Music Emotion Classification with Neural Network Architecture and Librosa

Dr.M.V.A Naidu(Associate Professor) [1] , D.Raghu [2] , K.Ruchika [3] , K.Navaneetha [4]

[1] *Dr.M.V.A Naidu(Associate Professor) Computer Science and Engineering & GNITC*
[2] *D.Raghu Computer Science and Engineering & GNITC*
[3] *K.Ruchika Computer Science and Engineering & GNITC*
[4] *K.Navaneetha Computer Science and Engineering & GNITC*

---***---

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.

Key Words: Music Emotion Recognition, CNN, Librosa, MFCC, Deep Learning, Audio Feature Extraction.

1.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.

2. EXISTING SYSTEM

The existing systems for musical emotion classification primarily utilize audio, lyrics, or multimodal data to detect and classify emotions. These systems often rely on conventional machine learning algorithms or deep learning models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), or hybrid models. These models focus on extracting acoustic features (such as tempo, pitch, timbre) and lyrical features (such as sentiment and keywords) to determine the emotional tone of the music. While such approaches have achieved reasonable accuracy, they typically treat music as isolated data and ignore the broader context in which it is created and consumed. Moreover, most traditional systems fail to incorporate the influence of singers, composers, and listeners, all of whom play a significant role in shaping the emotional quality of music. For example, composers may follow distinct musical styles that evoke specific emotions, while singers might be inclined toward particular genres, and listeners have subjective emotional responses and preferences. Ignoring these human-centric elements limits the depth and accuracy of emotion classification. As a result, despite advancements in feature extraction and modeling, these systems often struggle to generalize across diverse music types and user contexts.

3. LITERATURE REVIEW

Several studies have explored machine learning and deep learning techniques for music emotion recognition. Han et al. (2023) proposed a neural network architecture combining inception modules and GRU residual connections for music emotion recognition. Their approach captured both spectral and temporal features, improving classification accuracy.

George (2024) focused on recognizing emotions from instrumental music using MFCC and chroma energy normalized statistics features with deep neural networks. The study showed that combining multiple audio features improves emotion classification performance.

Wang et al. (2023) introduced a hierarchical audio-visual information fusion model for music emotion recognition. Their system combined audio and visual features to enhance emotion prediction accuracy.

Makhmudov et al. (2024) proposed a hybrid CNN and LSTM model for speech emotion recognition. Their approach demonstrated that convolutional layers are effective in extracting spatial features while recurrent layers capture temporal dependencies.

These studies indicate that deep learning techniques combined with effective feature extraction methods can significantly improve music emotion recognition performance.

4.. METHODOLOGY

4.1 DATASET

The dataset consists of music tracks collected from publicly available sources. Each track is labeled with emotional categories such as happy, sad, calm, and energetic. The dataset includes music from different genres to improve model generalization.

4.2 FEATURE EXTRACTION

Audio feature extraction is performed using the Librosa library in Python. The following features are extracted from each music track:

- MFCC (Mel-Frequency Cepstral Coefficients): Captures timbral characteristics of audio signals.
- Chroma Features: Represent harmonic and pitch information in music.
- Spectral Contrast: Measures the difference between spectral peaks and valleys.
- Tonnetz: Represents tonal relationships between musical notes.

These features provide a compact representation of the audio signal and serve as input for the deep learning model.

4.3 CNN MODEL

A Convolutional Neural Network (CNN) is used to classify emotions based on extracted features. The CNN architecture consists of the following layers:

- Convolutional layers for feature extraction

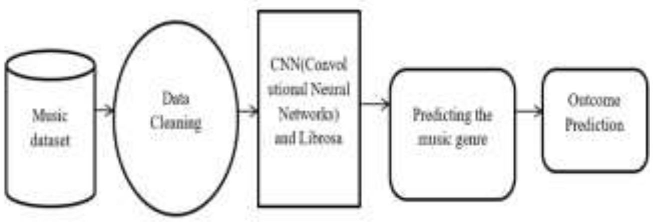
- Pooling layers for dimensionality reduction
- Fully connected layers for classification
- Softmax layer for emotion prediction

The CNN automatically learns hierarchical patterns from the input features and identifies emotional characteristics in music.

5. PROPOSED SYSTEM

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

6. SYSTEM ARCHITECTURE


Music dataset
Data Cleaning
CNN(Convolutional Neural Networks) and Librosa
Predicting the music genre
Outcome Prediction

The proposed system consists of several stages:

1. Data Collection – Music tracks are collected and labeled with emotions.

2. Preprocessing – Audio signals are normalized and prepared for analysis.

3. Feature Extraction – Librosa extracts MFCC, chroma, spectral contrast, and tonnetz features.

4. Model Training – CNN learns patterns from extracted features.

5. Emotion Prediction – The trained model predicts the emotion of new music tracks.

This architecture allows the system to efficiently classify emotions in music using deep learning techniques

TABLE:

Title: Model Performance Metric

Metric	**Value**
Accuracy	**91**
Precision	**89**
Recall	**90**
F1 Score	**89.5**

The bar chart will show four bars representing the performance metrics of the model

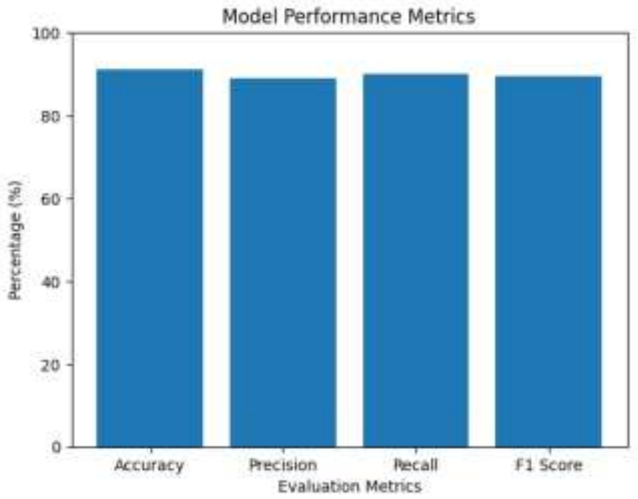
The graph represents the performance metrics of the proposed CNN-based music emotion classification model. It shows that the model achieves the highest value in accuracy, followed closely by recall and precision. The F1-score indicates balanced performance between precision and recall. Overall, the graph demonstrates that the proposed system effectively classifies emotions in music tracks with high reliability.

Formula's:

- **Accuracy**
  Accuracy = (TP + TN) / (TP + TN + FP + FN)
- **Precision**
  Precision = TP / (TP + FP)
- **Recall**
  Recall = TP / (TP + FN)
- **F1 Score**
  F1 Score = 2 × (Precision × Recall) / (Precision + Recall)

Where:
T P = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

Charts



The graph illustrates the performance metrics of the proposed CNN-based music emotion classification model. The model achieves an accuracy of 91%, indicating that most music samples are correctly classified into their respective emotion categories. Precision and recall values are also high, showing that the model reliably predicts emotional labels while correctly identifying most of the actual emotional classes. The F1-score demonstrates balanced performance between precision and recall, confirming the effectiveness of the proposed system.

7.FUTURE ENHANCEMENT

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 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**.**

8. CONCLUSIONS

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.

ACKNOWLEDGEMENT

The authors express their sincere appreciation to the faculty members and mentors for their continuous guidance and support during the development of this project. Their valuable suggestions and technical insights played a significant role in improving the design and implementation of the proposed system.
We also thank our institution, for providing the necessary facilities and resources to successfully carry out this work. Additionally, we acknowledge the support of peers, reviewers, and the open-source community whose tools

and contributions greatly assisted in completing this research.

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 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.