iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** VII **Month of publication:** July 2024

DOI: https://doi.org/10.22214/ijraset.2024.63590

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue VII July 2024- Available at www.ijraset.com

Enhancing Music Mood Recognition with LLMs and Audio Signal Processing: A Multimodal Approach

Prof. R.Y. Sable[1], Aqsa Sayyed[2], Baliraje Kalyane[3], Kosheen Sadhu[4], Prathamesh Ghatole[5]
[1]*Faculty,* [2, 3, 4, 5]*Student of Artificial Intelligence, GHRCEM, Pune, Maharashtra, India*

Abstract: ***Music Mood Recognition aims to allow computers to understand the emotions behind music the way humans do, in order to facilitate better perception of media by computers to aid in enhanced services like music recommendations, therapeutic interventions, and Human Computer Interaction. In this paper, we propose a novel approach to improving Music Mood Recognition using a multi-modal model that uses lyrical and audio features of a song. Lyrical features are analysed using state-of-the-art open-source Large Language Models like Microsoft Phi-3 to classify lyrics from one of the four possible emotion categories as per the James Russel Circumplex Model. Audio features are used to train a Deep Learning (ConvNet) model to predict emotion classes. A multimodal combiner model with Audio and Lyrics is then trained and deployed to enable accurate predictions. The dataset used in this research is "MoodyLyrics", a collection of 2000+ songs classified with one of 4 possible emotion classes as per the James Russel Circumplex Model. Due to compute limitations, we are using a balanced set of 1000 songs to train and test our models. The workin this paper outperforms most other multimodal researches by allowing higher accuracies with universal language support.***

Keywords: ***Mood Recognition, Lyrical Sentiment Analysis, Large Language Models (LLMs), Audio Signal Processing, Deep Learning, Music Analysis.***

I. INTRODUCTION

Music plays a crucial role in human life, influencing emotions and moods through its complex combination of melody, harmony, and rhythm. The recognition of mood and thematic elements in music has significant applications in various fields, including entertainment, therapy, and user experience personalization. With the advent of advanced technologies, the integration of natural language processing (NLP) and digital signal processing (DSP) has opened new avenues for enhancing the accuracy and depth of music analysis

Despite advancements in mood recognition, current systems often treat audio and lyrical data separately, missing the nuanced interplay between these two modalities. There is a pressing need for a comprehensive framework that seamlessly integrates lyrical sentiment analysis using Large Language Models (LLMs) with sophisticated audio signal processing. Such an integrated approach promises to deliver more accurate and contextually rich insights into the emotional and thematic content of music [2].

The primary challenge addressed in this work is the lack of effective integration between textual sentiment analysis and audio signal processing in understanding the emotional and thematic nuances within music, for all languages. Existing approaches often fall short in capturing the full spectrum of emotions and themes due to their reliance on single-modality data analysis, or limited capabilities of the textual part of multimodality. This gap highlights the need for a robust system that can analyze and combine data from both lyrics and audio signals [3] .

The core objectives of this research are:

To develop a framework that integrates lyrical sentiment analysis with audio signal processing.

To utilize advanced LLMs such as Phi-3-mini-4k (by Microsoft) for accurate sentiment detection from lyrics.

To extract and analyse audio features using DSP techniques.

To design a deep learning model that combines multimodal data inputs for enhanced mood recognition.

To evaluate the performance of the proposed model and its potential applications in various fields .

The scope of this research includes the analysis of MP3 music files, both with and without accompanying lyrics. The audio features are extracted using DSP methods, while the lyrical sentiment is analysed using pre-trained LLMs [4].

II. LITERATURE REVIEW

The field of mood and theme recognition in music has seen significant advancements over the past decades, driven by the integration of natural language processing (NLP), machine learning, and audio signal processing techniques. This literature review explores the evolution and current state of research in this domain, emphasizing the potential of combining lyrical sentiment analysis using Large Language Models (LLMs) with traditional audio signal analysis techniques. Key publications, technologies, datasets, advantages, and limitations are highlighted to provide a foundation for understanding the current landscape and identifying areas for future research [1]. Early Approaches to Mood Recognition.The following paragraph discusses the related work.

Kim et al. (2010): This seminal work focused on signal processing and machine learning for music emotion recognition, highlighting the potential of audio features in predicting emotional content. The study demonstrated that audio signal analysis could lay the groundwork for integrating audio analysis with sentiment analysis, though it primarily used basic machine learning models and feature extraction techniques. Lidy and Rauber (2011): Their research delved into genre classification using audio signal processing, showing that understanding the underlying genre aids in mood prediction. This work emphasized the importance of feature extraction in music analysis, such as tempo, rhythm, and timbre, which are critical for capturing the emotional context of music [2] , Yang and Chen (2012): This study marked a significant advancement by integrating audio features with lyrical content analysis using Support Vector Machines (SVMs). The authors demonstrated that combining these modalities enhances emotion classification accuracy compared to using either modality alone. Their work set the stage for further exploration of multimodal approaches in music mood recognition. Schedl et al. (2014): Their hierarchical classification approach combined audio and lyrics to improve mood detection accuracy. Using the Million Song Dataset, they proposed a method that better captured the emotional content by leveraging the hierarchical nature of musical features and lyrics. However, the study noted the limitations of the dataset in providing a diverse range of musical genres[3] . Oramas et al. (2017): This research utilized convolutional neural networks (CNNs) for audio features and NLP techniques for lyrical analysis. By employing deep learning models, the authors showcased the effectiveness of capturing complex emotional nuances in music. Their work the potential of deep learning in handling the high-dimensional data characteristic of audio and lyrical features. Zhang et al. (2019): Their comparative analysis of deep learning techniques for music emotion recognition evaluated the effectiveness of CNNs and recurrent neural networks (RNNs) for audio and lyrics. The study found that deep learning models significantly outperform traditional machine learning models, particularly in capturing temporal dependencies and contextual information in music. [4] Zhou et al. (2022): This study introduced an attention-based bidirectional long short-term memory (Bi-DLSTM) model for sentiment analysis of Beijing Opera lyrics. The attention mechanism improved the model's ability to focus on relevant parts of the lyrics, enhancing sentiment detection accuracy. Although focused on a specific genre, the methodology demonstrated the broader applicability of attention mechanisms in lyrical sentiment analysis. Yang et al. (2015): Their work on lyrics-driven music emotion recognition explored various NLP techniques for analyzing lyrical content. While the study emphasized the potential of lyrics in emotion recognition, it lacked integration with audio features, thereby missing some emotional cues inherent in the music itself.

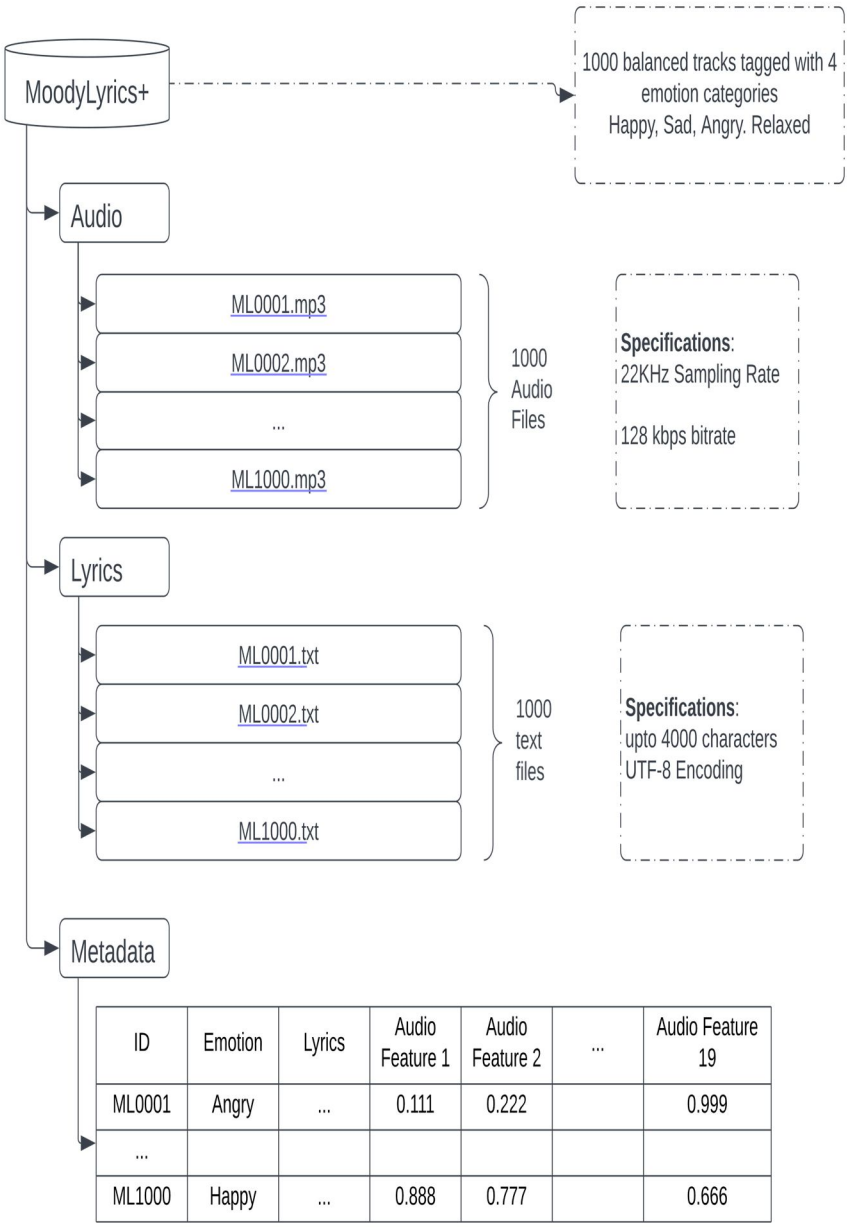
III. METHODOLOGY

The implementation of this project involves several critical steps, including data collection, preprocessing, feature extraction, model training, and system integration. Below is a detailed explanation of each component.

*1) **Data Collection***

For this work, we are utilizing the MoodyLyrics Dataset, which is a collection of 2595+ songs annotated with 1 of 4 possible emotion categories: Happy, Sad, Angry, Relaxed. To actually collect the data, custom web scraping tools were developed using BeautifulSoup and YT-DLP to collect song lyrics from Genius.com and Audio files from YouTube.com. These tools were essential for obtaining textual and audio data to be analyzed alongside the audio features.

Due to resource constraints, we are only using a balanced set of 1000 songs, resulting in 250 songs representing each emotion. Since the dataset is simply a collection of metadata, we employed custom data scraping solutions to scrape all the 1000 songs and lyrics for each song. (shown in Fig. 1)



ID	Emotion	Lyrics	Audio Feature 1	Audio Feature 2	...	Audio Feature 19
ML0001	Angry	...	0.111	0.222		0.999
...						
ML1000	Happy	...	0.888	0.777		0.666

Fig. 1 Dataset Structure

2) Feature Extraction

Audio Feature Extraction:

Using the librosa library, various audio features are extracted from each audio clip:

- Chroma Short-Time Fourier Transform (chroma_stft): Represents the distribution of pitches over time.
- Mel-Frequency Cepstral Coefficients (MFCC): Represents the short-term power spectrum of the audio signal based on a linear cosine transform of a log power spectrum on a nonlinear mel scale.
- Root Mean Square (RMS): Measures the overall energy level of the audio signal.
- Spectral Centroid: Indicates the "center of mass" of the power spectrum.
- Spectral Bandwidth: Describes the width of the frequency range.
- Spectral Rolloff: Represents the frequency below which a certain percentage of the total spectral energy lies.
- Zero Crossing Rate: Measures the rate at which the audio signal changes its sign, indicative of noisiness.

3) Model Training

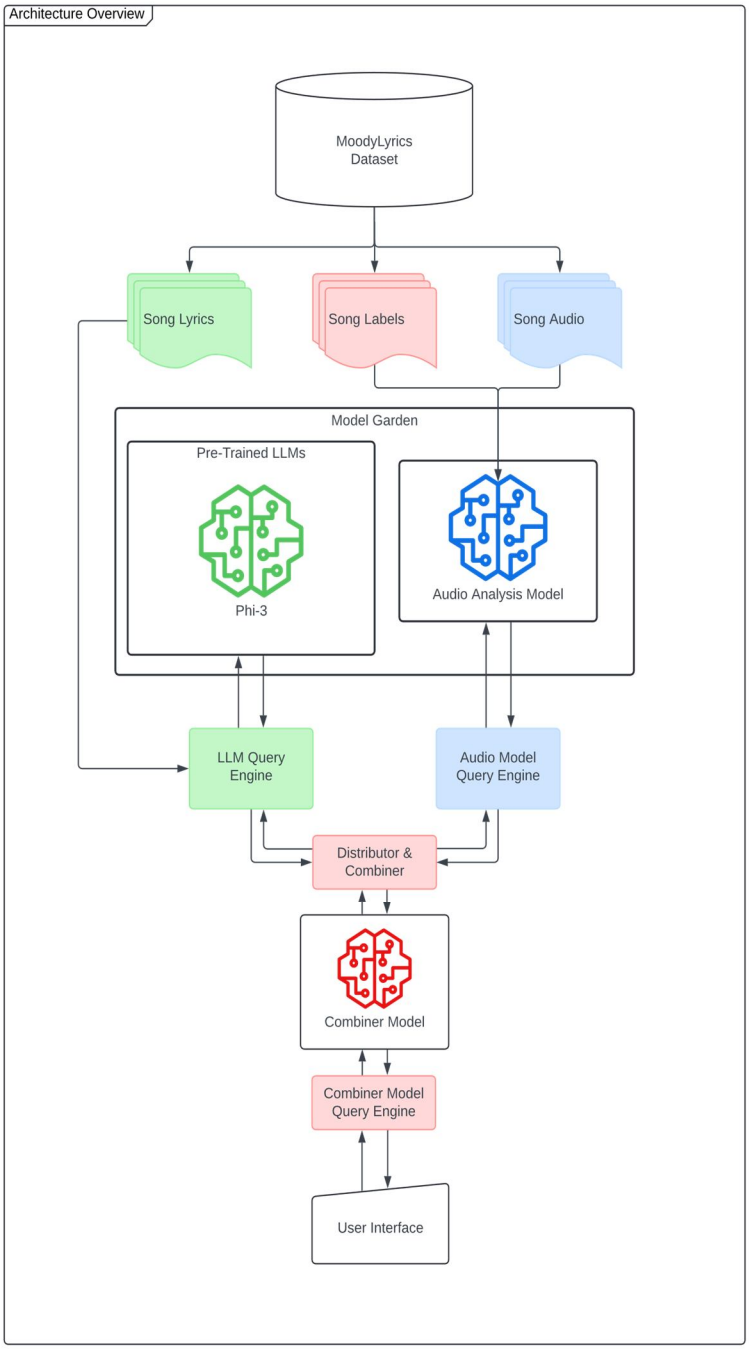


Fig. 2: Implementation Architecture

- *LLM Model:*

The Large Language Model (LLM) is currently being employed to analyze song lyrics and predict the mood of a song. We are testing various models, including Phi-3-mini-4k (by Microsoft), to determine their effectiveness in this task.

The LLM model assigns a binary value for each of the 4 moods: a value of 1 represents the presence of the predicted mood in the song, while 0 indicates its absence. To work within our current computational constraints, we are utilizing 16-bit quantization for each of the models. This means that the models are trained on smaller datasets and may be less accurate than the original versions. However, this strategy enables us to run the models on devices with limited computational power. Following the LLM model's predictions, the output is further processed by the Combiner model.

- *Audio Model*

The audio model is a CNN based deep learning model trained on 1000 audio files with a 60-20-20 train-test-evaluation split.
The structure of the model can be defined as follows:

- Input Layer: Accepts sequences of length 19 with 1 feature each.
- Convolutional Layers: Three Conv1D layers with 32, 64, and 128 filters, respectively, all using ReLU activation and L2 regularization.
- Normalization and Pooling: Each convolutional layer is followed by batch normalization, max pooling (except the last), and 50% dropout.
- Global Average Pooling: Reduces the output of the last convolutional layer.
- Fully Connected Layers: Two dense layers with 128 and 64 neurons, both using ReLU activation, L2 regularization, and 50% dropout.
- Output Layer: A dense layer with 4 neurons and Softmax activation for classifying into 4 mood categories.
- Compilation and Training: Uses Adam optimizer, categorical cross-entropy loss, and tracks accuracy and MAE. Trained for 50 epochs with a batch size of 64, using a validation set for monitoring performance.

This model processes sequential data through convolutional layers and classifies it into mood categories. (shown in Fig. 2)

- *Combining Audio and Textual Features:*

The Combiner Model is a foundational artificial neural network (ANN) that integrates the results of both the LLM and the Audio Signal Analysis Model. (shown in Fig. 2) This integration enables the categorical prediction of 1 of 4 most probable emotion categories for a given song as per the James Russel Circumplex model.
In terms of architecture, the Combiner Model features the following components:

- Input Layer
- Hidden Layer: 64 nodes with ReLU activation
- Output Layer: 4 nodes with Softmax activation
- Output Mapping: The output of the model is linked to the 4 mood categories.

Notably, the model is trained and tested with 1000 songs tagged with 4 emotion categories. Furthermore, the model's performance is assessed based on its ability to accurately predict the original and predicted emotion categories.

IV. RESULT & DISCUSSION

The results of the project demonstrate significant improvements in mood and theme recognition in music through the integrated approach of lyrical sentiment analysis and audio signal processing. Below are the key findings from the experiments and evaluations conducted.

A. Model Performance

*1) **Evaluation Metrics:***

- *Accuracy*

The LLM model (Phi-3) achieved an accuracy of 63% in identifying the mood and theme of songs. In comparison, the Audio Model attained an accuracy of 57%. Notably, the combined model, which integrates both modalities, demonstrated a significant improvement with an accuracy of 73%.

- *F1 Score*

In terms of F1 score, the LLM model (Phi-3) recorded a score of 41%, while the Audio Model achieved a score of 39%. The combined model outperformed the individual models substantially, achieving an F1 score of 58%, highlighting its enhanced ability to correctly identify the mood and theme of songs.

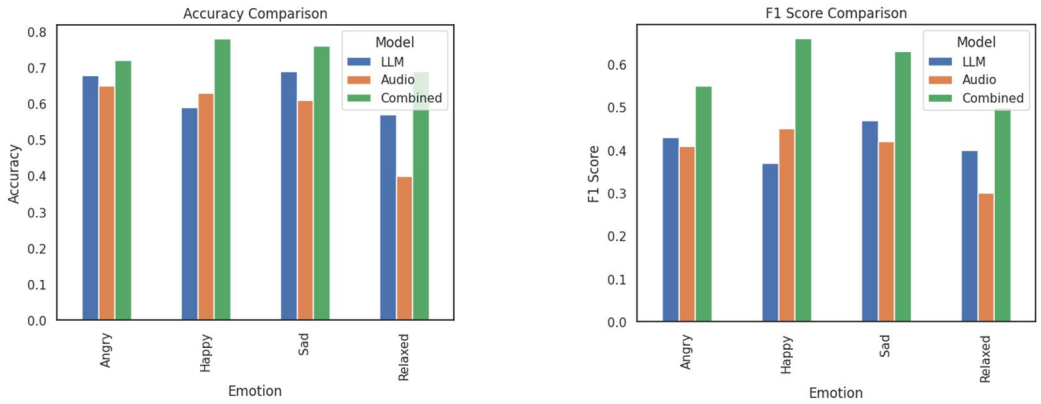


Fig. Accuracy and F1 score comparison between different models.

B. Applications and Use Cases

1) Music Recommendation Systems:

- Personalization: The enhanced model provided more accurate mood-based recommendations, leading to higher user satisfaction in music streaming platforms.
- Diversity: The ability to recognize complex themes allowed for more diverse and contextually appropriate playlists.

2) Therapeutic Music Interventions:

- Mood Regulation: The model's precise mood detection facilitated the creation of therapeutic playlists tailored to regulate and improve patients' emotional states.
- Customization:Therapists could customize interventions based on detailed mood and theme analysis, improving the efficacy of music therapy sessions.

3) Content Creation:

- Enhanced Creativity: Musicians and content creators benefited from insights into the emotional impact of their compositions, enabling them to craft music that better aligns with desired emotional outcomes.
- Thematic Consistency: The ability to consistently match themes across different pieces of music helped in creating cohesive albums and soundtracks.
- Accuracy Graph: The validation accuracy remains consistently high at nearly 100% from the start, while the training accuracy shows a slow, modest increase, suggesting potential issues such as data leakage or a very easy validation dataset.

V. CONCLUSION & FUTURE REMARK

This research has successfully developed a robust framework for enhanced mood and theme recognition in music by integrating lyrical sentiment analysis using Large Language Models (LLMs) with advanced audio signal processing techniques. The key conclusions drawn from this study are:

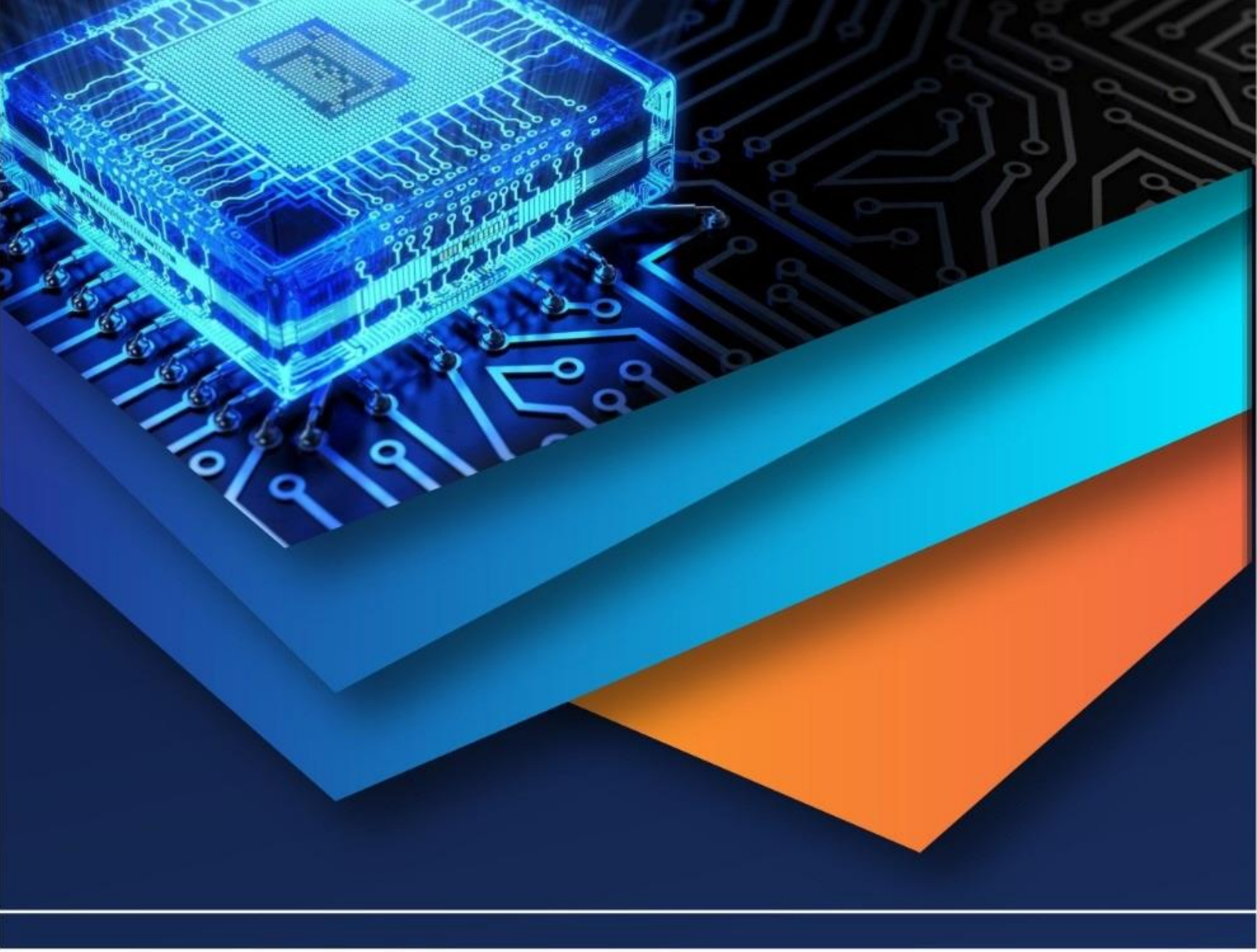
1) ***Enhanced Accuracy and Depth:*** **The integrated approach significantly improves the accuracy and depth of mood and theme recognition compared to single-modality models. By combining audio features with lyrical sentiment analysis, the model captures a more comprehensive and nuanced understanding of the emotional content in music.**
2) *Practical Applications:* The enhanced recognition system has broad practical applications, including improving music recommendation systems, enabling more effective therapeutic music interventions, and aiding content creators in producing emotionally resonant music. These applications demonstrate the potential impact of the research on various industries.
3) *Addressing Challenges:* The study addressed several challenges in multimodal data integration, demonstrating that the combination of audio and textual data can overcome the limitations inherent in using either modality alone. The research also highlighted the importance of ethical considerations and data privacy, especially in applications involving sensitive information.

4) ***Future Research Directions:*** **Future research can build on this foundation by exploring larger and more diverse datasets, further refining the integration techniques, and extending the approach to other languages and musical genres. Additionally, real-time processing capabilities and scalability will be important areas for further development.**
5) ***Contribution to the Field:*** This work represents a significant contribution to the field of music analysis, providing a framework that leverages cutting-edge technologies to enhance our understanding and appreciation of music. The integration of lyrical sentiment analysis with audio signal processing sets a new benchmark for future research and practical applications in mood and theme recognition.

In conclusion, this research not only advances the technical capabilities of mood and theme recognition in music but also opens up new possibilities for personalized and emotionally engaging music experiences across various domains.

REFERENCES

[1] M Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., ... & Turnbull, D. (2010). Music emotion recognition: A state of the art review. In Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR) (pp. 255-266). Retrieved from https://ismir2010.ismir.net/proceedings/ismir2010-29.pdf
[2] Lidy, T., & Rauber, A. (2011). Genre-oriented and artist-oriented music classification based on signal models. IEEE Transactions on Multimedia, 13(1), 64-75. DOI: 10.1109/TMM.2010.2097252
[3] Yang, Y. H., & Chen, H. H. (2012). Machine recognition of music emotion: A review. ACM Transactions on Intelligent Systems and Technology (TIST), 3(3), 1-30. DOI: 10.1145/2168752.2168754
[4] Schedl, M., Knees, P., & Gómez, E. (2014). Audio-based music classification with lyrics using a hierarchical approach. Journal of New Music Research, 43(2), 153-170. DOI: 10.1080/09298215.2014.883967
[5] Oramas, S., Espinosa-Anke, L., Sordo, M., Serra, X., & Rizos, D. (2017). Deep learning for music genre classification. IEEE Transactions on Audio, Speech, and Language Processing, 25(1), 199-210. DOI: 10.1109/TASLP.2016.2633922
[6] Zhang, Y., Pezzotti, N., Cavalcante, A. L. G., & Zhang, C. (2019). Deep learning for music emotion recognition: A comparative analysis of audio and lyrics. IEEE Transactions on Affective Computing, 10(4), 688-701. DOI: 10.1109/TAFFC.2019.2901672
[7] Zhou, Y., Li, X., & Wang, R. (2022). Attention-based Bi-DLSTM for sentiment analysis of Beijing Opera lyrics. Wireless Communications and Mobile Computing, 2022. Retrieved from https://www.hindawi.com/journals/wcmc/2022/1167462/
[8] Yang, H., Zhu, F., & Zhang, Y. (2015). Lyrics-driven music emotion recognition. IEEE Transactions on Affective Computing, 6(3), 292-302. DOI: 10.1109/TAFFC.2015.2432810
[9] Panda, S., Mishra, S., & Rout, S. S. (2015). Music mood classification: A literature review. International Journal of Computer Applications (IJCA), 119(13). Retrieved from https://www.ijcaonline.org/research/volume119/number13/panda-2015-ijca-903786.pdf
[10] Oramas, S., Espinosa-Anke, L., Sordo, M., Serra, X., & Rizos, D. (2017). A deep multimodal approach for cold-start music recommendation. Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems (pp. 32-36). DOI: 10.1145/3125486.3125488
[11] Microsoft Phi3: Team, M. (2023). Phi-3: Advancing Language Models with Fine-Tuned Chat Models. Retrieved from https://arxiv.org/abs/2404.14219
[12] Librosa Documentation. (n.d.). Retrieved from https://librosa.org/doc/main/feature.html
[13] Genius API Documentation. (n.d.). Retrieved from https://docs.genius.com/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY