International Journal of Interdisciplinary Innovative Research & Development (IJIIRD)
ISSN: 2456-236X
Vol. 11 Special Issue 03 | 2026

Music Genre Classification and Music Emotion Recognition using Multimodal Learning

Abhilasha Ishware[1], Jidnyasa Naik[2], Gauri Kadam[3], Prof. Mandar Joshi[4]
Department of Information Technology
Finolex Academy of Management and Technology, Ratnagiri Maharashtra, India

DOI: 10.5281/zenodo.20631155

ABSTRACT

Music has always been close to people's emotions as well as their emotional state, and music also influences the mood and well-being of an individual. There are many digital music platforms, and many are growing as new songs are emerging, which leads to demand for systems that can categorize, recommend, and fetch based on an individual's mood or favorite genre . The paper we propose contains a multimodal approach that is used to classify the genre of a particular music and recognize its emotion . This is done by combining the features extracted from the audio and lyrics of the music, and then those features are given input to the deep learning model to find both spectral and temporal patterns from the music's signal. The LSTM (BiLSTM) learning model of RNN is used in the model, as it can be used to analyze connected data. Lyrics are processed using BERT, which converts the raw lyrics into 768-dimensional semantic embeddings. In short, the text is converted into numerical representation. These features are combined with audio features using late fusion to create a single unified vector representation [3]. This vector is used as an input to a model that will predict the music genre and emotion based on arousal and valence. The results indicate that the combination of two inputs improved the performance. This multimodal system can be used for various purposes like music recommendation and emotion-aware music search.

Keywords—Music Genre Classification, Music Emotion Recognition, Multimodal Learning, Audio Feature Extraction, Lyrics Analysis, BiLSTM, BERT, Valence-Arousal Model, Deep Learning, Music Recommendation System

1. INTRODUCTION

The rapid growth of digital music and streaming services has led to an increase in music creation [5]. This has created a demand for tools that can classify, organize, and recommend music. Since the 1990s, much research has focused on genre classification [1]. As a result, users can easily search for music in various genres, helping them find songs that match their taste. However, manual tagging is subjective, time-consuming, and difficult to scale for large music collections.Research in music processing expanded in the early 2000s. Music Emotion Recognition (MER) is used to identify theemotional qualities of music and songs [2]. It deals with positive emotions like happiness and sadness, as well as more neutral ones like calmness and energy. Recognizing musical emotions is valuable for applications in mood-based playlists, personalized recommendations, and even music therapy and mental health[6]–[9]. Many existing systems rely on a single method, which may lead to an incomplete understanding of the music.To address this limitation, this study proposes a multimodal learning approach that integrates audio and textual information to facilitate a deeper understanding [3]. Features such as Mel-spectrogram, tempo, chroma, and MFCCs are extracted and processed using a CNN and BiLSTM architecture with an attention mechanism. At the same time, BERT is used to tokenize song lyrics and generate text embeddings to extract contextual features. By integrating multiple types of information, the proposed system improves the accuracy and reliability of music genre classification and emotion recognition compared to traditional single-method approaches [1]–[4]. This also supports more effective music recommendations and applications that consider emotional aspects [7]–[10].

2. MOTIVATION AND SIGNIFICANCE OF MULTIMODAL LEARNING

Comprehending a piece of music depends on many factors (e.g. rhythm, melody, harmony, lyrics, and cultures). When we listen to music, humans naturally combine audio perception with lyrical meaning to interpret musical genres and emotional expression. Nonetheless, in many cases, the music analysis systems are fully automated and work on a single modality. Music genres that have similar audio features, including tempo, pitch, and timbre, are hard to identify, such as pop mimicking classical and R&B. Some setups just handle sound, paying attention to volume, pitch, or timing. On the flip side, others dig into words alone - ignoring beat, tune, or how

instruments play together. Because of that narrow view, single-source approaches often misread feelings in music, particularly when emotion runs deep or shifts quietly. Merging layers - sound plus text - lets systems notice more, drawing from both what's sung and how it's delivered. Audio data gets analyzed by a system built on CNN and BiLSTM methods, along with an attention layer that picks up timing and sound patterns. Lyric content goes through language processing tools such as word vectors and a BiLSTM structure to uncover meaning in context. Instead of treating them separately, combining these two streams improves results - especially for sorting songs into genres or detecting emotions they convey. Because it handles voice tone and words together, this method fits well in apps tuned to sense feelings from music's full expression.

3. PROJECT OVERVIEW AND BACKGROUND

With the tremendous growth of digital music platforms and streaming services, the need for systems that automatically organize and recommend music based on music genre and emotional mood has increased significantly. The system can analyze the song utilizing two different sources: the audio features of it and the corresponding lyrics.The audio features specialize in the treatment of sound signals as well as the musical properties of songs while the lyrics are the text that provide insight into the emotional meaning of song. The system classifies songs in two ways, genre classification which organises music according to style. For example, pop, jazz, hip-hop categorisation makes it easier for users to explore different type of music. With the aid of Itunes, we will . Emotion recognition is identifying the mood of a song like happy, sad, energetic, etc which can help in mood-based music recommendations and personalized experience.As digital music platforms and streaming services become more and more ubiquitous, there exists an ever-increasing demand for systems that can automatically classify and understand music. The classification of musical genres has been studied since the 1990s to allow users to explore songs according to their musical style and music characteristics. In the early 2000s, researchers proposed a research field called music emotion recognition (MER), whose objective is to identify the emotional tonality of a music piece, such as happy, sad, calm, etc. Most earlier systems, however, rely on unimodal approaches that either analyze audio signals or song lyrics only, which limits their ability to elucidate the meaning and emotion of music. To overcome this limitation, this project incorporates a learning approach that integrates audio features and lyrics so that the system can attain a richer and more human-like understanding of music.

4. OBJECTIVES AND SCOPE

⇒ Objectives:
- Extract meaningful features from both audio signals and lyrics.
- Classify music into genres (e.g., pop, rock, and jazz).
- Recognize emotions using valence-arousal or mood cate-gories (e.g.,happy, sad)
- The accuracy of the unimodal systems was improved by combining multiple data sources. It enables personalized music recommendations.

⇒ Scope:
- The system can automatically classify music into pre-defined genres, reducing the need for manual tagging and helping maintain better consistency in large music databases.
- It can accurately recognize the emotional state expressed in music, such as happy, sad, calm, or energetic, which can support mood-based music recommendation systems.
- By using a multimodal learning approach that combines audio features and lyrics (text), the system improves classification accuracy compared to traditional methods that rely on only one type of data.
- The model can also be extended for real-time music analysis on streaming platforms, helping generate playlists and providing more personalized listening experiences for users.
- In addition, this approach creates a strong foundation for future improvements, such as analyzing lyrics in multiple languages, recognizing a wider range of emotions, and in-tegrating the system with commercial music applications.

5. LITERATURE REVIEW

The literature review discusses the progress in the line of music genre classification and emotion identification using machine learning and deep learning. But they have limitations. Dwivedi and Islam (2024) proposed a framework using a GAN along with Fourier and wavelet transforms for genre classification [1]. Their framework learns features that are robust. However, this method doesn't scale because it requires large training datasets and generalizes poorly across different music collections.A comprehensive survey of methods, datasets and challenges related to music emotion recognition was given by Patil et al. 2025 [2]. The study surveys the use of machine learning, deep learning and ensemble methods to perform music emotion recognition.

Although the review provides a thorough overview of the MER ecosystem, it focuses on open issues like the scarcity of standardized datasets and lack of usable real-time emotion recognition systems. Yang et al. (2023) proposed COSMIC, a hierarchical multimodal model that fuses music structure with audio–lyrics interaction for emotion recognition [3]. According to the study, multi-modal fusion enhances emotion prediction accuracy but it limits themselves only to emotion recognition. They evaluate their work upon a narrow data set which pertains only to Chinese pop music which limits applicability. Hu and colleagues, in 2025, addressed the problem of datasets that hindered research into new algorithms for music emotion classification [4]. The authors constructed a large-scale multilingual music emotion dataset with valence–arousal annotations. Even though they can benchmark MER methods using this dataset, they do not consider the detailed alignment of lyrics and audio. The reviewed studies indicate there is a research gap for a generalized multimodal system capable to classify the genre of music and recognize the emotion simultaneously with a high degree of accuracy, real-time performance and cross-dataset robustness.

6. PROBLEM STATEMENT

This project will improve both the performance of music genre classification and emotion recognition and reduce the dependency on manually annotated (subjective) data. In doing so, this project aims to create a completely automated system based upon machine learning models that analyze the features in audio data to improve our understanding of what makes up the musical properties.A larger benefit is to be able to automate organizing very large music libraries more quickly and to be able to provide users with personalized music recommendations.

7. CHALLENGES IN MUSIC GENRE AND EMOTION CLASSIFICATION

The categorization of music types and feelings poses many difficulties. Music styles blending together creates difficulty separating sounds, especially if rhythms and chords feel alike across types. Still, what defines a genre isn't fixed - it shifts with era and location, adding confusion. Feelings heard in music aren't clear cut either each person feels something unique. A single track might spark joy in one listener, sadness in another, shaped by personal setting. Though instruments may lack emotional clarity, words could twist meaning through irony or hidden symbols. Not every dataset matches how it gets labeled - forms differ. Because crowd opinions define emotions, valence and arousal shift unpredictably. To cut through confusion, systems blend inputs from separate sources in layered ways.

8. PROPOSED SYSTEM

1. System Architecture: The system processes both audio and lyrics to better understand a song. Audio features such as MFCCs and Mel-spectrograms are extracted, while the lyrics are converted into embeddings through text processing. These features are combined using a multimodal model based on LSTM or Transformer networks. The final module predicts the song's genre and emotion and displays the results to the user.

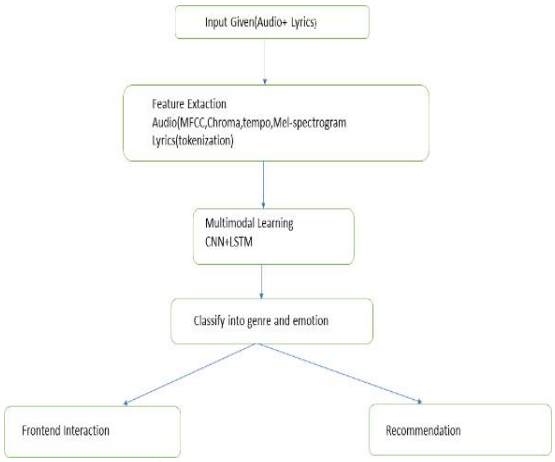

Fig. 1. System Architecture for Multimodal Music Genre and Emotion Classification

2. Detailed System Architecture Explanation The system uses a modular multimodal architecture to process audio signals and song lyrics individually before merging them for the final prediction. The model can learn the required modality-specific features with this design while maintaining audio and text information. As it passes through the audio processing pipeline, a raw music signal becomes a mel-spectrogram that captures time–frequency characteristics (e.g., pitch, rhythm, and spectral patterns). Through a convolutional neural network

(CNN) and then a bidirectional long short-term memory (BiLSTM), the authors process the spectrograms for acoustic feature extraction and temporal modeling of music signals. Afterwards, attention was applied to highlight the most informative time frames. In the lyrics processing pipeline, the song lyrics were tokenized and transformed into embedding vectors that represented the semantic relations. In order to capture the contextual meanings and emotional expressions, a bidirectional LSTM encoder processes these embeddings. The representations from both encoders were fused using a late fusion technique by concatenating audio and lyric features to get a single multimodal representation. This unified portrayal allowed for both genre recognition and understanding of musical emotion. Moreover, the modular design is optimally employed not only to accommodate future extensions but also, for example, to add additional modalities and more elaborate encoder models.

3. Data Flow: Input → Preprocessing → Feature Extraction
→ Fusion → Classification → Output

⇒ Input: Audio (.mp3) + Lyrics (.txt)

⇒ Processing: Audio → Mel-Spectrogram → audio Encoder

- Lyrics → Token Embeddings → LSTM Encoder

⇒ Fusion: Combine both latent features

⇒ Output:

- Genre Label (Categorical)
- Emotion Metrics (Valence and Arousal or Emotion Cate-gory)

2. Dataset Description:

⇒ Audio datasets used:

For the dataset in this project, we used a custom multimodal dataset with audio files, song lyrics, and metadata annotations. This software is meant to help in music-genre classification and emotion recognition task. The dataset was structured by means of a CSV metadata file. Each song has a song_id which corresponds with the audio file as well as the lyrics text. By using this systematic design, accuracy in matching audio and textual information for the audio text retrieval task is ensured. Music tracks in the form of audio in MP3 files covering multiple genres like pop, rock, Jazz etc. The Librosa library processed each audio file where the raw audio signals were loaded and converted to mel-spectrogram representations. These spectrograms of music capture important acoustic features, such as timbre, pitch and rhythm which enable one to identify musical styles and detect the emotional characteristics of songs.

⇒ Lyrics/text data sources:

The lyrics and text data for each song are stored in distinct text files. A character-level tokenizer was used to preprocess the
lyrics so that the model could learn the semantic and emotional cues present in the song lyrics. This textual modality enhances audio features by offering contextual and emotional meaning that sound alone might not be able to fully convey.

⇒ Emotion and genre labels:

Genre and emotion labels were attached to each song in the dataset. While emotion labels are provided using the valence–arousal model, where valence denotes emotional positivity and arousal denotes emotional intensity, genre labels represent categorical music classes (e.g., pop, rock, and jazz). Together with the song_id, these labels are stored in the CSV file, allowing supervised learning for both genre classification and emotion recognition in a single multimodal framework.

3. Feature Extraction:

- Input: Takes text with lyrics (.txt) and audio files (.wav). mp3).
- Audio feature extraction: Librosa was used to transform audio into mel-spectrograms.
- Audio Encoding: Captures spectral and temporal patterns by processing spectrograms using CNN + BiLSTM with Attention.
- Lyric Processing: BiLSTM is used to process the lyrics after they are tokenized and transformed into embeddings.
- Multimodal Fusion: Late fusion is used to concatenate au-dio and lyrics features to create a cohesive representation.
- Prediction: Genre (classification) and emotion (Valence & Arousal regression) were predicted using the combined features.
- Output: Shows the anticipated genre and emotional characteristics of music.

4. Multimodal Fusion Strategy

The proposed system employs multimodal fusion which increases prediction performance. The model employs late fusion where audio and lyrics features are learnt separately and fused on the final prediction step [3]. The

model incorporates information from each modality after they have learned their properties. The audio encoder receives the music signal and generates a high-level representation of it using CNNs, BiLSTMs, and attention. Architectures of the CNN layers extract spectral patterns from the Mel-spectrograms, the BiLSTM captures temporal dependencies in the audio sequence, and the attention layer highlights the most informative segments of the signals. The lyrics encoder creates a representation of the song's lyrics. The lyrics were converted into word embeddings and processed using a Bidirectional LSTM, which captures the semantic relationships between words and identifies emotional patterns in the text. By concatenating the feature representations of the two encoders, a common multimodal embedding was devised representing the association of acoustic features and lyrical meaning [3]. After being subjected to dropout regularization, this fused representation passed through fully connected prediction layers to output final music genre, valence and arousal estimations. With this fusion method, it improves the prediction accuracy [1]–[4]. Also, it is robust and provides flexible modeling. A system can also be made more flexible with output fusion.

5. Modules:

- Data Acquisition: We retrieved audio files along with lyrics to complement the audio files for each song to help us form a paired dataset with audio and text for all songs.
- Processing: The audio data was converted into Mel-spectrograms, which captured relevant time-frequency features of the music. At the same time, the lyrics were tokenized and converted to text embeddings so that the system could evaluate the semantic and sentimental meaning of the lyrics.
- Model: A deep learning model that processes audio and lyric features. The analysis of audio features was conducted with CNN and BiLSTM with an attention mechanism, while the lyrics were processed with a BiLSTM encoder. The fusion of the two modalities was then performed to classify the song genre and emotion
- Security: To keep the system reliable and block any uploads that are malformed or harmful, submissions by users are restricted to audio only (.wav). MP3 Format It guarantees that input data will be processed safely and consistently.

6. Methodology:

A deep learning architecture based on a CNN–BiLSTM framework is a proposed solution to merge two different data types, audio signals and lyrics [3]. This learning method enables the model to acquire supplementary information from both modalities, thereby enhancing the accuracy of identifying the music genre and emotion

7. Audio Features:

The audio features were obtained through mel-spectrograms using Librosa. These are the essential time-frequency characteristics of music. The local spectral patterns were learned with the extracted features using a convolutional neural network (CNN). Further analysis of CNN outputs was performed using a Bidirectional Long Short-Term Memory (BiLSTM) network to capture the temporal relationships in the music. For indicating the parts of the audio signal that contain the most information, an attention mechanism was applied.

8. Text Features:

Textual features were collected from the song lyrics. Initially, the lyrics underwent tokenization, which was followed by embedding vector mapping. The Bidirectional LSTM encoder processed these embeddings to achieve contextual understanding and representation of the lyrics, along with identifying the emotional aspect of the lyrics.

9. Project Modules

⇒ A script called "dataset.py" handles pre-processing.

- Reads metadata from CSV files (for example, song_id, genre, valence, arousal).
- It matches the song_id in the CSV with the audio file and lyrics file.
- Librosa is used to load audio and compute mel-spectrograms.
- It loads lyric files and tokenizes them into sequences of numbers.

⇒ train.py

- Loads directories for CSV, audio, and lyrics.
- This function creates a MusicDataset object and wraps it in a PyTorch DataLoader with a collate function.
- The training loop runs for 10 epochs and computes the multitask loss for training.

⇒ model script

- The model uses a CNN + BiLSTM with attention to process Mel-spectrogram features.
- Lyrics Encoder – Embedding layer + BiLSTM for extract-ing contextual lyric features.
- Concatenating audio and lyrics representations.

⇒ Prediction Heads

- Valence Head, Arousal Head, and Genre Head

10 Technology stack:

- Frontend: React.js.
- Backend: Python, PyTorch, librosa, FastAPI.

- Libraries/Frameworks: PyTorch, Librosa, Scikit-learn
- Storage: Local files. Frontend: React.js Backend: Python, FastAPI

Libraries/Frameworks: PyTorch, Librosa, Scikit-learn
Storage: Local files

11. Assumptions/Constraints:

Dataset includes paired audio and lyrics for each track.

- Model accuracy depends on quality and balance of dataset.
- Network latency or storage constraints may limit real-time prediction for long audio files.

11. Dataset Used:

- Audio Dataset – GTZAN Genre Collection
- Lyrics Dataset – Kaggle Lyrics Emotion Dataset

9. MODEL TRAINING AND OPTIMIZATION

The suggested multimodal model is trained through super-vised multitask learning where system learns to predict the music genre and emotion at the same time [1]–[4]. While training, audio spectrograms and lyric token sequences were processed by distinct neural encoders in parallel.

The audio encoder is made of CNN Layers and a BiLSTM network with an attention mechanism [3]. The CNN layers get the valuable spectral patterns from the Mel-spectrograms. The BiLSTM gets the latent temporal relationship in the music signal. The attention layer helps the model summon at the most useful parts of the audio sequence. The lyrics encoder takes care of the textual content of the songs. Through the implementation of embedding layer, the words firstly get converted into a numerical representation followed by passing it through Bidirectional LSTM network, receiving information with consideration for both past and future data points means. The emotional detection of the lyrics takes place at this stage.

The two encoders yield representations that are fused via concatenation to form a multimodal feature representation. The dropout layer provided the combined representation in an effort to increase generalization; the resulting output was then passed to the task-specific prediction heads for genre classification and emotion prediction. Different loss functions for optimization.

- The genre classification was learned through a category. Loss function.
- The Mean Squared Error is used for valence and arousal prediction. (MSE) loss since they are the continuous emotion values.

The total training objective is the sum of all these, individual losses, enabling the model to learn all tasks jointly, within a multi-task learning framework

10. EXPERIMENTAL SETUP

This section describes the experimental setup that assesses the proposed multimodal deep learning framework that classi-fies music genres and recognizes emotions. The experimental arrangement obtained is such that one can get reliable results and reproduce them for fair evaluation of the model in audio and text.

The experiments were run on a multimodal music dataset comprising 4,492 audio tracks, each sample containing an audio file, corresponding song lyrics, emotion annotations (valence and arousal), and genre label. To prepare the dataset for genre classification, the original fine-grained genre labels were grouped together to finally form eight broad genres: pop, rock, hip-hop, R&B, electronic, jazz, classical, and country. Grouping enables us to make the classification task easier but also enables us to evaluate the data uniformly for different music styles. In this case, 90% of data has been used for training, which allows the data pattern to learn meaningfully from the dataset, whereas the rest 10% were used for testing. The training process did not utilize the test set in any way, thereby ensuring that this test set was entirely kept separate from the training set. When splitting the dataset, the stratified sampling strategy was applied [2]. This method ensures that the genre class distribution is retained for the training and testing subsets. Hence, the test set included a representative sample of all genre categories, which helped in mitigating the effect of class imbalance on the results.

The model's parameters were learned using the training set, and the test set is used only to test the model's accuracy in genre classification and capacity of emotion recognition [1]–[4].

11. RESULTS AND DISCUSSION

The MoodTune result interface displays predicted genre, valence, and arousal to indicate the song's mood and intensity.It also provides recommended songs based on similar genre and emotional features. The interface is intuitive, presents results in real time, and clearly communicates insights from com-bined audio and lyric analysis. Additionally, it enhances user experience by enabling quick and informed music exploration.

A. Test Results

- Valid test samples used: 450

- Genre Accuracy: 47.78

Fig. 2. Training log displaying dataset size, training samples, and epoch-wise loss values, with automatic saving of the best-performing model weights.

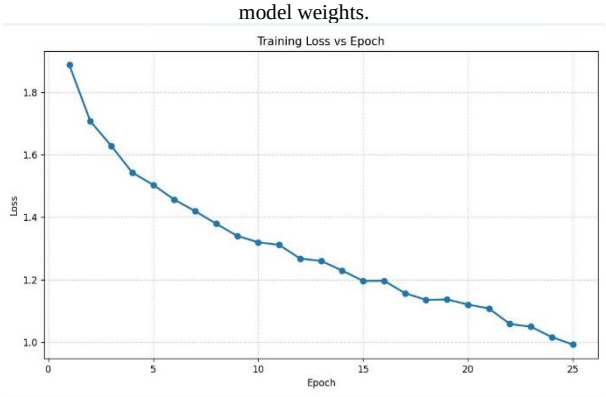

Fig. 3. Training loss curve showing decreasing loss across epochs, indicating effective learning and model convergence.

B. Classification Report

C. TABLE I CLASSIFICATION REPORT FOR GENRE PREDICTION

Genre	Precision	Recall	F1-Score	Support
Rock	0.6250	0.0943	0.1639	53
Hip-Hop	0.6889	0.6200	0.6526	50
R&B	0.2750	0.6346	0.3837	52
Electronic	0.5410	0.4648	0.5000	71
Jazz	0.4833	0.4462	0.4640	65
Classical	0.7400	0.6981	0.7184	53
Country	0.4828	0.5490	0.5138	51
Accuracy	-	-	0.4778	450
Macro-Avg.	0.5290	0.4816	0.4707	450
Weighted-Avg.	0.5273	0.4778	0.4700	450

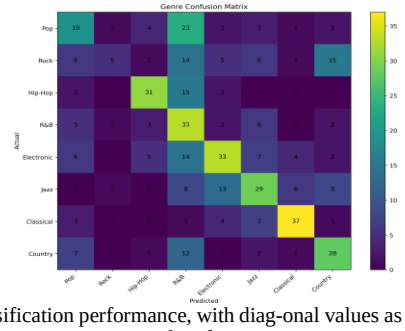

Fig. 4. Confusion matrix showing genre classification performance, with diag-onal values as correct predictions and off-diagonal values as misclassifications.

D. Emotion Prediction Error

The performance of the emotion prediction model is evalu-ated using mean absolute error (MAE):

- Valence MAE: 0.1506
- Arousal MAE: 0.1415

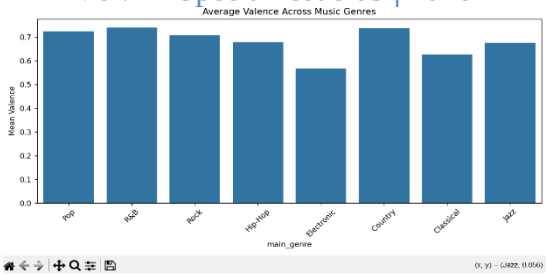

Fig. 5. Average valence scores across music genres indicating the relative positivity of songs.

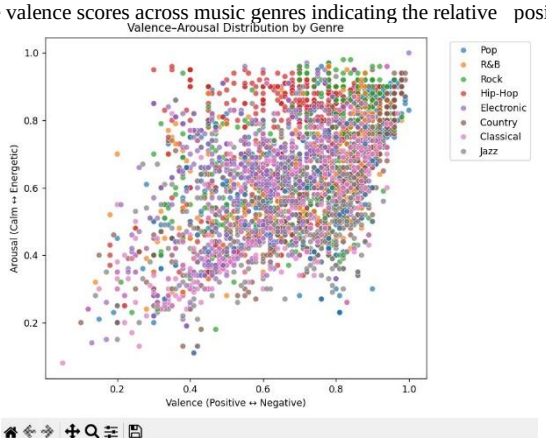

Fig. 6. Valence–arousal distribution of songs across genres, where each point represents a song's emotional characteristics.

User Interface and System Visualization

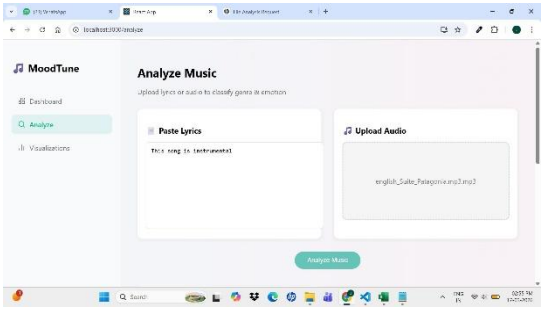

Fig. 7. MoodTune interface for audio or lyric input to classify genre and emotion, with visualization and dashboard access.

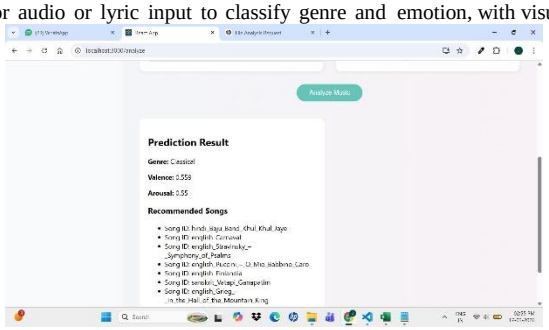

Fig. 8. MoodTune results interface showing predicted genre, valence, arousal, and recommended songs based on the input.

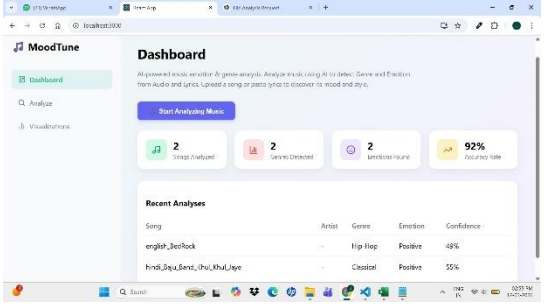

Fig. 9. The MoodTune dashboard shows an analysis overview, including songs analyzed, genres, emotions, accuracy, and recent results.

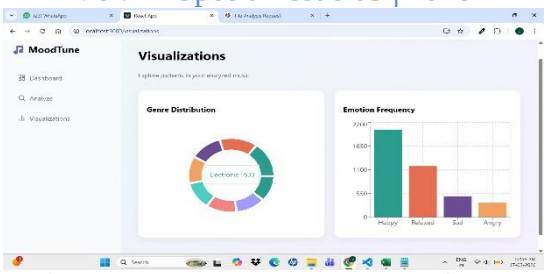

Fig. 10. MoodTune visualization interface showing genre distribution and emotion frequency through charts for pattern analysis.

Genre classification performance: The music genre classi-fication result of the proposed multimodal learning model is impressive by leveraging information captured from audio and lyrics [1]–[4]. The model's understanding of acoustic patterns such as rhythm, pitch, and timbre is aided by audio features extracted from mel-spectrograms. These features also provide semantic information based on text through lyric embeddings. The combination of text and audio features captures music as well as text characteristics of songs and improves genre prediction accuracy when compared to systems that use only audio or only text [3]. Emotion recognition performance: The valence-arousal emotion representation is used by the model to predict an audio emotion [2, 4]. The multimodal architecture learns from the audio signals and expression curves in the lyrics [3]. The model can make a more consistent and reliable prediction about emotion. This will help the system better distinguish the emotions happy, sad, calm, and energetic, which are reflected in sound and lyrics [6]–[9]. Evaluation metrics (accuracy, precision, recall, and F1-score): Commonly used classification metrics, which include accuracy, precision, recall, and F1-score, were used to evaluate the performance of the developed system (RS) [2]. To assess the overall correctness of the model's predictions, accuracy is used. When there is more than one genre or emotion category, checking precision and recall helps to gain better insights into model performance for each class. The F1 score, which balances precision and recall, is particularly useful for evaluating performance when the dataset may contain imbalanced classes [2]. Comparison with unimodal approaches: A comparison was performed between audio-only, lyrics-only, and multi-modal models. The results show that unimodal approaches have limitations in understanding the full context of music composition [2]. Audio-only models focus on acoustic properties but cannot capture the semantic mean-ing expressed in the lyrics, whereas lyrics-only models ignore important musical elements such as rhythm, tempo, and instrumentation. In contrast, the multimodal model combined both sources of information and consistently achieved better performance across all evaluation metrics [1]–[4].

Analysis of results

The experimental results demonstrate that multimodal learning significantly improves genre classification and emotion recognition [1]–[4]. By in-tegrating complementary information from audio and lyrics, the system can reduce ambiguity and produce more reliable predictions. Although the model performed well overall, some variations in performance were observed across different genres and emotion categories. These differences are mainly due to dataset imbalance and the subjective nature of emotion labeling in the datasets [2, 4]. Nevertheless, the results highlight the effectiveness of the proposed multimodal approach and support its use for intelligent music analysis and recommendation systems [5].APPLICATIONS The proposed multimodal learning system can be effectively applied to music recommendation systems to provide more personalized suggestions by accurately identifying both the genre and emotional characteristics of songs. By understanding these aspects, streaming platforms can recommend music that better matches a user's listening preferences and emotional contexts. The system also supports moodbased playlist generation by automatically grouping songs according to detected emotions, such as happy,sad,calm,or energetic. This allows users to easily select music that suits their current mood, activity, or emotional state without manual tagging.On music streaming platforms, the model can enable automatic tagging of genres and emotions, enhancing the organization of music, searching for content, and discovering it. Through motivation, the analysis of the audio and lyrics multimodally, and importantly less reliance on the metadata of things such as human-created tags of manual metadata, the system allow the platforms the effective managing and analyzing of large libraries of music.Further, the proposed method can find applications in music therapy and affective computing, wherein music selection, which is aware of emotion, can assist in emotional analysis, stress management and mental health applications.

12.HARDWARE AND SOFTWARE REQUIREMENTS

⇒ Hardware Requirements:
- Processor: Intel Core i3/i5
- RAM: Minimum 8 GB
- Minimum 50 GB (for datasets, model checkpoints, and logs)

⇒ Software Requirements:
- Operating System: Windows 11
- Python Version: 3.8 or above
- Key Libraries and Tools:PyTorch, Librosa, Scikit-learn, VS Code, Matplotlib

13. RELEVANCE OF THE PROJECT

⇒ Social Relevance
This project allows users to save time in accessing music on online music platforms by tapping the music genre and emotion automatically. As a result, users will be able to listen to music that fits their mood or wine of the day. This will help with emotional well-being. Furthermore, it allows for easier access to large collections of music without needing to manual-tag each song.
⇒ Industrial Relevance
In the music streaming and entertainment sectors, precise genre and emotion classification is vital for personalized recommendations and efficient management of audio and video content. The multimodal system proposed seeks to automate the task of tagging music in order to cut down on human effort and enhance the accuracy of recommendations with the aim of supporting a scalable deployment in real-world conditions.
⇒ Educational Relevance
The development project combines expertise in machine learning, deep learning, audio signal processing and natural language processing with the practicality of a learning platform. Students are provided hands-on experience with multimodal AI systems through this, strengthening the conceptual under-standing and research skills in music information retrieval.

14. SUSTAINABLE DEVELOPMENT GOALS (SDGS)

- SDG 3: Good Health and Well-being Proposed multi-modal music analysis system allows for emotion-based song categorization which will help in terms of mental wellbeing, mood regulation and therapeutic music.
- SDG 9: Industry, Innovation, and Infrastructure The initia-tive proposes the development of an AI-based Multimodal Learning Model taking audio and lyrics into account to foster innovation to Emotion Aware Music Systems and Intelligent Content Management in the music industry.
- SDG 4: Quality Education By fusing digital signal processing, natural language processing, and machine learning, this project offers a hands-on teaching platform that encourages AI and data-driven learning in engineering and computer science.

15. LIMITATIONS OF THE PROPOSED SYSTEM

-Less Availability of Datasets
The dataset including the audio and lyrics is not easily available to be processed. The Dataset which is available is less in size which may restrict the model's learning capability.
-Emotion Perception
The emotion of a particular song may evoke many other different emotions as per subject.
-Language Dependency
The performance of the model varies based on the song's language, as lyrics are one of the inputs. The overall working of the model can get affected and may decrease.
-High Computational Requirement
Models that are multimodal typically combine multiple different deep learning architectures. This may include convolutional neural networks, recurrent neural networks, or transformers. There is a rise in computational cost, training time, and memory usage.
-Feature Fusion
Combining the features extracted from different inputs is challenging. Both pieces of information must be taken into consideration.
-Noise in Audio
Low-quality audio, background noise, or compressed audio can affect the accuracy as well as the learning of the model and can affect the extraction of audio features like Mel-spectrograms and MFCCs.
-Overlap Between Music Genres
Numerous musical pieces exhibit traits of more than one genre. One can combine several different genres such as pop and rock
or jazz and fusion. The model is challenged to pick just one genre label since the genres overlap in the text.
-Not suitable for real-time processing
As multimodal systems need to process the audio as well as the text, they may take longer for processing. It makes it difficult for live applications like live music analysis and instant recommendations.

- Model's dependency on preprocessing

Low quality preprocessing will make the model struggle to learn patterns, including audio feature extraction and lyrics text preprocessing.

- Limited Model Interpretability

Deep learning models used in multimodal learning are often a black box. Because of this, it can be hard to specify the reasons for a song being in a particular genre or emotion.

16. CONCLUSION

We will examine the song's audio signal and the lyrics. The system that utilizes machine learning, audio signal processing and natural language processing aims to understanding various understandings of music which will enhance the organization and recommendation of songs on digital platforms. A system like this can help improve personalized music recommendations for the users and may generate mood-based playlists too. It can also help in applications like emotion-aware music retrieval music therapy and mental well-being where understanding the emotional impact of music is essential.

17. REFERENCES

[1] Pulkit Dwivedi and Benazir Islam, "Generative Adversar-ial Networks Based Framework for Music Genre Classification," Springer, October 8, 2024.

[2] Sheetal Patil, Rudragoud Patil, Shweta Goudar, S. Sangani, and R. H. Goudar, "Review on Music Emotion Analysis Using Machine Learning: Technologies, Methods, Datasets, and Challenges," Springer, January 2025. ¡en-us-grammar¿[3] Liang Yang, Zhexu Shen, Jingjie Zeng, Xi Luo, and Hongfei Lin, "COSMIC: Music Emotion Recog-nition Combining Structure Analysis and Modal Interaction," Springer, July 8, 2023.¡/en-us-grammar¿

[4] Qiong Hu, Masrah Azrifah Azmi Murad, and Qi Li, "Advancing Music Emotion Recognition: Large-Scale Dataset Construction and Evaluator Impact Analysis," Springer, January 29, 2025.

[5] M. Darvish and M. Bick, "The Role of Digital Technolo-gies in the Music Industry—A Qualitative Trend Analysis," Information Systems Management, 2023.

[6] L. Rebecchini, "Music, Mental Health, and Immunity," Brain, Behavior, and Immunity – Health, 2021.

[7] S. Bhandarkar, B. V. Salvi, and P. Shende, "Current Scenario and Potential of Music Therapy in the Management of Diseases," Behavioural Brain Research, 2024.

[8] K. Adiasto, D. G. J. Beckers, L. M. L. van Hooff, K. Roelofs, and S. A. E. Geurts, "Music Listening and Stress Recovery in Healthy Individuals: A Systematic Review with Meta-Analysis of Experimental Studies," PLoS ONE, 2022.

[9] M. de Witte, A. da Silva, G.-J. S. Pinho, X. Moonen, A. E. R. Bos, and S. van Hooren, "Music Therapy for Stress Reduction: A Systematic Review and Meta-Analysis," Health Psychology Review, 2020.

[10] A. C. Feneberg et al., "Perceptions of Stress and Mood Associated with Listening to Music in Daily Life During the COVID-19 Lockdown," JAMA Network Open, 2023.