Music Genre Classification with Multi-Modal Properties of Lyrics and Spectrograms

Janitha Madushan, Ruwan Weerasinghe
Informatics Institute of Technology, Sri Lanka

***Abstract*—Music Genre classification i s w idely u sed i n online music streaming platforms. Deep learning has enabled extracting musical information more effectively, and there have been various research works done to improve their accuracy with power spectrogram images and lyrical features. This paper evaluated the optimum usage of multiple modalities such as lyrics and spectrogram images based on the richness of their features. Furthermore, it proposes a hybrid-fusion-based deep learning multi-modal, multi-class classifier, t hat e mploys Mel Spectrograms, Mel-Frequency Cepstral Coefficients, a nd Lyrics to classify musical genres more accurately. Finally, the proposed model benchmarked with 3 previous studies, with a prepossessed dataset from the Music4All dataset with country, jazz, metal, and pop genre classes and obtained the highest F1-Score of 0.72 for the proposed model.**

***Index Terms*—Music Genre Classification, M usic Information Retrieval, Lyrical Information, Audio Information, Hybrid Fusion**

I. Introduction

There is a higher correlation between humans' emotions and mental status with music, and homo sapiens' specific cerebral circuit can be stimulated by music with its melody and rhythm, helping work-life balance, stress reduction, social bonding, emotional and cognitive behaviour, pain management, and improving sleeping patterns [1]. Also, they would like to listen to different kinds of music in various genre categories [2], where instrumentation, rhythmic structure, and harmonic content are the common characteristics that define a music genre [3]. Jazz, pop, classical, and metal can be classified as the most common genre classes, and they have been formed on various musical features enabling psychological and sociological research to find t he c orrelation between homunculus sensors and musical similarities. Psychological experiments have proven that different genre classes can stimulate different homunculus sensors in humans' sensory systems. For example, pop releases adrenaline and invokes aggression, whereas classical evokes calm and peace, jazz evokes depth and melancholy with nostalgia, and metal brings joy [4]. Also, the study [5] reveals, that there is a strong correlation between brain wave effects such as electroencephalography (EEG) when listening to different categories of genres of music.

With the advancement of information technology over the past two decades, an abundance of music content is added daily to online music streaming platforms. These platforms provide features to search for music content primarily by genre as the most discriminative factor of music [6]. These musical platforms can improve customer satisfaction through advanced music retrieval, Music Recommendation Systems (MRS), and indexing music content more efficiently [7]. Thus, advanced genre classifiers improve the revenues of music streaming platforms, helping music content creators to get their intellectuals to go viral, and facilitating platform listeners to find their desired music piece.

In earlier days, music classification was done manually, mostly with the support of platform users. Even though this approach is effective, manual annotation takes a long time to get sufficient tags, known as the cold start problem, the issue of music content taking quite a while to go viral. Also, humans' accuracy in recognizing genre classes of songs is 70% [6]. To answer these questions Machine Learning (ML) and deep learning-based approaches were introduced in Music Genre Classification (MGC). However, Deep Learning (DL) have overtaken ML approaches, as they provide more accurate results, with their ability to extract hidden musical features. However, DL approaches consume a lot of computation power and require a huge amount of data and storage to learn [7], so optimizing computation power and feeding an optimal number of features would be a huge advantage for researchers and the industry to build their models. The paradigm of Transfer Learning is used widely in DL approaches which can produce more accurate predictions with the lowest computation costs as they have been already trained with large datasets.

Recently, Spectrogram-based power image and Lyrics analysis with DL approaches have become quite popular in Music Information Retrieval (MIR) applications. Multi-modal-based DL models have been introduced to extract genre information from audio and lyrics more effectively in MGC. However, employing multiple spectrogram power images along with lyrics has not been evaluated on the same scale to the best of our knowledge, at the time this research was written.

So, the objective of this study is to evaluate the effectiveness of multi-modal features with different types of spectrogram power images such as Short-time Fourier transform (STFT), Mel-Frequency Cepstral Coefficients (MFCC), Mel Spectrograms (MSpec), and Lyrical features contribution in MGC with deep learning-based classifiers with an extracted sub-

Correspondence: Janitha Senadeera (E-mail: janithasen@gmail.com)
Received: 16-06-2024 **Revised:** 12-08-2024 **Accepted:** 09-09-2024
Janitha Senadeera is from Informatics Institute of Technology (janithasen@gmail.com) and A.R. Weerasinghe is from University of Colombo School of Computing (arw@ucsc.cmb.ac.lk)
DOI: https://doi.org/10.4038/icter.v18i2.7292
The 2025 Special Issue contains the full papers of the abstracts published at the 24th ICTer International Conference.

dataset from the Music4All dataset. Finally, developing a late fusion-based DL classifier employing the optimum number of multi-modal features provides higher accuracy than existing state-of-the-art (SOTA) models.

II. Related work

MIR can be divided into three main categories, namely, music creation, classification, and manipulation [8], whereas features used in MGC can be divided into three main categories such as high-level (genre, social tags, mood, lyrics), mid-level (melody, pitch, chorus), and low-level (temporal, timbre) [9]. Also, MGC systems consist of two main components, such as feature extraction and classifier [7].

Pop songs have catchy choruses, simple structures, easy dance rhythms, and less complicated lyrics. Jazz music consists of unusual harmonies based on African rhythms and European harmonies, and lyrics are often made up during the improvisation sometimes syllables without meanings and just vocals. The lyrics of rock music are mostly like pop music while, music consists of high energy levels, electric guitar sounds, and drums. Folk music is based on historical lyrics, styles or melodies with modern songs. Modern folk music consists of contemporary sounds with traditional elements, and lyrics are often storytelling and extensive. Metal is a sub-genre of Rock music where metal has monumental, distorted sound with loudness and high energy levels. However, lyrics are often distinct from one to another genre [10].

Spectrogram image analysis-based Fast Fourier Transformation (FFT) is widely used in MIR because these two-dimensional power images can extract deep musical features, and provide higher accuracies in MIR tasks, also hand-crafted audio frequencies are much more difficult to feed into deep-learning networks [6] [11]. Also, SOTA natural language processing (NLP) models have gained significant interest in MGC [10].

Nowadays, there are many MIR applications have been published based on multiple modalities, transfer and deep learning concepts [12] [13] [14] [6]. Multi-modal fusion mechanisms enable employing multiple features making more accurate models with diverse features. There are mainly two fusion approaches, namely early and late fusion. Early fusion combines data into a single classifier whereas late fusion extracts features and combines them. There are mainly three types of fusion methods, such as feature concatenation, decision weighting and hybrid fusion. The ability to identify deep musical features, and employ multiple musical features in different modalities improves the accuracy of MIR tasks with DL and multi-modal approaches. The Transfer Learning (TL) concept is popular because it helps to develop models based on previous knowledge, which increases the accuracy of models and saves computation power, enabling everyone to research their domains [6], and its usage is emerging in MIR tasks.

Music Auto Tagging (MAT) systems were introduced to supplement the shortage of social tags which causes the cold start and label sparsity problem, which adversely affects the collaborative filtering (CF) algorithms depending on social tags [9]. The study [9] employs multi-task and multi-modal learning utilizing lyrical and audio features. Audio features were extracted as MSpec and utilized CRNN network with Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), whereas lyrical features were extracted with Word2Vec and Glove embeddings with an attention network, and studied the correlation between tags.

Music Emotion Detection (MED) is widely used in Music Recommendation Systems (MRS) where the recommendation results are customized based on the emotional effect of users [2]. The study [2] proposes a personalized music recommendation framework based on previous preferences and emotions based on Joslin's theory. Acoustic characteristics such as beats, tempo, pitch, and MFCC were utilized in the study as features while utilizing a Long Short Term Memory (LSTM) based model based on spectrograms, genre and emotional features to make predictions. The study [15] combines audio and lyrics to propose a multimodal Music Emotion Recognition (MER) model, which jointly predicts arousal values and valence features. This study created a hierarchical structure fusing multiple modalities with cross-model attention (CAM) mechanism. CovNet Model was used to extract features of MSpec, while pre-trained BERT was used to extract lyrical features. Finally, high-dimensional features were concatenated using the middle fusion mechanism.

There are many studies done in MGC using uni-modals. The study [11] evaluated multiple DNN models such as ResNet50, DenseNet121, MobileNetV2, VGGNet16, and NASNetMobile against four datasets containing genres and emotions, such as FMA, GTZAN_4, GTZAN_10, and EMA, with audio features were extracted using MSpec. This study proved Resnet50, was outperforming while the best-performing model was Resnet50_trust based on an improved balanced loss function. The study [16] employed the GTZAN_10 dataset to extract audio data with MFCC. Initially, the data was categorized using a CNN network, and fed into three different architectures such as CNN, Multi-Layer Perception (MLP) and LSTM. CNN was found as the best-performing model. The study [17] employed the GTZAN dataset for extracting audio features as MSpec. The ATMGCM model proposed is based on Resnet32, the model is an ensemble model for regression and classification, and it was benched marked with SVM and Random Forest (RF). The study [7] showcases the effectiveness of MScale for Sgrams. It employed Karnatic, Hindustani, Homberg, and GTZAN audio datasets, where the Xception model was outperforming with MSpec. The study [6] utilized the GTZAN dataset with multiple ImageNet-based models and showed the Resnet34 model was outperforming.

In MIR, Audio, lyrics, album covers, text, etc, are utilized when performing multi-modal learning. The study [8] employed the GTZAN dataset to extract audio features such as MFCC, MSpec, and STFT to develop a late fusion-based classifier, and they obtained 1.0 accuracy for their model for this less noisy dataset. The study [12] developed a multi-modal classifier utilizing the BERT-base and CNN to extract lyrical and spectrogram-based information. The MetroLyrics dataset was employed, and the audio content was downloaded where the study utilized jazz, country, folk, hip-hop and metal genre

classes. Also, they benchmarked their study with BOW, Glove, HAN, SVM, and ViLBert, where their hybrid fusion model outperformed. The study [14] utilized the Music4All dataset to build a multi-modal classifier for developing a cross-attention-based model utilizing audio with MSpec and lyrics with a BERT-base and CNN-based model. It has mainly three main modules such as learning, fusion and genre correlation extraction module using a Graph Convolution Network (GAN). Also, the study benchmarked two previous studios based on the dataset and obtained the highest accuracy. The study [13] proposes a multi-modal, multi-task, and muti-label network for MGC and a novel Emotion regression approach based on the Music4All dataset only for the English songs. The study divided music genre and emotions into 44 coarse and 255 fine-grain categories. Also, the study proposes filter and channel separable convolution to reduce the complexity of the dense residual network employing lyrics and audio data. MSpec is used to extract audio information while GloVe pre-trained embedding was used for lyrical information with some pre-processing on the text. The Hierarchical Attention Network (HAN) was used to extract lyrical details. Finally, the study benched marked with one of the previous studies based on the dataset. The study [18] proposed a novel approach to employ encoded values of textual and musical features using MSpec. The CNN network generated spectrogram encoding and the BERT network generated the lyrical encoding, where the features combined using multiple co-attention networks (DCN), employing QKV attention (Question answer Key) value pairs to generate lyrical embedding pairs. Finally, the proposed model benched marked with XGB (Extreme Gradient Boosting), Multi-frame KNN, Ensemble approaches, and BERT-based multi-class classifier.

TABLE I: Transfer Learning Models used in MIR

MIR	Model	Study
MGC	BERT	[14] [12] [18] [10] [19]
	Resnet	[7] [6] [17] [4] [20] [11]
	VGG	[21] [22] [18]
	Inception	[23]
Movie Genre Classification	Resnet	[23]
	Inception	[24] [25]
MED	BERT	[26]
	Resnet	[22] [11]

Recent studies show deep learning approaches outperform traditional methods in MIR [8] [10] [12] [11]. Also, the Table I shows recent studies done using transfer learning models in MIR, where BERT and Resent-based models have been widely used in recent studies.

The Bidirectional Encoder Representations from Transformers (BERT), is a transfer learning model introduced by the Google AI team in 2008 [27]. It is used in many MIR studies for extracting information from lyrics, reviews, etc. The self-attention mechanism is employed in the BERT model to generate queries, values and keys, where the mechanism uses each token's representation, so that the mechanism allows each token to attend every other token in the sequence helps to capture contextual information bi-directionally.

The study [12] shows that BERT can extract the semantic meaning of lyrics so that it performs better than HAN, and Glove networks. They benchmarked accuracies of BERT-base over Glove, SVM, HAN, and ViLBert. The study [10] shows BERT outperforms traditional ML classifiers such as SVM, Logistic Regression, Naïve Bayers, Random Forest, etc with Glove and Word2Vec embeddings in MGC. They also stated that BERT performs better with English data whereas XLM-RoBERT performs better in multilingual data. The study [28] trained a fusion model with BERT and CNN where they freeze-trained BERT on lyrics when training the fusion model. The study [19] shows that BERT can understand lyrical information and they used a BERT-base model. All these studies show the capability of BERT lyrical information extraction, so the BERT-base model used in this study consists of 768 hidden units, 12 attention mechanism heads and 12 hidden layers. The optimum learning rate of 2e-5 was used to train the BERT model used in the study [12].

CNN-based models are widely used in MGC because of the ability to extract spectrogram image-based musical features [17]. Traditional approaches namely, KNN, Naïve bayers, SVM and Decision trees are less performant in MGC [10].

The ImageNet challenge held in 2015, introduced many CNN-based architectures based on the ImageNet dataset, such as MobileNet, Resnet VGG, DenseNet, AlexNet, SqueezeNet, GoogleNet, ShuffleNet, InceptionV3 having different architectures. Among them, the Restnet50 model was the best architecture where the inception blocks address the vanishing gradient problem that occurs when the networks go deeper [17] [6]. The study [10] outlines the ability to distinguish overlapping genre classes of residual blocks. Also, the survey conveys an inception block and can capture low-level features in the spectrogram images. The study achieved higher accuracy [17] with Resnet32 over other traditional methods. The study [6] status of the optimal learning rate for Resnet50 is 2e-5 in MGC for newly added layers.

TABLE II: Recent Studies utilizing Spectrogram images

Study	Research	Features
[2]	MEC	MFCC
[9]	MAT	MFCC
[8]	MGC	STFT, MFCC, MSpec
[14]	MGC	MSpec
[15]	MGC	MSpec
[12]	MGC	MSpec
[6]	MGC	MSpec
[4]	MGC	Timbral, Chroma, and Harmonic Percussive
[13]	MGC, MEC	MSpec
[20]	MGC	MSpec
[23]	MGC	MFCC, Chroma, Tempogram
[16]	MGC	MSpec
[22]	MEC	MSpec
[28]	MEC	MSpec
[17]	MGC	STFT, MFCC, MSpec, SGram, SC
		CQT, HMS, Chroma, Tonnetz
[18]	MGC	MFCC, Chroma, Spectral Centroid,
		Spectral BW, Spectral Contrast

The Table II shows studies done recently in MIR using spectrogram images along with the features employed, where MSpec, MFCC and STFT have been widely used. Spectrograms determine details of flux and pitch with their frequency changes of songs while colour reflects musical energy [11], and they are generated by transforming audio

frequencies into power-based images using Discrete Fourier transformation [17]. In 1937 Newman and Volkmann invented the Mel Scale to ignore higher frequencies but to keep lower frequencies where it does the same in the human peripheral system [6]. Also, these features can be categorized into mid-level (Tonnetz, Chromogram), and low-level (STFT, MFCC, MSpec) features [7]. The study [8] shows that MFCC, STFT, MSpec, Tempo, and Chromogram features are widely used in MIR, and MFCC can mimic the human auditory peripheral system [7]. Also, the study [7] ranked different power spectrograms based on their effectiveness in MGC. This study did a model performance comparison with multiple datasets, spectrogram images, and various deep learning-based models. The results show that MSpec and MFCC are on the highest ranks, while Spectral contrast (SC), STFT, and Constant-Q transform (CQT) show no significant difference in ranking. Even though Swaragram (SGram) is in the third position in the ranking, it is not supported in the Librosa Library used in the literature. Considering all the factors MFCC, MSpec and STFT Spectrograms have been chosen to conduct this research work.

$$s = A_1 \sin(2\pi f_1 t + \phi_1) + A_2 \sin(2\pi f_2 t + \phi_2) \quad (1)$$

Fourier transformation is the process of converting a sinusoidal signal into the frequency domain from the time domain. The equation 1 shows the equation used in Fourier Transformation, where the original signal (s) can be defined as a combination of multiple sine waves with different amplitudes and phases, and this is the fundamental process of generating spectrogram images [16].

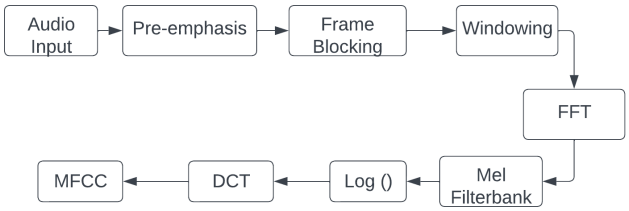

Fig. 1: Block Diagram of the MFCC Feature Extraction

MFCC can mimic the human auditory peripheral system [7]. Figure 1 shows the block diagram of generating MFCC spectrogram from the audio sound. The pre-emphasis filter is applied to the audio signal to boost higher frequencies, also in the framing, the continuous signal is divided into chunks using overlapping frames to keep the signal waveform constant. In windowing, typically the Hamming window is applied to reduce the spectral leakage. Then Fast Fourier Transformation (FFT) is applied to the waveform to convert it into the frequency domain, and the magnitude of the spectrum passes through the Mel filter bank where it has a series of triangular filters that can approximate human ears' responses to different frequencies. From the logarithm operation, human ears' perception of loudness is approximated. Finally, Discrete Cosine Transform (DCT) is applied to the outputs of the filter bank energies, where it is a crucial step to reduce the number of coefficients to keep only significant information. MSpec representation perceives a higher-dimensional time-frequency information whereas MFCC only keeps the most significant coefficients by the DCT transformation step. So, the DCT is ignored during the MSpec generation. STFT provides the linear representation of frequencies over time producing a 2D array of time vs frequencies, and it contains precise spectral information without any transformation. So, the process stops at FFT when generating STFT spectrograms.

TABLE III: Properties used when generating spectrograms in different studies.

Study	Sampling Frequency	Hop length	Mel Filters	MFCC
[9]	0 12 kHz	256	96	
[13]	22050 Hz	512	128	
[22]	22050 Hz	1024	128	
[18]		256	96	
[28]	44.1 kHz	320	64	
[16]	22050 Hz	2048	128	
[14]	22050 Hz	512	128	
[12]	12Hz	512	96	
[3]				20

Studies in Table III show properties used when extracting audio features when generating spectrogram images in previous studies where 22050 Hz of sampling frequency, 512 hopping length, 128 Mel filters, and 20 MFCC are commonly used.

The study [29] shows even a 250ms track length is sufficient for extracting sufficient information from audio, as the humans' average length for extracting information is 1.6s so they have used 3s soundtracks in their study, also they show even 2s audio track lengths are sufficient for recognizing human feelings such as fear, happiness, and sad. The study [20] utilized 3s of soundtrack length to generate spectrograms whereas the study [11] used 5s of track lengths providing higher accuracies. So, it is clear that 3s of soundtrack length is sufficient for generating spectrogram images. Also, the study [9] used (96, 1366) as the spectrogram image size whereas studies [21], [30] [6] used (224, 224) in their studies. And, the study [6] shows (224, 224) image size is optimum for training Imagenet-based models.

III. Methodology

A. Dataset

Even though there had been multiple datasets used in MIR studies, such as Million Songs and GTZAN, none of them had both lyrics and audio data. However, the Music4All [31] dataset contains both lyrics and audio data, and the authors of the dataset, generously gave us the dataset for this study. Only English songs were selected that had both lyrics and audio tracks. And, the study was initially set up to study the most common ten genre classes such as classical, jazz, country, disco, hip hop, rock, blues, reggae, pop, and metal. Each song in the dataset had been tagged with multiple genre classes using a string like "hard rock, rock, classic rock", so that cosine similarity was obtained to determine the genre class with the primary classes that had been set initially in the study, also programmatically the classes that have a smaller number

of data points were given higher priority when determining the genre class to treat the class imbalance to some extent.

TABLE IV: Pre-processed Genre classes

Genre	Count
Rock	13570
Pop	12029
Metal	8075
Country	850
Jazz	631
Hip Hop	336
Blues	235
Reggae	200
Classical	102
Disco	64

The table IV shows the intermediate dataset after extracting genre classes but it is highly imbalanced. Experimentally, it was found, that employing pop and rock classes provided lower accuracy in the classification and, with the support of the literature, only the pop class was selected because they have similar characteristics [10]. Finally, only the pop, jazz, country, and metal classes were chosen for this study. However, for further treatment for class imbalance, pop, metal, and county classes were downsampled, whereas the Jazz class was upsampled to 800. As the upsampling technique, the lyrics were duplicated, and spectrograms were generated from different offsets (time wrapping) of the soundtrack because spectrograms have unique properties to maintain where general image augmentation techniques don't support [32].

MSpec, MFCC, and STFT spectrograms were generated using the Librosa open-source library with 22050 Hz of the sampling frequency, 512 hop length, 128 Mel filters, 20 MFCC coefficients, and 3s soundtrack lengths, as they were common and optimum values found in the literature review. And, images were resized to (224, 224) [6]. Finally, the dataset was split into 0.7, 0.2 and 0.1 for train, test, and validation datasets.

B. Neural Networks

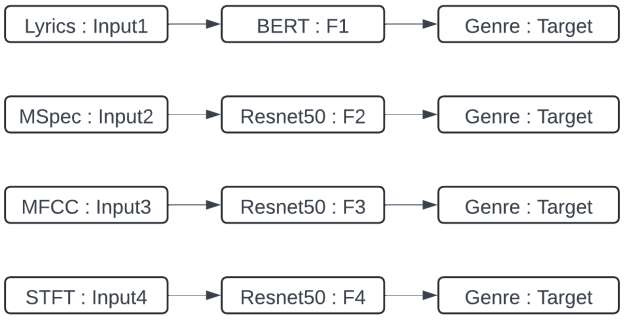

Fig. 2: Uni-Models

The CLS output of the BERT-base model was obtained by an encoder function where the layers of the base model were frozen. Then 0.1 dropout was added followed by a dense layer with 512 features before the output layer. The Adam optimiser was set with the 2e-5 learning rate [12] and categorical_crossentropy as the loss function.

An encoder function was written to return Resnet50 model layers with "ImageNet" weights and freeze the base models' layers. Then the encoded layer was followed by a GlobalAveragePooling2D layer, a dense layer with 1024 features, a 0.49 dropout layer, and an output layer. The Adam optimiser was set with a 2e-5 learning rate [6] with categorical_crossentropy loss function.

The BERT model was trained for 100 epochs with lyrics extracted from the dataset, and 64 as the batch size, whereas MSpec, MFCC, and STFT spectrograms were trained using three models with the same Resnet50 model configurations, while MSpec was trained for 74, MFCC for 77 and STFT for 47 epochs. Figure 2 shows the high-level architecture of each uni model.

C. Feature Selection

$$\text{Pred}_{\text{fusion}} = \alpha.\text{Pred}_{\text{Lyric}} + \beta.\text{Pred}_{\text{MSpec}} + \gamma.\text{Pred}_{\text{MFCC}} + \delta.\text{Pred}_{\text{STFT}} \quad (2)$$

Decision weighting was performed with multiple features used to determine the effectiveness of the combination of features. The Equation 2 is used to predict the final output based on multiple predictions of models where α + β + γ + δ = 1, each constant determines the contribution of each feature on the classification. However, this study assumes each feature contributes the same amount to the final prediction so that according to the experiments, equal values were given to the coefficients.

TABLE V: Weighted Prediction Results

	Lyrics α	MSpec β	MFCC γ	STFT δ	F1-Score
1	0	1	0	0	0.63
2	1	0	0	0	0.62
3	0	0	1	0	0.56
4	0	0	0	1	0.57
5	0.25	0.25	0.25	0.25	0.68
6	0.33	0.33	0.33	0	0.71
7	0.5	0.5	0	0	0.70
8	0.5	0	0.5	0	0.66
9	0.5	0.5	0	0	0.70
10	0.5	0	0	0.5	0.67
11	0.33	0.33	0	0.33	0.67
12	0.33	0	0.33	0.33	0.67
13	0	0.5	0.5	0	0.65
14	0	0.5	0	0.5	0.61
15	0	0	0.5	0.5	0.61
16	0	0.33	0.33	0.33	0.63

The Table V shows different F-values obtained when using different coefficient values for features in the decision weighting stage, and the maximum accuracy, which is 0.71 was obtained when employing MSpec, MFCC, and Lyrics combination.

D. Feature Concatenation

According to Figure 2, after training each uni-model, high-dimensional F1, F2 and F3 features have been extracted from BERT and Resnet50 models of Lyrics, MSpec, and MFCC,

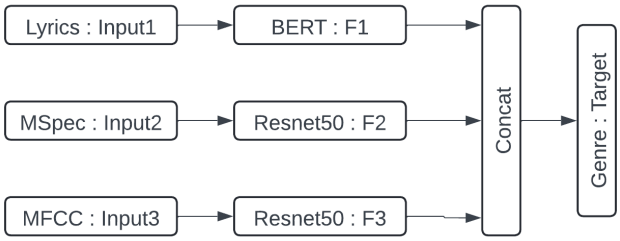

Fig. 3: Feature concatenation

according to the Figure 3 which shows the high-level block diagram of the feature concatenated model. The features (F1, F2, F3) were extracted where base model layers have been frozen, concatenated and a target was added after a dense layer with 256 features and dropout of 0.3. In the literature, it was found 2e-5 learning rate is common for both BERT and Resnet50 in MGC tasks so the same rate has been set with the Adam optimizer, the categorical_cross_entropy function used with the batch size of 32 and trained for 50 epochs. Finally, an extended version of Equation 2 is used to perform a hybrid-fusion with the same assumptions made on the coefficients at the decision-weighting stage.

TABLE VI: Prediction results of different fusion strategies.

Method	F1 - Score	Precision	Recall
Decision Weighting	0.71	0.71	0.71
Feature Concatenation	0.70	0.71	0.71
Hybrid Fusion	0.72	0.73	0.73

TABLE VII: F1-Scores of each fusion method & genre classes

Method	Country	Jazz	Metal	Pop
Decision Weighting	0.71	0.68	0.84	0.60
Feature Concatenation	0.69	0.67	0.84	0.61
Hybrid Fusion	0.71	0.68	0.85	0.64

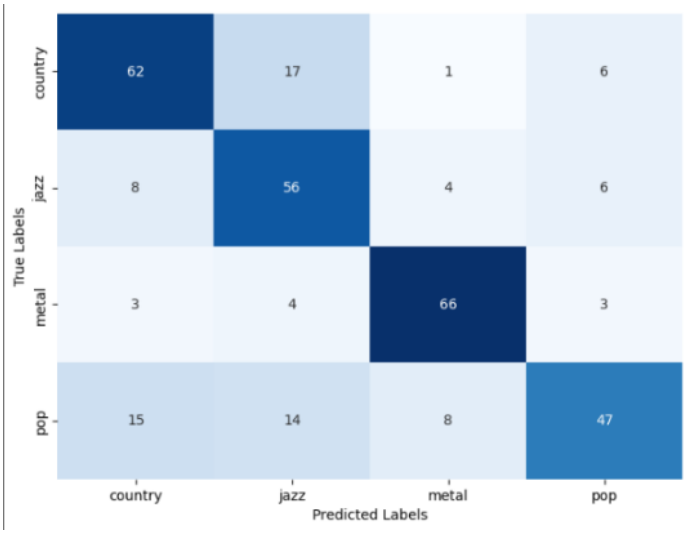

Fig. 4: Confusion Matrix of Hybrid Fusion Method

Table VII shows the accuracy measures of each fusion-based method used in the study whereas Table VII shows accuracy measures of each genre class against different fusion methodologies. And, Figure 4 shows the confusion matrix obtained of the hybrid-fusion approach.

E. Baseline Studies

Three main recent studies were selected for benchmarking based on different criteria. The proposed model based on the hybrid-fusion mechanism in the study [12] slightly changed due to the overfitting observed, so the last two convolution blocks and strides were removed from the CNN encoder, and dropouts were set to 0.5. Also, in the original study, the optimum values of the contribution coefficients of features were determined by multiple iterations, but this study has given the same priority to all the features. The proposed fusion method of MSpec, MFCC, and STFT in the study [8] slightly changed by adding L2 Kernel Regulation with 0.02 experimentally to avoid over-fitting. The recent study [14] is also based on the Music4All dataset, genre correlations extraction module for genre corrections (GCEM) was ignored because the extracted dataset only supported multi-class problems, and A-L loss was removed as this study is given the same precedence to all features.

TABLE VIII: Prediction results of different models.

Research	F1 - Score	Precision	Recall
[12]	0.67	0.68	0.68
[8]	0.55	0.54	0.56
[14]	0.44	0.39	0.46
Proposed Model	0.72	0.73	0.73

TABLE IX: Prediction results of each genre class.

Method	Country	Jazz	Metal	Pop
[12]	0.71	0.62	0.83	0.48
[8]	0.45	0.59	0.80	0.35
[14]	0.20	0.54	0.67	0.02
Proposed Model	0.71	0.68	0.85	0.64

Table VIII shows performance measures of each baseline study based on the extracted dataset of the Music4All dataset whereas Table IX shows classification results of each genre category of baseline studies.

IV. Conclusion

The proposed model is at least 5% more accurate than the baseline models tested, and this study outperforms the latest study done based on the Music4All dataset related to the MGC. The study [8] employed MSpec, MFCC, and STFT but according to the Table V the overall accuracy is 0.63 for the particular combination whereas the highest accuracy was obtained for the combination of MSpec, MFCC and Lyrics, where inclusion of STFT drops the overall model's accuracy due to distractions, as it is a more machine-friendly spectrogram failing to capture audio features comply with humans' auditory system. Also, the comparative analysis shows a notable variation in performance matrices using F1-Score, Precision and Recall. The results indicate that the proposed

model outperforms other methods significantly concerning the performance matrices discussed, and this shows the ability of the proposed method to capture features accurately, highlighting the importance of selecting and combining the right modalities and techniques for the MGC task.

The proposed method shows higher performance across all genres, particularly Jazz, Metal, and Pop. Baseline study [12] also shows a higher performance but it is lagging when predicting the pop genre. Even though the study [8] shows higher accuracy in Metal genre classification, lagging behind other genres. The study [14] shows the lowest accuracy of all, even though Jazz and Metal show higher accuracy on the classification, classifying Country and Pop was not that accurate, adversely affecting the overall accuracy of the model, indicating the effectiveness of the GCEM module proposed in the study.

The proposed method demonstrates higher classification accuracies for each genre class (Country, Jazz, Metal, and Pop) while maintaining the overall accuracy. Thus, these observations show the models' generalizability, robustness, and versatility, and suggest the models' ability to capture unique characteristics of different genre classes. Also, the model shows low variance with F1-scores ranging from 0.64 (Pop) to 0.85 (Metal) whereas the other studies showed higher variances. However, the confusion matrix in Figure 4 shows the highest confusion between Jazz and Country classes, the reason for this is Jazz initially included genre classes like Country but later different styles were made up by focusing on relaxation, variation, and improvisation [12].

This study and benched marked previous studies showed nearly higher accuracies of predicting, proposing the prepossessing strategies were quite accurate and efficient and helped to reduce the dataset drastically while keeping necessary features. The Jazz class was up-sampled with a unique strategy and Table VII shows the test accuracy of this genre class is also higher.

When considering Decision Weighting, Feature Concatenation, and Hybrid fusion, each model scored quite close results, indicating each method is relatively effective. Also, balanced scores of F1, Precision and Recall suggested neither false positives nor false negatives adversely affect the results, which makes the proposed method suitable for applications that require balanced results. The feature concatenation also maintains high precision and recall but the hybrid fusion method shows superior performance showing that it can capture more musical information. Also, most of the hyper-parameters were set with the support of the literature review so that most of the computation power was saved, also this study shows the effectiveness of utilizing transfer learning approaches on MGC. And, the higher accuracies obtained for each modality and their combinations on the Table V depicts the effectiveness of Resnet50 and BERT-base transfer learning models in MGC.

As a future improvement, the proposed model can be evaluated with a large dataset with multiple genre classes at least the basic 10 genre classes shown in this study on the table IV. Also, this study is particularly done for MGC but muti-task learning approaches can also be utilized to extend this study for emotion detection. Previous studies were also mentioning the Music4All dataset is a bit noisy, and highly imbalanced, so the proposed model can be evaluated with a quality and a large dataset. Also, different offset values can be used to augment spectrogram images, and emerging Generative approaches can be used to generate slightly different lyrics from the original lyrics to treat the imbalance of the Music4All dataset. Furthermore, this study mainly showed the effectiveness of features by giving the same precedence for all the features but the decision weighting step can be optimized with a sequence of values to find the optimum values for coefficients also, when setting up these coefficients multiple datasets can be used to determine more generalised values.

REFERENCES

[1] S. G. Abhyankar, S. S. Bharadwaj, G. S. Rani, P. G. Karigiri, S. Srikanth, and S. Gurugopinath, "A survey on music genre classification using multimodal information processing and retrieval," in *2023 International Conference on Recent Trends in Electronics and Communication (ICRTEC)*, 2023, pp. 1–6.

[2] Z. Liu, W. Xu, W. Zhang, and Q. Jiang, "An emotion-based personalized music recommendation framework for emotion improvement," *Inf. Process. Manage.*, vol. 60, no. 3, may 2023. [Online]. Available: https://doi.org/10.1016/j.ipm.2022.103256

[3] D. S. Lau and R. Ajoodha, "Music genre classification: A comparative study between deep learning and traditional machine learning approaches," in *Proceedings of Sixth International Congress on Information and Communication Technology*, ser. Lecture Notes in Networks and Systems, X.-S. Yang, S. Sherratt, N. Dey, and A. Joshi, Eds. Springer, 2022, pp. 239–247. [Online]. Available: https://doi.org/10.1007/978-981-16-2102-4_22

[4] D. Sharma, S. Taran, and A. Pandey, "A fusion way of feature extraction for automatic categorization of music genres," *Multimedia Tools and Applications*, vol. 82, no. 1, pp. 25 015–25 038, 2023. [Online]. Available: https://doi.org/10.1007/s11042-023-14371-8

[5] V. S. Saravanarajan, R.-C. Chen, N. A. Saravanarajan, M.-Z. Liu, C.-Y. Lin, and L.-S. Chen, "Effect of different genres of music on brain waves," in *2021 Emerging Trends in Industry 4.0 (ETI 4.0)*, 2021, pp. 1–5.

[6] J. Mehta, D. Gandhi, G. Thakur, and P. Kanani, "Music genre classification using transfer learning on log-based mel spectrogram," in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, 2021, pp. 1101–1107.

[7] Y. Singh and A. Biswas, "Robustness of musical features on deep learning models for music genre classification," *Expert Systems with Applications*, vol. 199, p. 116879, 03 2022.

[8] S.-H. Cho, Y. Park, and J. Lee, "Effective music genre classification using late fusion convolutional neural network with multiple spectral features," in *2022 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, 2022, pp. 1–4.

[9] H.-C. Wang, S.-W. Syu, and P. Wongchaisuwat, "A method of music autotagging based on audio and lyrics," *Multimedia Tools and Applications*, vol. 80, no. 10, pp. 15 511–15 539, 2021. [Online]. Available: https://doi.org/10.1007/s11042-020-10381-y

[10] A. Marijić and M. Bagić Babac, "Predicting song genre with deep learning," *Global Knowledge, Memory and Communication*, 2023. [Online]. Available: https://doi.org/10.1108/GKMC-08-2022-0187

[11] J. Li, L. Han, X. Li, J. Zhu, B. Yuan, and Z. Gou, "An evaluation of deep neural network models for music classification using spectrograms," *Multimedia Tools and Applications*, vol. 81, no. 4, pp. 4621–4647, 2022. [Online]. Available: https://doi.org/10.1007/s11042-020-10465-9

[12] Y. Li, Z. Zhang, H. Ding, and L. Chang, "Music genre classification based on fusing audio and lyric information," *Multimedia Tools and Applications*, vol. 82, no. 13, pp. 20 157–20 176, 2023. [Online]. Available: https://doi.org/10.1007/s11042-022-14252-6

[13] Y. R. Pandeya, J. You, B. Bhattarai, and J. Lee, "Multi-modal, multi-task and multi-label for music genre classification and emotion regression," in *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, 2021, pp. 1042–1045.

[14] G. Ru, X. Zhang, J. Wang, N. Cheng, and J. Xiao, "Improving music genre classification from multi-modal properties of music and genre correlations perspective," in *ICASSP 2023 - 2023 IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.

[15] J. Zhao, G. Ru, Y. Yu, Y. Wu, D. Li, and W. Li, "Multimodal music emotion recognition with hierarchical cross-modal attention network," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 2022, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/9859812

[16] Preetham, Manoj and Panga, Jemimah Beulah and Andrew, J. and Raimond, Kumudha and Dang, Hien", editor="Peter, J. Dinesh and Fernandes, Steven Lawrence and Alavi, Amir H., "Classification of Music Genres Based on Mel-Frequency Cepstrum Coefficients Using Deep Learning Models," in *Disruptive Technologies for Big Data and Cloud Applications*. Singapore: Springer Nature Singapore, 2022, pp. 891–907.

[17] C. Chen and X. Steven, "Combined transfer and active learning for high accuracy music genre classification method," in *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, 2021, pp. 53–56.

[18] L. Wadhwa and P. Mukherjee, "Music genre classification using multimodal deep learning based fusion," in *2021 Grace Hopper Celebration India (GHCI)*, 2021, pp. 1–5.

[19] V. R. Revathy, A. S. Pillai, and F. Daneshfar, "Lyemobert: Classification of lyrics' emotion and recommendation using a pre-trained model," *Procedia Computer Science*, vol. 218, pp. 1196–1208, 2023, international Conference on Machine Learning and Data Engineering. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050923000984

[20] P.-C. Chang, Y.-S. Chen, and C.-H. Lee, "Ms-sincresnet: Joint learning of 1d and 2d kernels using multi-scale sincnet and resnet for music genre classification," in *Proceedings of the 2021 International Conference on Multimedia Retrieval*, ser. ICMR '21. New York, NY, USA: Association for Computing Machinery, Sep. 2021, pp. 29–36.

[21] N. Purnama, "Music genre recommendations based on spectrogram analysis using convolutional neural network algorithm with resnet-50 and vgg-16 architecture," *JISA(Jurnal Informatika dan Sains)*, vol. 5, pp. 69–74, 06 2022.

[22] G. Tong and B. Ding, "Multimodal music emotion recognition method based on the combination of knowledge distillation and transfer learning," *Sci. Program.*, vol. 2022, jan 2022. [Online]. Available: https://doi.org/10.1155/2022/2802573

[23] H. Ding, W. Song, C. Zhao, F. Wang, G. Wang, W. Xi, and J. Zhao, "Knowledge-graph augmented music representation for genre classification," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[24] R. B. Mangolin, R. M. Pereira, A. S. Britto, C. N. Silla, V. D. Feltrim, D. Bertolini, and Y. M. G. Costa, "A multimodal approach for multi-label movie genre classification," *Multimedia Tools and Applications*, vol. 81, no. 14, pp. 19 071–19 096, 2022. [Online]. Available: https://doi.org/10.1007/s11042-020-10086-2

[25] Y. R. Pandeya and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video," *Multimedia Tools and Applications*, vol. 80, no. 2, pp. 2887–2905, Jan. 2021. [Online]. Available: https://doi.org/10.1007/s11042-020-08836-3

[26] T. T, E. Abdurachman, Y. Heryadi, and A. Zahra, "Bert transformer model for indonesian mood music dataset," in *2022 IEEE 8th International Conference on Computing, Engineering and Design (ICCED)*, 2022, pp. 1–4.

[27] G. Liu and Z. Tan, "Research on multi-modal music emotion classification based on audio and lyirc," in *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, vol. 1, 2020, pp. 2331–2335.

[28] B.-H. Sung and S.-C. Wei, "Becmer: A fusion model using bert and cnn for music emotion recognition," in *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, 2021, pp. 437–444.

[29] M. J. Lucia-Mulas, P. Revuelta-Sanz, B. Ruiz-Mezcua, and I. Gonzalez-Carrasco, "Automatic music emotion classification model for movie soundtrack subtitling based on neuroscientific premises," *Applied Intelligence*, vol. 53, no. 22, p. 27096–27109, sep 2023. [Online]. Available: https://doi.org/10.1007/s10489-023-04967-w

[30] S. Jang and Y. Kim, "Dual resnet-based environmental sound classification using gan," in *2023 17th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, 2023, pp. 1–6.

[31] I. A. Pegoraro Santana, F. Pinhelli, J. Donini, L. Catharin, R. B. Mangolin, Y. M. e. G. da Costa, V. Delisandra Feltrim, and M. A. Domingues, "Music4all: A new music database and its applications," in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2020, pp. 399–404.

[32] B. K. Iwana and S. Uchida, "Time series data augmentation for neural networks by time warping with a discriminative teacher," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 3558–3565.