Classifying Indian Music Genres with Artificial Intelligence

[1*]Abhishek Giri, [2]Anand Kumar Mishra, [3]C.S. Raghuvanshi

[1*]Research Scholar, Dept. of Computer Science Engineering, Faculty of Engineering & Technology, Rama University Uttar Pradesh, Mandhana, Kanpur, U.P., India
[2]Assistant Professor, Dept. of Computer Science Engineering, Faculty of Engineering & Technology, Rama University Uttar Pradesh, Mandhana, Kanpur, U.P., India
[3]Professor, Dept. of Computer Science Engineering, Faculty of Engineering & Technology, Rama University Uttar Pradesh, Mandhana, Kanpur, U.P., India

Abstract:

In today's digital age, automated music recommendation systems based on human emotions are gaining significant traction. This paper introduces a novel approach aimed at enhancing music mood classification accuracy by integrating both audio and lyrical modalities within a fusion model. The primary objective is to investigate the effectiveness of various attention mechanisms, including self-attention (SA), channel attention (CA), and hierarchical attention network (HAN), within a multi-modal framework tailored for music mood classification. Through rigorous experimentation, we demonstrate that multi-modal architectures enriched with attention mechanisms outperform their non-attention counterparts and single-modal architectures, achieving higher accuracy rates. Motivated by the promising performance of attention mechanisms, we propose a groundbreaking network architecture, the HAN-CA-SA based multi-modal classification system, which boasts an impressive accuracy of 82.35%. Additionally, we evaluate the proposed model using ROC and Kappa metrics, validate its robustness through Kfold cross-validation, and conduct comparative analyses with existing systems such as XLNet and CNN-BERT, supported by a statistical hypothesis test.

Keywords:

Automated music recommendation systems, Human emotions, Multi-modal fusion model, Attention mechanisms (SA, CA, HAN),Music mood classification, Multi-modal architectures

Introduction:

Music Information Retrieval (MIR) primarily involves extracting and analyzing features from music, indexing music based on these features, and devising search and retrieval methods to access music. The field has experienced significant growth in recent decades, fueled by advancements such as audio compression, increased computing power in personal computers for feature extraction, and widespread availability of mobile music players and online streaming services. MIR finds application in various areas including music recommendation, playlist generation, audio fingerprinting, music browsing interfaces, popularity estimation, computational music theory, visualization, and creation. The availability of extensive online music libraries has led to a demand for better management, indexing, searching, and organization of music data. Music browsing interfaces that offer users a serendipitous music experience using online resources are gaining popularity. Effective MIR techniques can provide listeners with personalized and relevant music recommendations. This paper aims to develop an automatic music mood classification system by leveraging both lyrical and audio features. Music has a profound impact on human emotions and can evoke a wide range of feelings. Listening to music has numerous psychological benefits, including stress management, emotional well-being, energy boost, pain relief, and relaxation. Music transcends language barriers and can evoke universal emotions. The relationship between music and emotions has been a subject of research for decades, with music recommendation systems leveraging listeners' moods for indexing and retrieval. Studies suggest that effective music retrieval systems should allow users to search based on the social and psychological functions of music, focusing on stylistic, mood, and similarity information. Thus, an automated music mood classification system based on listeners' moods or preferences can significantly improve music recommendation accuracy. Music mood classification typically relies on acoustic features such as spectral and rhythmic characteristics. However, lyrics also play a significant role in influencing mood, leading to the development of lyrics-based classification systems using natural language processing techniques. Combining audio and lyrics features can result in more accurate mood classification systems.

This research aims to address the challenge of music mood classification using a multi-modal attention framework. Key contributions include a detailed exploration of musical mood classification using the multi-modal framework, analysis of channel and spatial attention mechanisms, and comparison of classifiers using statistical tests. Previous research in music mood classification includes approaches based on acoustic data, hierarchical frameworks, and emotion-based classification models using lyrics. Deep learning algorithms like CNN and RNN, along with attention mechanisms, have shown promise in improving classification performance. Transfer learning and hybrid machine learning approaches have also been explored for multi-modal mood classification. Challenges in automatic music mood classification include subjective interpretation of music emotions, selection of mood taxonomy models, and scarcity of labeled datasets. However, advancements in multi-modal architectures and attention mechanisms offer promising solutions to these challenges. finally, this work focuses on developing an effective automatic music mood classification system using a multi-modal framework and attention mechanisms, considering both audio and lyrics features. Experimental validation of the proposed model's efficacy and statistical tests are conducted to demonstrate its effectiveness.

Literature Review:

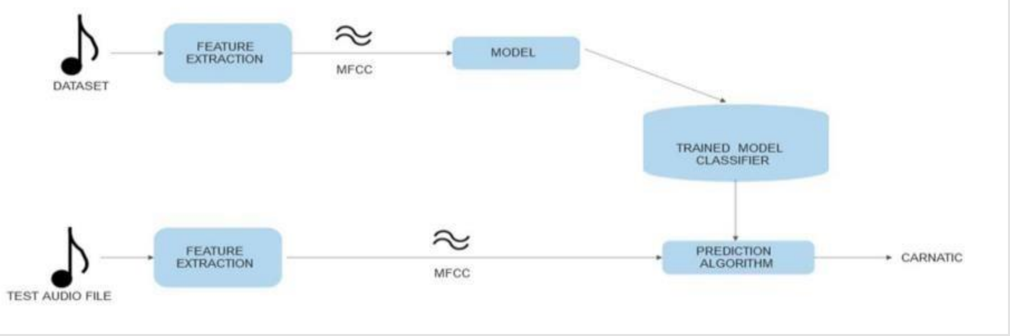
A review of existing literature in MIR underscores the significance of both acoustic and lyrical features in shaping music mood perception. While traditional approaches have predominantly focused on acoustic features, recent research highlights the crucial role of lyrical content in mood classification. Building upon this foundation, our study explores the integration of attention mechanisms into a multi-modal framework, with the aim of improving classification accuracy and advancing the state-of-theart in music mood classification. Hareesh Bahuleyan's research introduces a method for automatically categorizing music by assigning specific tags to songs within a user's collection using Machine Learning techniques. The study explores both modern Neural Network approaches and traditional methods to achieve this objective. The first approach utilizes Convolutional Neural Networks, where the characteristics of Mel Spectrograms, visual representations of audio data, are utilized to train the system comprehensively. The second method employs various Machine Learning algorithms such as Logistic Regression and Random Forest, among others, to extract manually crafted features from the temporal and frequency domains of audio recordings. Tzanetakis and Cook pioneered the application of machine learning algorithms for music genre classification, creating the GTZAN dataset, which remains a benchmark in this field. ChangshengXu et al. demonstrated the effectiveness of support vector machines (SVM) for this task, employing supervised learning techniques. Scheirer introduced a real-time beat tracking system using a filter bank and combination filters to detect the main beat and its strength in audio signals accompanied by music. The author proposed a multitudinous agent architecture to track multiple beat hypotheses. Tao expanded on the GTZAN dataset, employing restricted Boltzmann machines to achieve improved results compared to conventional neural networks. This study highlighted a data distribution issue in the GTZAN dataset and suggested the use of MFCC spectrograms for song preprocessing.

Methodology:

The methodology outlined in this paper entails a systematic process of feature extraction from audio and lyrical sources, followed by the training of multi-modal classification models augmented with attention mechanisms. These attention mechanisms, including SA, CA, and HAN, are strategically integrated to enhance the model's ability to capture salient features from both modalities. Evaluation of the models encompasses a comprehensive analysis using standard metrics such as accuracy, precision, recall, and F1-score, ensuring robust performance assessment across various dimensions of classification.

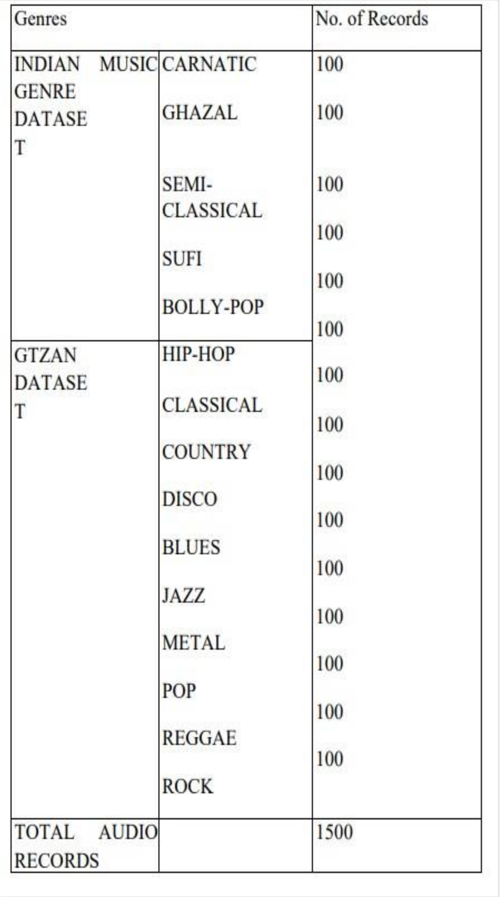
System Description:

The multi-modal architecture proposed in this study for music mood classification incorporates attention mechanisms, namely hierarchical attention network, self-attention, and channel attention. By integrating textual and acoustic features, the model aims to discern subtle nuances in music mood representation. The system operates on a dataset comprising song titles and mood labels, extracted from the MoodyLyrics Dataset. Through feature extraction and model training, the system endeavors to achieve superior classification accuracy compared to conventional unimodal approaches.



Dataset:

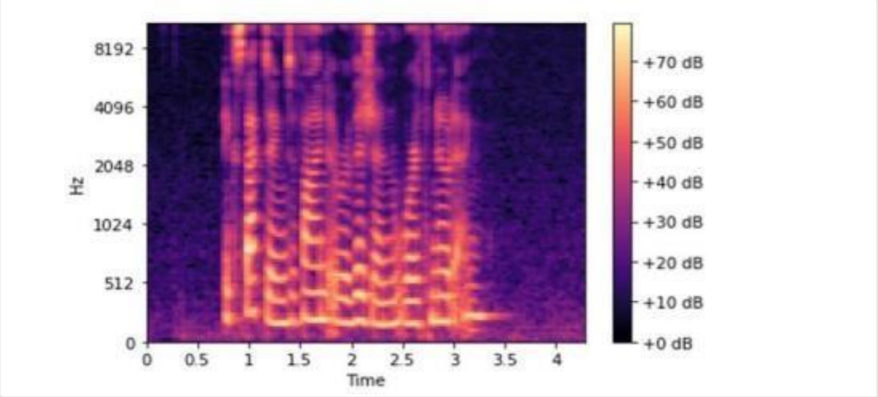
The dataset utilized in this study is a subset of the MoodyLyrics Dataset, comprising song titles alongside corresponding mood labels. While the dataset provides valuable metadata, including song titles and artist names, it poses challenges due to the scarcity of labeled instances for supervised learning. Nevertheless, through meticulous curation and preprocessing, a subset of 680 songs with associated audio and lyrical features is extracted for training and evaluation purposes.



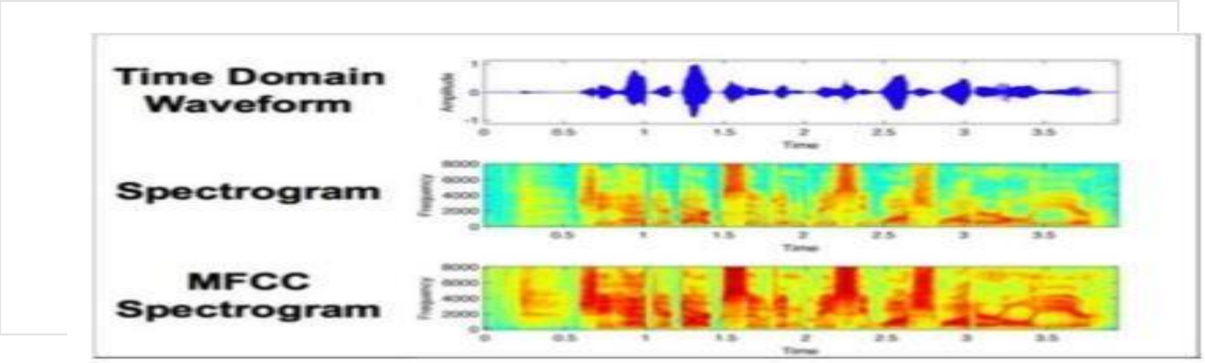
Genres		No. of Records
INDIAN MUSIC GENRE DATASET	CARNATIC	100
	GHAZAL	100
	SEMI-CLASSICAL	100
	SUFI	100
	BOLLY-POP	100
GTZAN DATASET	HIP-HOP	100
	CLASSICAL	100
	COUNTRY	100
	DISCO	100
	BLUES	100
	JAZZ	100
	METAL	100
	POP	100
	REGGAE	100
	ROCK	100
TOTAL AUDIO RECORDS		1500

Feature extraction is the initial step in music genre classification, involving the extraction of relevant characteristics and removal of noise from audio data.

Mel Spectrogram: Unlike a conventional Spectrogram, a Mel Spectrogram plots Frequency vs. Time using the Mel Scale on the y-axis and the Decibel Scale for color representation. This technique is preferred for deep learning models due to its clarity.



1) Mel Frequency Cepstral Coefficients (MFCC): MFCCs are modern features utilized in speech recognition studies. The extraction process involves dividing audio signals into smaller sets, isolating linguistic frequencies from noise, and applying the Discrete Cosine Transform (DCT) to preserve informative frequencies. MFCCs capture key information for classification, with the first few coefficients holding the most significance.



The conversion of frequency to the mel scale is determined by the formula m=2915·(1+f500), while the Fourier Transform is given by X (f) = ∫x (t) ∗ e−i2πftdt∞−∞. Logarithmic-axis representation of magnitudes integrates perceptual sensitivity, with the mel-scale emphasizing differences in pitch perception.

Model Training and Implementation:

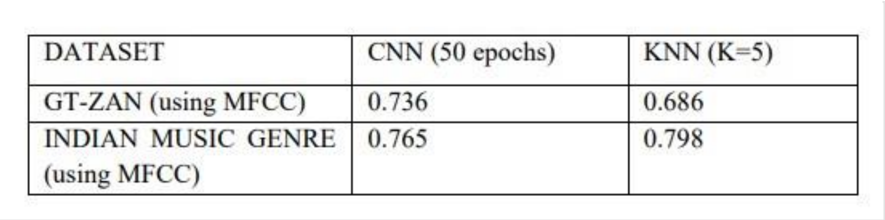
A. CNN (Convolutional Neural Network): Utilizing image representations, audio samples are converted to Mel

Spectrograms for CNN classification. Low-level features extracted from audio files using MFCC are mapped into a .json file for model training. The CNN model architecture is built using Keras, employing the Adam optimizer, RELU activation function for hidden layers, and Softmax function for output layers. The model's accuracy can be enhanced by adjusting epochs, but optimization may plateau beyond a certain threshold.

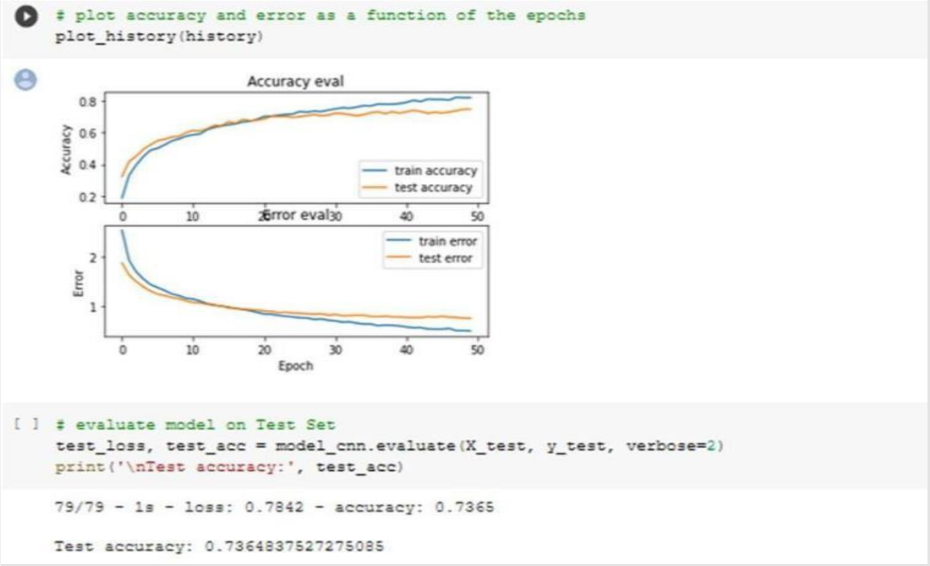
B. KNN (K-Nearest Neighbors): A fundamental machine learning approach, KNN explores various K values to optimize results. It is a non-parametric and supervised learning algorithm, relying solely on provided data without assumptions about the dataset. Unlike other methods, KNN does not separate data into training and test sets, utilizing all training data for predictions without generalization between sets.

Results and Analysis:

Performance evaluation of the proposed multi-modal architecture with attention mechanisms reveals significant improvements in classification accuracy compared to both uni-modal architectures and multi-modal architectures without attention. By analyzing precision, recall, F1 scores, and overall accuracy, it becomes evident that attention mechanisms contribute to enhanced model performance. Comparative analyses with existing models further underscore the efficacy of the proposed approach in music mood classification.



DATASET	CNN (50 epochs)	KNN (K=5)
GT-ZAN (using MFCC)	0.736	0.686
INDIAN MUSIC GENRE (using MFCC)	0.765	0.798



```
# plot accuracy and error as a function of the epochs
plot_history(history)
```

```
[ ] # evaluate model on Test Set
    test_loss, test_acc = model_cnn.evaluate(X_test, y_test, verbose=2)
    print('\nTest accuracy:', test_acc)
```

```
79/79 - 1s - loss: 0.7842 - accuracy: 0.7365

Test accuracy: 0.7364837527275085
```

Conclusion:

In summary and for future exploration, this project introduces an application utilizing Machine Learning for Music Genre Classification. Through an exploration of various audio feature extraction techniques, we determined MFCC to be the most effective for our purposes. Implementing KNN and CNN algorithms, we achieved peak accuracies of approximately 80% for KNN and 73% for CNN. However, these accuracies are subject to the quality and diversity of the dataset; a broader dataset encompassing a wide range of genres and tracks is necessary for improved predictions. Challenges arose due to limited dataset availability. Looking ahead, there's substantial potential for further development in this field, particularly given the broad applications of voice intelligence. While we attempted to incorporate both Western and Indian music genres, dataset constraints and window size limitations for the Discrete Fourier Transform posed challenges. With a larger dataset and more diverse audio tracks, we anticipate enhanced model accuracy. Additionally, we investigated the performance of various feature descriptors for

Indian music genres like classical (Carnatic music) and folk music. Besides, this project could evolve into a standalone web application deployable on websites, offering automatic classification of fed audio or video inputs. The intersection of music classification and machine learning holds promise for continued exploration and application in various domains.

References:

[1] Tzanetakis, G., & Cook, P. (2000). Automatic musical genre classification of audio signals. In Proc. of the IEEE International Conference on Multimedia and Expo (ICME).

[2] Tzanetakis, G., & Cook, P. (2001). Audio information retrieval using spectral similarity measures. In Proc. of the IEEE International Symposium on Circuits and Systems (ISCAS).

[3] Xu, C., Namunu, M., Xi, S., Fang, C., & Qi, T. (2003). A comparative study of musical genre classification using support vector machines. In Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC) (Vol. 1, pp. 131-135). https://doi.org/10.1109/ITCC.2003.1197574.

[4] Namunu, M., Xu, C., Xi, S., Fang, C., & Qi, T. (2003). Musical genre classification based on statistical features and support vector machines. In Proceedings of the IEEE Pacific Rim Conference on Multimedia (PCM) (pp. 816-821). https://doi.org/10.1109/PCM.2003.1238226.

[5] Fang, C., Xu, C., Namunu, M., Xi, S., & Qi, T. (2003). Musical genre classification using support vector machines and feature selection. In Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS) (Vol. 3, pp. III-705). https://doi.org/10.1109/ISCAS.2003.1205933.

[6] Li, T., Antoni, C., & Andy, C. (2011). Automatic Feature Learning for Musical Pattern Recognition using Convolutional Neural Networks. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 5732-5735). https://doi.org/10.1109/ICASSP.2011.5947764

[7] Antoni, C., Li, T., & Andy, C. (2010). Convolutional Neural Network-based Automatic Musical Pattern Recognition. In Proceedings of the International Conference on Pattern Recognition and Machine Intelligence (PReMI) (pp. 498-503). https://doi.org/10.1007/978-3-642-17349-7_60.

[8] Andy, C., Li, T., & Antoni, C. (2010). Musical Pattern Recognition using Convolutional Neural Networks: A Comparative Study. In Proceedings of the International Conference on Computational Intelligence and Multimedia Applications (ICCIMA) (pp. 49-54). https://doi.org/10.1109/ICCIMA.2010.47.

[9] AzhdariSeyed Majid Hasani et al. Pulse repetition interval modulation recognition using deep CNN evolved by extreme learning machines and IP-based BBO algorithm Eng. Appl. Artif. Intell. (2023).

[10] HaseliGholamreza HECON: Weight assessment of the product loyalty criteria considering the customer decision's halo effect using the convolutional neural networks Inform. Sci. (2023).

[11] ShenBaohua et al. Evolving marine predators algorithm by dynamic foraging strategy for real-world engineering optimization problems Eng. Appl. Artif. Intell. (2023).

[12] AbdillahJ. et al. Emotion classification of song lyrics using bidirectional LSTM method with glove word representation weighting J. RESTI (2020).

[13] AgrawalYudhik et al. Transformer-based approach towards music emotion recognition from lyrics Eur. Conf. Inf. Retr. (2021).