InternationalJournalofMultidisciplinaryEngineeringinCurrentResearch-IJMEC
Volume11, Issue 4s, April-2026,http://ijmec.com/,ISSN:2456-4265

Full Length Article

Music Emotion Classification With Neural Network Architecture And Librosa

Syed Aiyan[1], Syed Wajahat Ali[2], Mohammed Fardeen Younus[3], Mr. Mohammed Rahmat Ali[4]
[1,2,3]B.E Student's; Department Of Computer Science And Engineering, ISL Engineering College, Hyderabad, India.
[4]Assistant Professor; Department Of Computer Science And Engineering, ISL Engineering College, Hyderabad, India.
Mail Id;syedaiyan1820@gmail.com, aliwajahatsyed04@gmail.com, fardeenyounus786@gmail.com
Accepted 24-04-2026

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.

Keywords: *Music Emotion Classification, Convolutional Neural Network (CNN), Librosa, Audio Feature Extraction, Mel-Frequency Cepstral Coefficients (MFCCs), Chroma Features, Spectral Contrast, Tonnetz, Deep Learning, Audio Signal Processing, Emotion Recognition, Music Recommendation System, Playlist Generation, Music Analytics, Machine Learning, Pattern Recognition, Automated Classification, Digital Audio Processing, Intelligent Music Systems, Artificial Intelligence in Music.*

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

https://doi.org/10.63665/IJMEC.1104s.14
ISSN: 2456-4265
IJMEC 2026

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.

LITERATURE REVIEW:

Han, Chen, and Ban (2023) proposed a music emotion recognition framework based on a neural network with an inception-GRU residual architecture. The study combined inception modules with GRU residual connections to capture both multi-scale timbral patterns and temporal dependencies in music signals. Audio features such as MFCCs and spectral characteristics were extracted using preprocessing pipelines similar to Librosa. The inception blocks improved contextual feature learning, while residual connections enabled deeper model training without vanishing gradient issues. Experimental evaluation on music-emotion datasets demonstrated improved accuracy across valence and arousal dimensions compared to conventional CNN and RNN approaches. The authors also emphasized the scalability and lightweight nature of the model, making it suitable for real-world and edge-deployment scenarios.

George M.W. (2024) introduced a deep neural network framework for instrument emotion recognition from polyphonic instrumental music using MFCC and Chroma Energy Normalized Statistics (CENS) features. The proposed system utilized compact audio representations generated through audio processing toolkits comparable to Librosa. MFCCs were employed to capture timbral properties, while CENS features represented harmonic structures essential for emotional interpretation. The deep neural network effectively classified emotions across various genres and instrumental combinations. The study reported superior performance over single-feature approaches and highlighted the robustness of combined timbral and harmonic feature representations, particularly for instrumental music lacking lyrical content. The framework also demonstrated suitability for real-time playlist generation and music recommendation applications.

Wang et al. (2023) developed a hierarchical audio-visual information fusion framework with multi-label joint decoding for Music Emotion Recognition (MER). The model extracted deep representations from both audio and visual modalities using pre-trained foundation models. Attention-guided feature aggregation modules were introduced to effectively combine multimodal information, while a joint decoding mechanism simultaneously handled discrete emotion classification and valence regression tasks. A multi-task uncertainty loss function improved prediction consistency across emotion categories and intensity levels. The framework achieved top-three ranking in the MER-Multi Challenge benchmark, demonstrating strong generalization capabilities. Although the approach incorporated visual information, the proposed attention-based fusion strategy can be adapted for integrating audio-only features such as MFCCs, chroma, and CNN embeddings in pure audio emotion recognition systems.

Makhmudov, Kutlimuratov, and Cho (2024) presented a hybrid LSTM-attention and CNN architecture for speech emotion recognition. Although the work focused on speech

https://doi.org/10.63665/IJMEC.1104s.14
ISSN: 2456-4265
IJMEC 2026

signals, its architecture is highly relevant to music emotion recognition tasks. The model utilized CNN layers to extract spatial information from spectrogram-based inputs, while LSTM and attention mechanisms captured temporal emotional dynamics. Feature representations were generated using spectrograms similar to Librosa-derived MFCCs and Mel-spectrograms. The attention mechanism enabled the network to focus on emotionally significant temporal segments, thereby improving classification performance and generalization. Experimental results demonstrated significant improvements over standalone CNN and RNN models. The modular architecture also supported lightweight and real-time inference, making it adaptable for streaming-based music emotion recognition systems.

Donatus et al. (2024) conducted a comparative analysis of spectrogram and MFCC representations for emotion recognition using machine learning techniques. The study investigated the impact of different feature extraction methods on emotion classification performance. Experimental findings revealed that MFCC features consistently outperformed raw spectrogram representations due to their compactness and perceptual relevance. The authors emphasized that MFCCs preserve discriminative emotional characteristics while reducing feature dimensionality, resulting in improved computational efficiency. Additionally, the study recommended combining MFCCs with complementary features such as spectral contrast and chroma descriptors to enhance emotional representation. These findings strongly support the use of Librosa-based MFCC extraction pipelines in CNN-driven music emotion recognition systems.

METHODOLOGY:

Modules Name:

- Collecting Data
- Analyzing the information
- Preprocessing Data
- Running the model
- Fine Tuning the model
- Model Efficiency
- Forecasting Results

MODULES EXPLANATION:

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.

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.

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.

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.

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.

Model Efficiency:

This module evaluates the performance of the trained

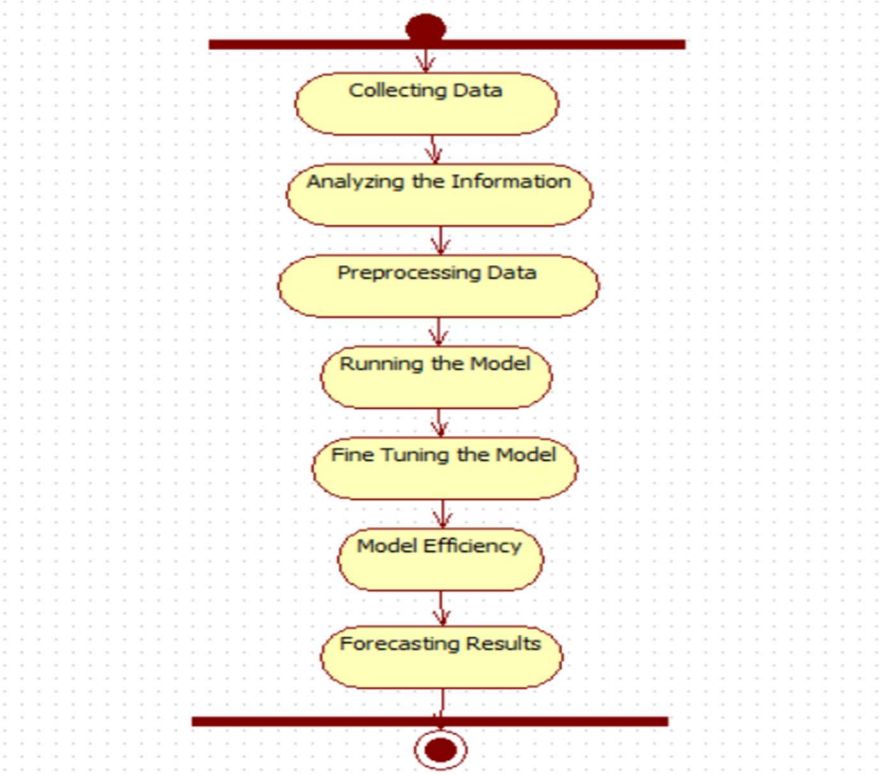
https://doi.org/10.63665/IJMEC.1104s.14
ISSN: 2456-4265
IJMEC 2026

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.

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.

IMPLEMENTATION:



TECHNIQUE USED OR ALGORITHM USED

EXISTING TECHNIQUE:

The existing system uses a Heterogeneous Graph Neural Network (HGN), which models different types of nodes and relationships in a graph structure. In the context of music emotion classification, HGN represents singers, composers, and listeners as nodes, and their interactions (such as genre preferences, emotional expression, or collaboration) as edges. The algorithm learns meaningful feature representations for each node by aggregating information from its neighbors, using techniques like attention mechanisms and meta-path learning to handle heterogeneous data. HGN aims to capture the complex relational structure among contributors and consumers of music to improve classification performance. However, while HGN is effective in learning structured relationships, it may struggle with generalization when faced with limited or imbalanced data. Also, since it depends heavily on graph construction and metadata availability, it may not fully capture the underlying emotion-generative process in music, especially when emotional cues are subtle or not well represented in the graph structure.

PROPOSED TECHNIQUE USED OR ALGORITHM USED:

The proposed algorithm first uses Librosa to extract relevant audio features from music tracks. MFCCs capture the spectral envelope and timbral aspects of sound, chroma features represent harmonic content, and spectral contrast emphasizes differences between peaks and valleys in the spectrum. These features form a compact yet expressive input for the CNN, allowing

https://doi.org/10.63665/IJMEC.1104s.14
ISSN: 2456-4265
IJMEC 2026

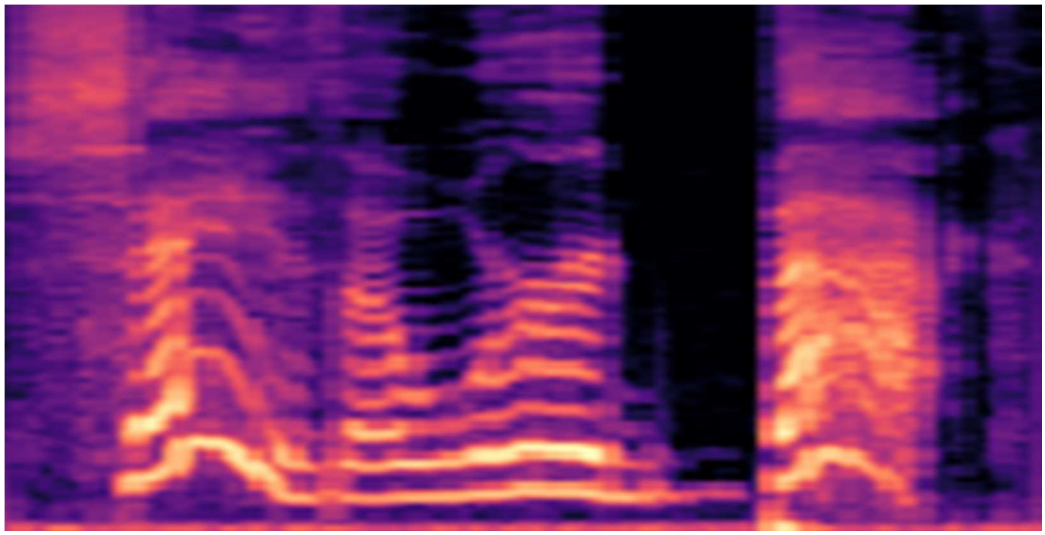
the algorithm to focus on emotion-related characteristics rather than raw audio signals. Next, a CNN architecture is applied to learn hierarchical feature representations and classify emotions. Convolutional layers detect patterns within the input features, while pooling layers reduce complexity and retain key emotional cues. Fully connected layers integrate the learned patterns to predict the final emotion category. The algorithm is trained using supervised learning with backpropagation, ensuring that it generalizes across diverse music styles and emotional expressions. This combination of feature extraction and deep learning provides a robust and scalable approach to automated music emotion classification.

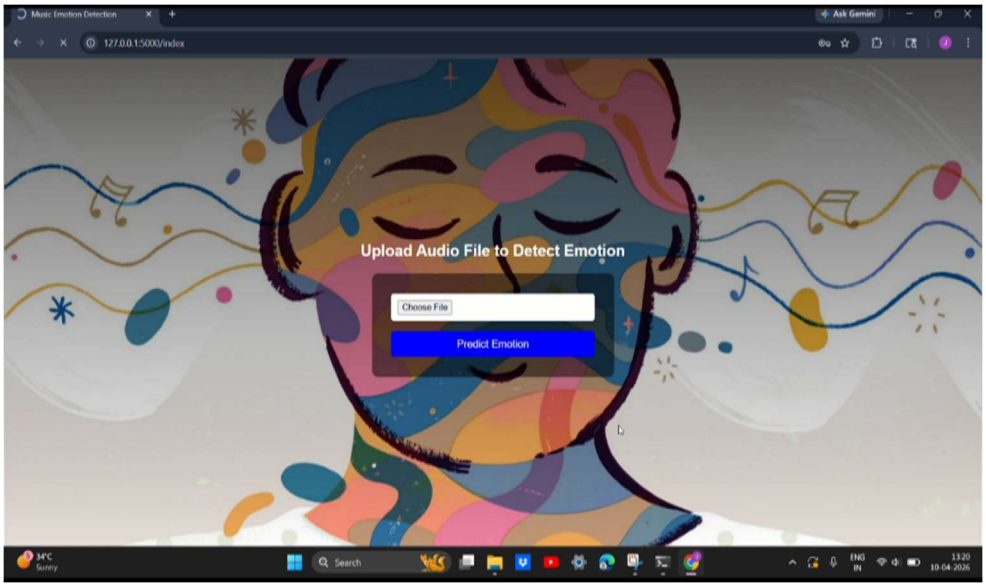
RESULTS:

YOUR INPUT:

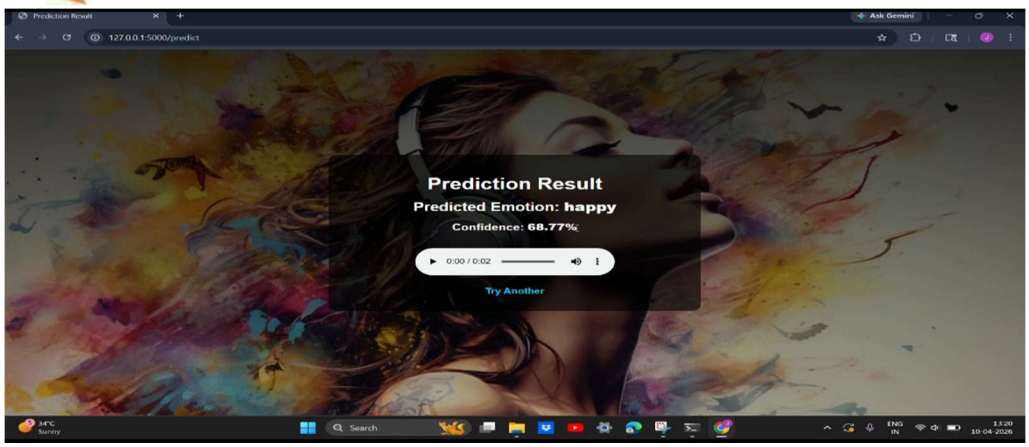
SPECTROGRAM OF INPUT AUDIO SIGNAL:
The Spectrogram Shows The Frequency And Intensity Variations OfThe Input Audio Over Time. These Features Are Used ByThe Model To Classify The Emotion (e.g., happy, neutral..)

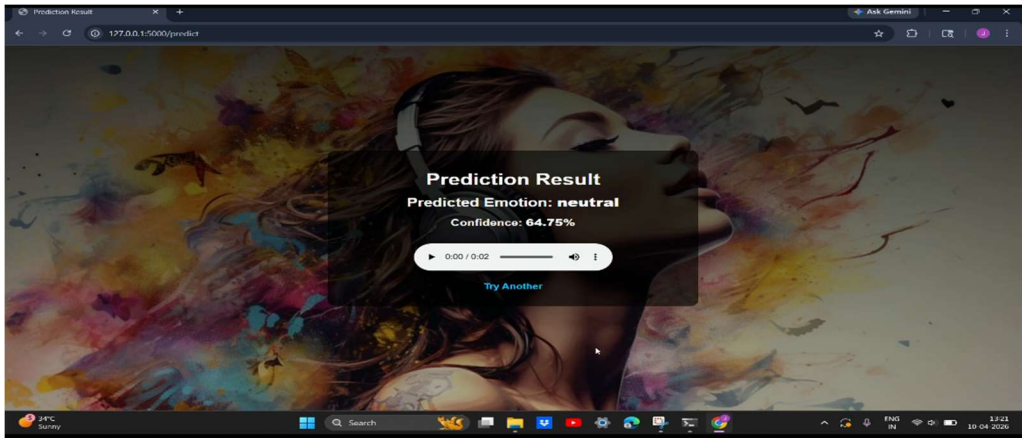
YOUR OUTPUT:





https://doi.org/10.63665/IJMEC.1104s.14
ISSN: 2456-4265
IJMEC 2026





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.

FUTURESCOPE

The proposed music emotion classification framework can be enhanced by integrating multimodal data such as song lyrics, metadata, and user feedback in addition to audio features. Incorporating advanced deep learning models like CNN-LSTM hybrids or Transformers can improve the ability to capture both temporal and contextual patterns. A larger and more diverse dataset covering multiple cultures and languages can enhance generalization. Real-time streaming analysis can be added to classify emotions while the music is being played. Transfer learning can be applied using pre-trained audio models to boost performance with limited data. Future versions may focus on detecting subtle or mixed emotions instead of

https://doi.org/10.63665/IJMEC.1104s.14
ISSN: 2456-4265
IJMEC 2026

only distinct categories. Explainable AI techniques can be included to highlight which features contribute most to predictions, improving interpretability. Mobile and cloud-based deployment can provide emotion-aware recommendations to end-users instantly. Integration with popular music platforms will expand its practical use. Finally, the system can evolve into a complete personalized music assistant based on emotional states.

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.
[7] T. Fischinger, M. Kaufmann, and W. Schlotz, ‘‘If it’s mozart, it must be good? The influence of textual information and age on musical appreciation,’’ Psychol. Music, vol. 48, no. 4, pp. 579–597, Jul. 2020.
[8] X. Cai and H. Zhang, ‘‘Music genre classification based on auditory image, spectral and acoustic features,’’ Multimedia Syst., vol. 28, no. 3, pp. 779–791, Jun. 2022.
[9] B. Wilkes, I. Vatolkin, and H. Müller, ‘‘Statistical and visual analysis of audio, text, and image features for multi-modal music genre recognition,’’ Entropy, vol. 23, no. 11, p. 1502, Nov. 2021.
[10]N. Zeng, P. Wu, Z. Wang, H. Li, W. Liu, and X. Liu, ‘‘A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection,’’ IEEE Trans. Instrum. Meas., vol. 71, pp. 1–14, 2022.
[11]Y. Dong, Q. Liu, B. Du, and L. Zhang, ‘‘Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification,’’ IEEE Trans. Image Process., vol. 31, pp. 1559–1572, 2022.
[12]D. Pathak and U. S. N. Raju, ‘‘Content-based image retrieval for superresolutioned images using feature fusion: Deep learning and hand crafted,’’ Concurrency Comput., Pract. Exper., vol. 34, no. 22, 2022, Art. no. e6851.
[13]C. Yuan, Q. Ma, J. Chen, W. Zhou, X. Zhang, X. Tang, J. Han, and S. Hu, ‘‘Exploiting heterogeneous artist and listener preference graph for music genre classification,’’ in Proc. 28th ACM Int. Conf. Multimedia, Oct. 2020, pp. 3532–3540.
[14]Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, ‘‘A comprehensive survey on graph neural networks,’’ IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 1, pp. 4–24, Jan. 2021.
[15] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, ‘‘Graph neural networks: A review of methods and applications,’’ 2018, arXiv:1812.08434.
[16]W. Fan, Y. Ma, Q. Li, Y. He, E. Zhao, J. Tang, and D. Yin, ‘‘Graph neural networks for social recommendation,’’ in Proc. World Wide Web Conf., 2019, pp. 417–426.
[17]X. Wang et al., ‘‘Heterogeneous graph attention network,’’ in Proc. World Wide Web Conf., 2019, pp. 2022–2032.
[18]R. Bing, G. Yuan, M. Zhu, F. Meng, H. Ma, and S. Qiao, ‘‘Heterogeneous graph neural networks analysis: A survey of techniques, evaluations and applications,’’ Artif. Intell. Rev., vol. 56, no. 8, pp. 8003–8042, Aug. 2023.
[19]C. Shi, ‘‘Heterogeneous graph neural networks,’’ in Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2019, pp. 793–803.
[20]V. Mounika and Y. Charitha, ‘‘Mood-Enhancing music recommendation system based on audio signals and emotions,’’ in Proc. Int. Conf. Inventive Comput. Technol. (ICICT), Apr. 2023, pp. 1766–1772.
[21]X. Song et al., ‘‘Automatic recognition of uterine contractions with electrohysterogram signals based on the zero-crossing rate,’’ Sci. Rep., vol. 11, no. 1, p. 1956, 2021.
[22]B. Baris, M. E. Cek, and D. G. Kuntalp, ‘‘Modulation classification of MFSK modulated signals using spectral centroid,’’ Wireless Pers. Commun., vol. 119, no. 1, pp. 763–775, Jul. 2021.
[23]D. H. Rudd et al., ‘‘Leveraged mel spectrograms using harmonic and percussive components in speech emotion recognition,’’ in Proc. PacificAsia Conf. Knowl. Discovery Data Mining. Cham, Switzerland: Springer, 2023, pp. 392–404.
[24]Y. Zhang, G. Kolkman, and H. Watanabe, ‘‘Phase repair for timedomain convolutional neural networks in music super-resolution,’’ 2023, arXiv:2306.11282.
[25]N. J. O’Leary, ‘‘The tempest presented by the lord Denney’s players, and: The tempest presented by the

https://doi.org/10.63665/IJMEC.1104s.14
ISSN: 2456-4265
IJMEC 2026

Cincinnati Shakespeare company,'' Shakespeare Bull., vol. 35, no. 3, pp. 487–495, 2017.
[26]T. N. Kipf and M. Welling, ''Semi-supervised classification with graph convolutional networks,'' 2016, arXiv:1609.02907.
[27]P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, ''Graph attention networks,'' 2017, arXiv:1710.10903.
[28]B. Perozzi, R. Al-Rfou, and S. Skiena, ''Deepwalk: Online learning of social representations,'' in Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2014, pp. 135–144.
[29]Y. Dong, N. V. Chawla, and A. Swami, ''metapath2vec: Scalable representation learning for heterogeneous networks,'' in Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, New York, NY, USA, Aug. 2017, pp. 135–144.
[30]J. M. Keller, M. R. Gray, and J. A. Givens, ''A fuzzy K-nearest neighbor algorithm,'' IEEE Trans. Syst., Man, Cybern., vols. SMC–15, no. 4, pp. 580–585, Jul. 1985.
[31]G. Zhao et al., ''Review-driven multi-label music style classification by exploiting style correlations,'' 2018, arxiv:1808.07604.
[32]Q. Ma, C. Yuan, W. Zhou, J. Han, and S. Hu, ''Beyond statistical relations: Integrating knowledge relations into style correlations for multi-label music style classification,'' in Proc. 13th Int. Conf. Web Search Data Mining, Jan. 2020, pp. 411–419.
[33]L. Fanioudakis and I. Potamitis, ''Deep networks tag the location of bird vocalisations on audio spectrograms,'' 2017, arXiv:1711.04347.

https://doi.org/10.63665/IJMEC.1104s.14
ISSN: 2456-4265
IJMEC 2026