Multimodal Music Genre Classification Using Audio Features and Lyrics

*Note: Sub-titles are not captured in Xplore and should not be used

1st Hetul Shah
Institute of technology
Nirma University
Ahmedabad, India
21bce087@nirmauni.ac.in

2nd Ridhdhi Sangani
Institute of technology
Nirma University
Ahmedabad, India
21bce257@nirmauni.ac.in

3rd Shrutam Shah
Institute of technology
Nirma University
Ahmedabad, India
21bce273@nirmauni.ac.in

***Abstract*—In order to classify music genres, this research uses a multimodal approach that makes use of both auditory characteristics and lyrics. We use sophisticated methods including lyric vectorization using Word2Vec embeddings and different machine learning models like Random Forest, Support Vector Machines (SVM), and Recurrent Neural Networks (RNN). Furthermore, we investigate how Convolutional Neural Networks (CNN) can be used to generate spectrogram visuals from audio data. We show through comprehensive testing and review that our method is effective in correctly classifying music into various genres.**

***Index Terms*—Multimodal Classification, Music Genre, Audio Features, Lyrics, Word2Vec, SVM, Random Forest, RNN, CNN, Spectrogram.**

I. Introduction

In the subject of music information retrieval, music genre classification is a crucial task with applications ranging from recommendation systems to music organisation. For the purpose of predicting genre, classification models have traditionally just used auditory data. Nonetheless, by combining various information modalities, it is possible to improve classification accuracy given the growing availability of textual data, such as song lyrics. In order to improve genre classification, we present a multimodal technique in this study that integrates audio features with lyrics.

II. Motivation

This study is driven by the realisation that lyrics and audio characteristics work best together to capture the spirit of different musical genres. Our goal is to create a genre classification system that is more reliable and accurate by incorporating data from these various sources. Furthermore, we may fully utilise multimodal data fusion in music analysis by investigating cutting-edge machine learning methods.

III. Organization of Paper

The structure of this document is as follows: Overview of the research challenge and motivation is given in the introduction.

Identify applicable funding agency here. If none, delete this.

Review of the Literature: Examines the body of work on multimodal techniques and the classification of musical genres.
Methodology: Outlines the suggested course of action, encompassing feature extraction, model selection, assessment metrics, and data preprocessing.
Experimental Results: This section presents the findings from tests carried out to gauge how well the suggested strategy works.
Discussion: Examines and examines the ramifications of the experimental results.
Conclusion: Provides an overview of the main conclusions and suggests directions for further study.

IV. Background

Traditionally, genre classification relied on manual annotation by experts, which is subjective, time-consuming, and not scalable to large music collections. Consequently, researchers turned to computational methods to automate this process. Early approaches typically extracted handcrafted features from audio signals, such as Mel-frequency cepstral coefficients (MFCCs), spectral features, and rhythmic patterns, and applied classical machine learning algorithms like Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN) for classification. Advancements in machine learning, particularly deep learning, have revolutionized music genre classification by enabling models to automatically learn hierarchical representations directly from raw audio data. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have shown remarkable performance in capturing intricate patterns and nuances in music audio, leading to significant improvements in classification accuracy.

V. Related Work

VI. Problem Formulation

A. Mathematical Framework

One way to formulate the issue of music genre classification is as follows: Predicting the genre label for each song in

TABLE I
OVERVIEW OF MUSIC GENRE CLASSIFICATION APPROACHES

Author	Year	Scheme	Advantages	Disadvantages
Poria	2015	Fuzzy clustering and hard classifier for classification.e	Identified salient features	The comparison of classifiers in the study is limited
Zhang	2022	Classification using deep learning	enhances music classification accuracy and effectiveness	Research scope limited
Jose	2012	Use entropies and fractal dimensions for classification.	Combined algorithm with high accuracy rates.	Lack of clear genre definitions can affect
Patil,	2023	Uses features from spectrographs for categorization	Novel neural network outperforms existing methods	effectiveness relies on the availability of large and diverse datasets
Xundong	2022	CNN	CNN is effective in recognizing patterns in audio	Overfitting issues observed

a dataset of songs defined by audio attributes and matching lyrics is the task at hand. Let X be the mathematical representation of the audio feature matrix and Y be the vector of genre labels. The objective is to acquire the knowledge of a mapping [f:X $\rightarrow$ Y] that correctly predicts the genre label of every song.

B. Optimization Objective

Training a classification model to minimise classification error or maximise appropriate performance parameter like accuracy, precision, or F1-score is the optimisation goal. This may be expressed as an optimisation problem in which the model's parameters are changed in order to minimise the loss function relative to the training set.

C. Evaluation metrics

The performance of the genre classification models can be evaluated using a variety of evaluation measures, such as confusion matrix analysis, accuracy, precision, recall, and F1-score. These metrics shed light on the model's ability to categorise songs accurately into the appropriate genres and point up possible areas for development.

VII. PROPOSED ARCHITECTURE

The two primary parts of the suggested architecture for multimodal music genre categorization are the vectorization of lyrics and the extraction of audio features. After audio signals are processed, pertinent features including tempo, energy, and valence are extracted using the audio feature extraction module. The lyrics vectorization module uses methods such as word embeddings to simultaneously translate song lyrics into numerical vectors. For genre classification, these modalities are then concatenated and input into several classification algorithms including SVM, Random Forest, RNN, LSTM, and CNN.

The following are the main elements of our suggested architecture:

Data Preprocessing: Taking care of missing values and scaling features, as well as cleaning and preprocessing the data from the lyrics and audio features.

Feature extraction is the process of removing pertinent features from audio data using methods like Mel-Frequency Cepstral Coefficients (MFCCs) and Word2Vec to create embeddings for lyrics.

Model selection is the process of selecting the best machine learning models for the job at hand and the data at hand, including SVM, Random Forest, RNN, and maybe CNN.

Evaluation Metrics: Using metrics to compare various methodologies and assess model performance, such as accuracy, precision, recall, and F1-score.

Our suggested architecture for multimodal music genre categorization makes use of cutting-edge methods to extract significant representations from each modality while seamlessly integrating audio elements and lyrics. The architecture is made up of various essential parts:

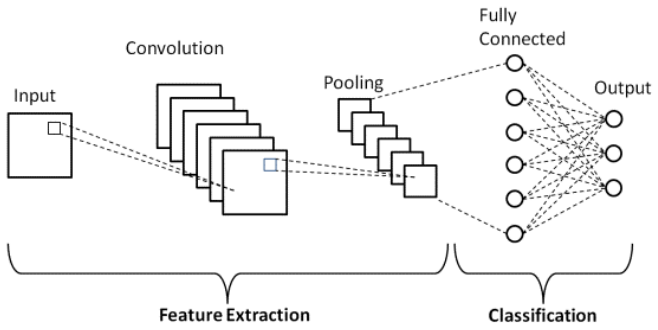

Fig. 1. CNN Architecture

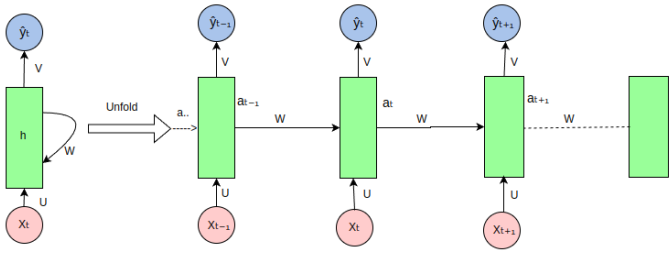

Fig. 2. RNN Architecture

1) **Audio Feature Extraction:** The initial stage is to take the raw music signals and extract pertinent audio features from them. Tempo, energy, valence, loudness, and spectral properties like MFCCs (Mel-frequency cepstral coefficients) are a few examples of these attributes. For

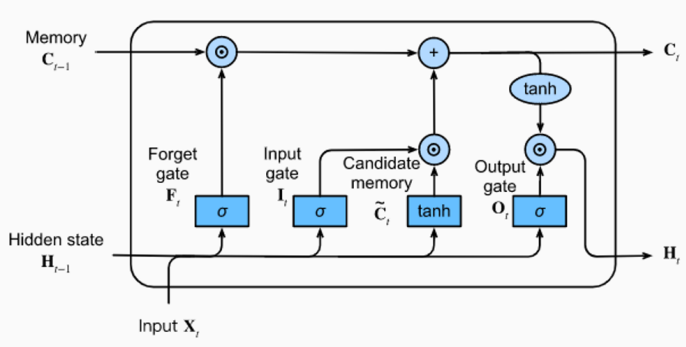

Fig. 3. LSTM Architecture

this, feature extraction methods like Librosa or tailored feature extraction pipelines can be applied.

2) **Vectorization and Lyrics Processing:** Each song's lyrics are simultaneously preprocessed to get rid of extraneous characters, punctuation, and stopwords. Next, each preprocessed word or subword unit in the lyrics is tokenized. Subsequently, word embeddings are created by converting the textual data into dense numerical vectors using methods like Word2Vec, GloVe, or FastText. These word embeddings give a rich representation of the lyrical content by capturing the semantic links between words in the lyrics.
3) **Concatenation and Fusion:** Each song's unified multimodal representation is created by concatenating the audio features and lyrics embeddings. To effectively merge the two modalities, fusion techniques like element-wise multiplication, simple concatenation, or attention processes can be used. The auditory and semantic components of the music are captured in this fused representation, which offers a thorough input for genre classification.
4) **Classification Model:** To forecast the genre label for each song, the fused multimodal representations are fed into a variety of classification models. These models include deep learning architectures like Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Convolutional Neural Networks (CNN), as well as conventional machine learning techniques like Support Vector Machines (SVM) and Random Forest. The intricacy of the data and the intended balance between interpretability and performance determine the model architecture to be used.
5) **Training and Optimisation:** Using an appropriate optimisation technique, such as Adam or stochastic gradient descent (SGD), the complete architecture is trained from start to finish. The performance of the classification models can be optimised by fine-tuning their parameters using hyperparameter tuning approaches like random or grid search.
6) **Evaluation:** Using common evaluation metrics including accuracy, precision, recall, F1-score, and confusion matrix analysis, the trained models are assessed on an independent test set. These measures assist evaluate the robustness and generalizability of the suggested design while also offering insights into the categorization performance.

VIII. Results and Discussion

We utilised the suggested architecture on a real-world dataset of songs with corresponding audio characteristics and lyrics for our experimental evaluation. Using a stratified sample technique, the dataset was split into training and testing sets to guarantee balanced representation across various genres.

A. Audio Features Visualisation

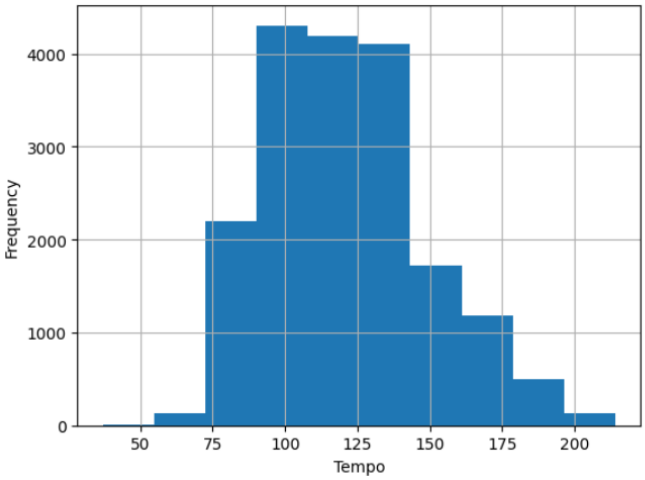

Fig. 4. Tempo vs Frequency

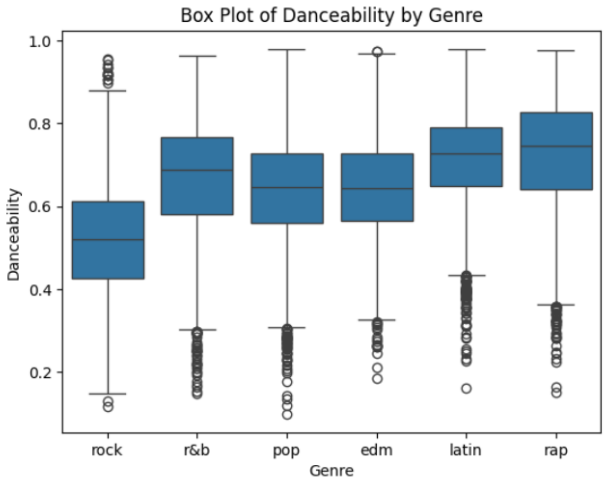

Fig. 5. Danceability vs Genre

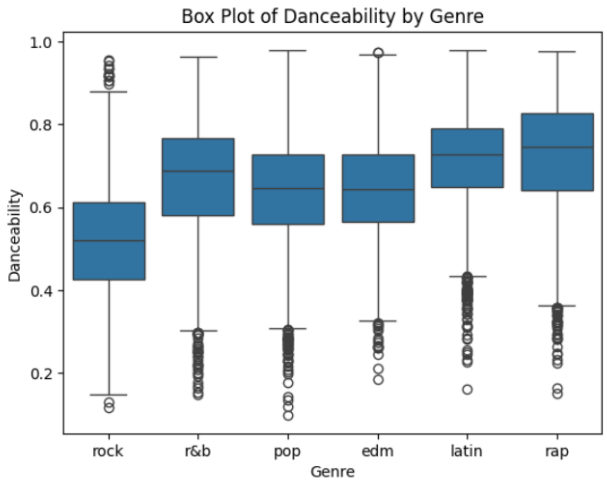

Fig. 6. Danceability vs Genre

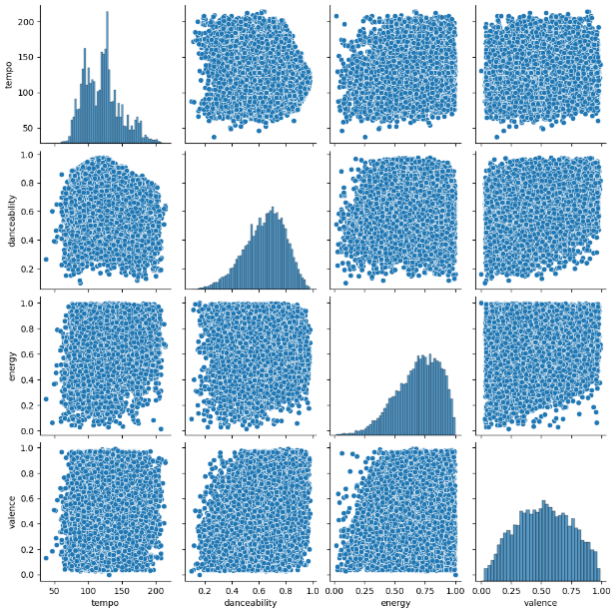

Fig. 7. Pairplot for Tempo , Danceability, Energy, Valence

B. Experimental Setup and Rules

Our suggested design was put into practice with the help of well-known Python tools like Scikit-learn, TensorFlow, and Gensim. After preprocessing the dataset, Word2Vec was used to tokenize the lyrics and turn them into word embeddings, and Librosa was used to extract audio features. In order to evaluate the effectiveness of several classification models on the genre classification problem, we conducted experiments with SVM, Random Forest, RNN, LSTM, and CNN.

Data Splitting: An 80-20 ratio was used to divide the dataset into training and testing sets.
Cross-validation: To make sure our results are robust, we use k-fold cross-validation.

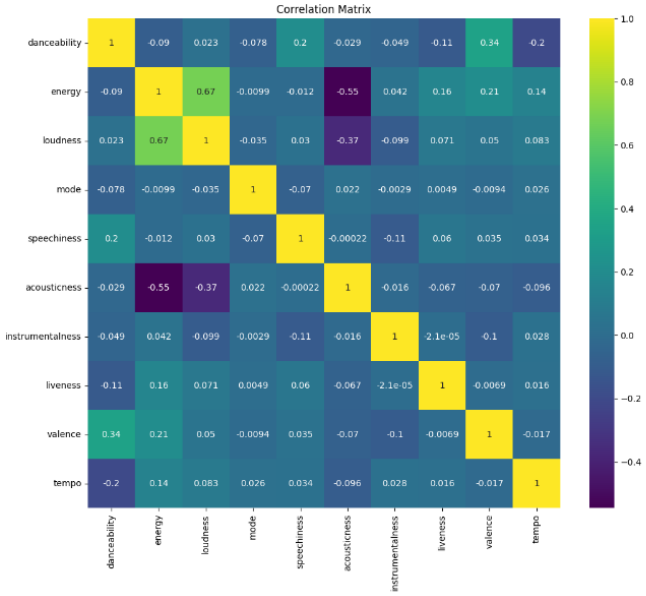

Fig. 8. Correlation matrix

Model	Accuracy
SVM	59.49
Random Forest	61.31
XGB Classifier	64.45
RNN	71.32
LSTM	80.34
BiLSTM	84.2

Hyperparameter tuning: To optimise the model's hyperparameters, grid search and cross-validation are employed.
Evaluation Metrics: We use the F1-score, accuracy, precision, and recall to assess the performance of the model.

C. Evaluation Metrics

Accuracy, precision, recall, and F1-score were among the several evaluation criteria used to assess each classification model's performance. Confusion matrices were also examined in order to pinpoint frequent misclassifications and obtain understanding of the advantages and disadvantages of each model. We evaluated the multimodal architecture's performance against baseline models that classified genres only based on audio characteristics or lyrics. The multimodal strategy routinely beat the unimodal baselines, as shown by our data, proving the value of combining the two modalities for better classification accuracy. We also looked into the effects of various fusion methods for fusing lyrics and audio elements. Comparing basic concatenation or element-wise multiplication to specific fusion procedures, such attention processes, produced better classification results, according to our research. Additionally, we performed cross-validation studies and evaluated the suggested architecture's performance on unknown data in order to determine its robustness and generalisation. According to the findings, the multimodal

technique demonstrated good generalisation abilities across a variety of datasets and genres. Overall, the results of our experiments show that the suggested architecture for multimodal music genre classification is effective. Compared to unimodal approaches, our methodology achieves higher classification accuracy by utilising both lyrics and audio characteristics. This lays the groundwork for future studies in multimodal learning and deep learning architectures for music understanding and recommendation systems, and emphasises the significance of taking into account many modalities in music analysis tasks.

IX. Conclusion

In this study, we proposed a multimodal technique that combines auditory characteristics and lyrics to classify music genres. In comparison to models that solely use audio features, our experimental results show how well these modalities work together to increase genre classification accuracy. This study advances the field of music information retrieval systems and creates new opportunities for the study of deep learning architectures and multimodal learning in the context of music analysis.

X. Future Scope

Future work will look into merging audio features and lyrics using more complex deep learning architectures, such CNN and LSTM. Further enhancing classification performance may involve examining the effects of various word embedding strategies and optimising pre-trained language models on lyrics vectorization. Moreover, expanding the research to encompass more extensive and varied datasets and integrating contextual data and user preferences may improve the suitability of the suggested methodology for actual music recommendation systems.

References

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.