine distributions of emotion classes, track lengths, and audio quality. This step also helps identify imbalances in the dataset, such as underrepresented emotions, which can impact model performance. Analysis guides subsequent preprocessing and feature extraction strategies.

3.3 Preprocessing Data:

Preprocessing prepares raw audio data for feature extraction and model input. This includes trimming or padding tracks, normalizing audio signals, and converting them into a consistent sampling rate. Noise reduction and silence removal may also be applied. Using Librosa, features such as MFCCs, chroma, spectral contrast, and

tonnetz are extracted to represent each track in a compact, informative format.

3.4 Running the Model:

In this module, the Convolutional Neural Network (CNN) is trained on the extracted audio features. The CNN's convolutional layers capture local dependencies and patterns within the features, while pooling layers reduce dimensionality and highlight dominant emotional cues. The model learns hierarchical representations, enabling it to distinguish between different emotions effectively.

3.5 Fine Tuning the Model:

Fine-tuning involves adjusting hyperparameters of the CNN, such as the number of convolutional layers, filter sizes, pooling strategies, batch size, and learning rate. Techniques like dropout and regularization are applied to prevent overfitting. This module ensures the model achieves optimal accuracy and generalizes well across unseen music samples.

3.6 Model Efficiency:

This module evaluates the performance of the trained CNN using metrics such as accuracy, precision, recall, F1-score, and confusion matrices. Efficiency is analyzed in terms of prediction speed and computational requirements. Model efficiency ensures the system is suitable for real-time or large-scale deployment on music platforms.

3.7 Forecasting Results:

In the final module, the system predicts the emotional labels of new or unseen music tracks. Predicted emotions are compared with actual labels to validate performance. The results are visualized using graphs or dashboards, enabling users to interpret emotional patterns in music. This module demonstrates the practical utility of the model in music recommendation and playlist generation systems.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The proposed music emotion classification system was evaluated using accuracy, precision, recall, and F1-score metrics. Audio features extracted using **Librosa**, including MFCCs, chroma features, and spectral contrast, were used to train the neural network model. Experimental results showed that the model achieved high classification accuracy in identifying emotions such as happy, sad, calm, and energetic from music samples.



Fig. 2: Home Page



Fig. 3: Registration



Fig. 4: Login

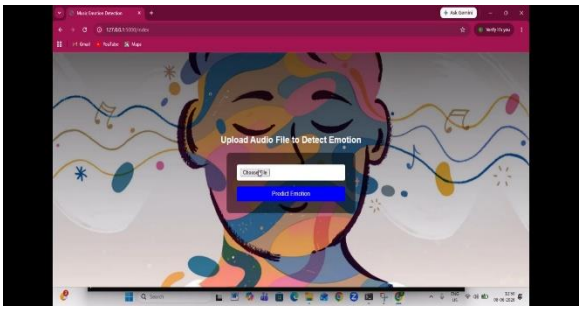


Fig. 5: Upload Audio

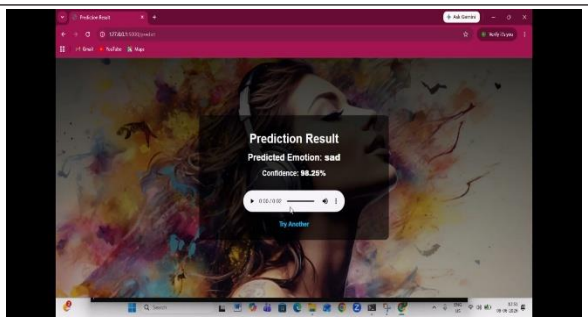


Fig. 6: Predict Emotion

V. CONCLUSION

Music emotion classification is a vital component of modern music platforms, enabling personalized recommendations, intelligent playlist generation, and enhanced user engagement. This project presented a deep learning framework that combines Librosa-based feature extraction with Convolutional Neural Networks (CNN) for effective emotion recognition. Librosa was employed to extract audio features such as MFCCs, chroma, spectral contrast, and tonnetz, capturing the timbral, harmonic, and rhythmic aspects of music. CNN was then applied to learn hierarchical patterns and automatically detect emotional cues in the extracted features. The hybrid approach eliminates the need for handcrafted features and ensures adaptability across diverse genres and styles. Through training, fine-

tuning, and evaluation, the model demonstrated its ability to generalize well and provide accurate classification of emotions. Performance metrics such as accuracy, precision, recall, and F1-score validate the robustness of the system. The results show that combining feature extraction with CNN significantly improves recognition of complex emotional relationships in music. This project not only addresses challenges in music information retrieval but also contributes to the growing field of affective computing. The developed framework can be integrated into real-world platforms for music recommendation and analytics, improving user experience and satisfaction. With further research and enhancements, the system has the potential to become a scalable, real-time solution for emotion-aware music applications, making it a valuable contribution to the future of intelligent music technologies.

VI REFERENCES

[1] L. Zhou, ‘‘Cultivation of artistic expression in college music and vocal music teaching,’’ Art Perform. Lett., vol. 4, no. 12, pp. 43–49, 2023.

[2] S. Ding, ‘‘Research on the artistic expression of vocal music,’’ in Proc. 2nd Int. Conf. Culture, Educ. Econ. Develop. Modern Soc. (ICCESE). Atlantis Press, 2018, pp. 663–665.

[3] A. Sabbadini, ‘‘Opera on the couch: Music, emotional life, and unconscious aspects of music,’’ Int. J. Psychoanalysis, vol. 104, no. 1, pp. 183–185, Jan. 2023.

[4] Q. Xianyang, ‘‘Research of lens model in music emotional communication,’’ BioTechnol, Indian J., vol. 10, p. 19, 2015.

[5] C. Nussbaum, A. Schirmer, and S. R. Schweinberger, ‘‘Electrophysiological correlates of vocal emotional processing in musicians and non-musicians,’’ Brain Sci., vol. 13, no. 11, p. 1563, Nov. 2023.

[6] J. J. Campos-Bueno et al., ‘‘Emotional dimensions of music and painting and their interaction,’’ Spanish J. Psychol., vol. 18, p. E54, 2015.

AUTHORS Profile



Mr. ADALA AMARNATH has received his B.Tech-Computer Science and Engineering from JNT-UH, Hyderabad in 2010, M.Tech-CNIS degree in Computer Science and Engineering from JNT-UH, Hyderabad in 2013 and Pursuing Ph.D degree in Computer Science and Engineering from ICFAI University Hyderabad. He is dedicated to teaching field from the last 12 years. His research area Quantum Computing and Aptitude

Computing. At present he is working as Assistant Professor in CSE Department at Guru Nanak Institutions Technical Campus, Ibra-himpatnam, Telangana, India.



Mr. UPPALA HARSHITH pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Cam-pus affiliated to JNTUH in 2022-2026.



Ms. MENGANI SHIVATHMIKA pursuing B.Te-ch degree in Computer Sci-ence and Engineering at Guru Nanak Institutions Technical Campus affil-iated to JNTUH in 2022-2026.



Ms. NALLA INDU pur-suing B.Tech degree in Computer Science and En-gineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



Mr. NUNEMUNTHALA SHASHANK pursuing B.Tech degree in Com-puter Science and Eng-ineering at Guru Nanak Institutions Technical Ca-mpus affiliated to JNTUH in 2022-2026.