IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 October 2023, accepted 18 December 2023, date of publication 25 December 2023,
date of current version 3 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3346883

RESEARCH ARTICLE

A Short Survey and Comparison of CNN-Based Music Genre Classification Using Multiple Spectral Features

WANGDUK SEO[1], SUNG-HYUN CHO[2], PAWEŁ TEISSEYRE[3,4], AND JAESUNG LEE[2,5]
[1]School of Computer Science and Engineering, Chung-Ang University, Seoul 06974, South Korea
[2]Department of Artificial Intelligence, Chung-Ang University, Seoul 06974, South Korea
[3]Institute of Computer Science, Polish Academy of Sciences, 01-224 Warsaw, Poland
[4]Faculty of Mathematics and Information Sciences, Warsaw University of Technology, 00-661 Warsaw, Poland
[5]AI/ML Innovation Research Center, Chung-Ang University, Seoul 06974, South Korea

Corresponding author: Jaesung Lee (curseor@cau.ac.kr)

This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) funded by the Korean Government (Ministry of the Science and ICT, MSIT) through the Artificial Intelligence Graduate School Program, Chung-Ang University, under Grant 2021-0-01341; and in part by the Chung-Ang University Research Grants, in 2023.

ABSTRACT The goal of music genre classification is to identify the genre of given feature vectors representing certain characteristics of music clips. In addition, to improve the accuracy of music genre classification, considerable research has been conducted on extracting spectral features, which contain critical information for genre classification, from music clips and feeding these features into training models. In particular, recent studies argue that classification accuracy can be enhanced by employing multiple spectral features simultaneously. Consequently, fusing information from multiple spectral features is a critical consideration in designing music genre classification models. Hence, this paper provides a short survey of recent studies on music genre classification and compares the performance of the most recent CNN-based models with a newly devised model that employs a late fusion strategy for the multiple spectral features. Our empirical study of 12 public datasets, including Ballroom, ISMIR04, and GTZAN, showed that the late fusion CNN model outperforms other compared methods. Additionally, we performed an in-depth analysis to validate the effectiveness of the late fusion strategy in music genre classification.

INDEX TERMS Music genre classification, convolutional neural network, spectral feature, late fusion strategy.

I. INTRODUCTION

Music information retrieval (MIR) can be broadly categorized into three main areas: music classification, manipulation, and creation [2]. Among these, music classification plays an integral role in everyday life and has experienced steady research growth. Specifically, this subfield involves the assignment of appropriate labels to musical clips based on genres or emotional moods, utilizing their musical characteristics. More particularly, given that most end users exhibit consistent tastes that do not deviate significantly from their established preferences [3], music classification facilitates music enjoyment by filtering out irrelevant genres. Further, while various applications are based on music classification [4], [5], [6], the music genre classification (MGC) task can provide a highly satisfying listening environment tailored to individual tastes.

Meanwhile, to improve the accuracy of MGC, considerable research has been conducted on extracting spectral features from music clips [7]. Specifically, these features contain critical information for MGC, which can subsequently be fed into training models. Particularly, notable extraction techniques include the short-term Fourier transform (STFT) [8], mel-frequency cepstral coefficient (MFCC) [9], mel-spectrogram (MLS) [10], tempo, and chromagram [11]. Early MGC studies [12] trained models using single-spectral

The associate editor coordinating the review of this manuscript and approving it for publication was Olutayo O. Oyerinde.

features derived from one of these extraction techniques. However, due to the inherent variety and ambiguity of music genres, more recent studies argue that the accuracy of classification tasks can be improved by employing and combining multiple spectral features obtained from various extraction techniques [28], [30], [31], [32]. This suggests two important considerations: first, selecting which spectral features and how to combine them can critically improve the accuracy of MGC. Second, the model architecture should be able to handle multiple spectral features simultaneously.

Because of the remarkable success of deep learning in various fields, various deep learning models such as a graph neural network [86], a convolutional neural network [60], and an attention model [87] can be applied to the MIR domain. Recently, convolutional neural networks (CNNs) have attracted huge research interests in the MIR domain due to their inherent ability to handle complex spectral features [38], [39], [41]. Specifically, these features, derived from music clips, are two-dimensional, making CNNs ideal for discerning spatial relationships among them, thereby capturing the temporal structures of music clips. Therefore, recent CNN-based MGC models are designed to handle either single-spectral features [43] or multiple spectral features that are concatenated as input data [66]. The latter approach is predicated on the assumption that the concatenation of multiple spectral features can provide more critical information than single-spectral features, thereby improving the accuracy of MGC. More specifically, when the combination of multiple spectral features is determined, the features are concatenated as cubic input data, and this method is called the early fusion strategy.

This paper begins by reviewing existing MGC methods in terms of their applied datasets, input features, and classification models. Specifically, this review allows us to identify the limitations of current techniques and guides us to devise a CNN-based MGC model to handle multiple spectral features. In particular, we explore a late fusion strategy that concatenates multiple spectral features after the extraction of individual information via convolutional operators to compare the performance with that of the early fusion strategy. Consequently, the devised model effectively handles multiple spectral features by extracting distinct information from each one and fusing only the essential information. Our experimental results, derived from 12 well-known music genre datasets, reveal the relative superiority of our method over conventional CNN-based methods for MGC tasks. The contributions of this study can be outlined as follows:

- We provide a short survey of existing MGC methods in terms of their applied datasets, input features, and classification models.
- We devise a late fusion CNN that utilizes multiple spectral features to establish baseline performance in comparison with conventional MGC methods.
- We conduct empirical experiments to validate the impact of the information fusion strategy for MGC tasks predicated on CNN.
- We conduct an in-depth analysis to underline the effectiveness of the late fusion strategy.

II. RELATED WORK

Many studies have thoroughly explored the MGC task, particularly focusing on spectral features selection and MGC across various music datasets [20]. Research detailing the investigation of the impact of traditional classifiers in developing an automatic MGC model [36] extracted rhythmical features from the widely used GTZAN dataset [49]. Specifically, the study employed principal component analysis for feature dimension reduction [24], [50]. Experiments were performed using conventional approaches like gradient boost [51], support vector machines (SVM), random forest [52], XG boost [53], decision tree (DT) [54], and k-nearest neighbor (kNN) [55]. Notably, the influence of linguistic content on accuracy was seldom separated from the audio [22]. Furthermore, the comparative performance of SVM and kNN revealed SVM having superior classification accuracy, closely followed by kNN [23]. In addition, a unique ensemble model of SVM and radial basis function (SVM-RBF) was proposed for music clip classification within the Spotify dataset [25]. Meanwhile, the Daubechies Wavelet Coefficient Histogram (DWCH) that concurrently captures local and global music signal information was also utilized for MGC tasks [37]. Moreover, three multi-linear subspace techniques were compared for processing large data tensors and deriving compact feature vectors [33]. Particularly, the research found that the algorithm based on the time domain was rapid, while that based on the frequency domain was accurate [35].

Furthermore, deep neural networks featuring multiple layers have exhibited performance improvements over traditional classifier techniques [56]. Specifically, these networks include CNN [57], recurrent neural network (RNN) [58], and the combined convolutional recurrent neural network (CRNN) or CNN-long short-term memory (CNN-LSTM) [59]. Additionally, the bidirectional recurrent neural network (BRNN) has also been investigated [27], with a model employing parallelized CNN attention, achieving a classification accuracy of 92.7%. Furthermore, long short-term memory (LSTM) was explored using time and frequency domain features [28]. Precisely, the highest achieved accuracy was 0.989 using the SVM classifier with the combined feature in an experiment with LSTM. Meanwhile, recent advancements in the natural language processing field inspired the use of a transformer classifier, which yielded significant results [40]. Specifically, the authors designed a multi-head attention mechanism and a feed-forward layer in the encoder. Despite RNN's intrinsic time series information modeling capability from music clips, recent studies have focused on CNN for MIR tasks. This is attributed to CNN's high classification

TABLE 1. Summary of existing MGC methods.

Ref.	# of employed Datasets	Dataset Names	Used Features	Baseline Model
[21]	1	(Manual)	Scalogram	CNN
[22]	1	GTZAN	MFCC	kNN & DT
[23]	1	GTZAN	4 Temporal Features & 4 Spectral Features & 3 Cepstral Features	kNN & SVM & DNN
[24]	1	GTZAN	MFCC & FFT	kNN & SVM & DT & RF & Gradient Boosting
[25]	1	Spotify	FC1 & FC2 & FC3 & FC4	SVM-RBF
[17]	1	GTZAN	2 Time Domain Features & 4 Frequency Domain Features	CNN
[27]	2	GTZAN & Ballroom-Extended	STFT	BRNN
[28]	1	GTZAN	ZCR & MFCC	LSTM
[29]	1	(Manual)	MFCC	CNN
[30]	1	GTZAN	MLS	ResNet-18
[31]	1	FMA	Pitch & Pulse Clarity & Tempo & Key & Scale	CNN
[32]	1	MIREX 2015	MLS	CNN & Parallel CNN
[20]	2	GTZAN & CAL500	14 Timbre Features & 3 Temporal Features	kNN & GMM & SVM & AdaBoost & DT & NC & SRC
[33]	2	GTZAN & ISMIR04	Spectro-Temporal Features	SVM
[34]	1	Last.fm	Timbral Features & Rhythmic Features & Tonal Features & Temporal Features	SVM & Logistic Regression & RF
[35]	-	-	3 Time-Domain Features & 7 Frequency-Domain Features	-
[36]	1	(Manual)	Instruments & M/K Changes & Instrument Classes & Notes Extension & Melodic Intervals	NB
[37]	2	(Manual)	Timbral Textural Feature & Rhythmic Content Features & Pitch Content Features	SVM & GMM & MPSVM & LDA & kNN
[38]	2	MagnaTagATune & Million Song Dataset	-	CNN
[39]	1	GTZAN	MLS	CNN
[40]	1	GTZAN	MLS	RNN
[41]	1	GTZAN	Visual Features & Audio Features	CNN
[42]	2	GTZAN & ISMIR04	Spectral Features & Cepstral Features	CNN
[43]	3	GTZAN & Ballroom & Ballroom-Extended	MLS	CNN
[44]	1	ISMIR04	MFCC-EMD & Fluctuation Pattern & Spectrum Histogram	Regularized Least-Squares
[45]	2	(Manual)	Harmonic & Instrumental	NB
[46]	3	GTZAN & ISMIR04 & Ballroom-Extended	Constant-Q & Harmonic & MLS & Percussive & Scatter Transform	CNN
[47]	3	GTZAN & ISMIR04 & HOMBURG	5 Auditory Images & 5 Spectral Features & 5 Acoustic Features	CNN
[48]	2	GTZAN & Ballroom-Extended	STFT	CRNN
[87]	2	GTZAN & Ballroom-Extended	MLS	CNN
[88]	1	GTZAN	Waveform	CNN
[89]	1	GTZAN	MLS	CNN
[90]	3	GTZAN & FMA & MTAT	Constant-Q	CNN
[91]	1	GTZAN	STFT & MLS & Constant-Q	CNN

accuracy when treating time-frequency information as the input image.

Particularly, the efficacy of a four-layer neural network model with extracted musical features from the GTZAN dataset was examined for the MGC task [17]. The results were compared with neural networks, SVMs, multi-layer perceptrons, and DTs using MFCCs, achieving a classification accuracy of 91%. In addition, the authors also experimented with ImageNet's pre-trained CNN model using three musical datasets: ESC-50 [60], UrbanSound8k [61], and GTZAN. Moreover, transfer learning was utilized to classify music genres [38]. The models pre-trained using two large-scale datasets, which are the Million Song and the MagnaTagATune datasets, achieved a classification accuracy of 0.88, although they struggled to classify pop and R&B. Further, in evaluating the effectiveness of the middle-level learning feature interaction method, visual and audio features were processed through visual and audio feature extraction modules, respectively [41]. More specifically, the combination of two visual and nine audio features boosted the model accuracy while decreasing training speed by 60% compared to the equivalent model. Besides, some innovative features were proposed based on long-term modulation spectral analysis [42]. In addition, an information fusion approach was developed, integrating feature-level and decision-level fusion.

Meanwhile, the design of a specialized network for music genre identification was proposed in [43]. Specifically, the model aimed to exploit the low-level information of Mel spectrograms for classification decision-making. More particularly, a distance-based MGC algorithm was created to adopt an information fusion framework [44]. Precisely, the fusion method improved the accuracy rate by 2% to 15% compared to using single ones. A novel combination model that fuses harmonic and instrumental information was also proposed [45], exhibiting that jointly processing both pieces of information improves accuracy. A multi-level feature coding network using a CNN with self-attention and NetVLAD learned high-level features for each low-level feature [46]. The proposed model achieved a high classification accuracy of 96.50% on the GTZAN dataset. A res-gated convolutional structure with an attention mechanism was also applied to the MGC task, and the model achieved a classification accuracy of 96.8% on the GTZAN dataset [87]. Parallel attention was applied to the CNN model to extract multiple features from the MLS features [89]. A new pre-training method was also proposed that applied a swin transformer to learn meaningful music representation of massive unlabeled music data [90]. A neural model reprogramming was employed as part of a transfer learning approach, in which the model was initially trained on a large-scale acoustic dataset and subsequently fine-tuned using the smaller GTZAN dataset [88]. Additionally, a locally activated gated network was employed to capture the different characteristics of music genres [91].

Furthermore, inspired by the auditory system and spectrogram features, music genres were classified through the late fusion method, combining auditory, spectral, and acoustic features [47]. Specifically, this proposed model achieved a higher classification accuracy than many state-of-the-art classification methods. Another study for mobile devices was conducted [48], where an STFT image passed through two separate blocks of convolution and bi-recurrent operations. In particular, the two blocks produced the same-sized output combined for genre classification. Table 1 summarizes the related works on the MGC task.

While CNN exhibited promising results in MGC tasks, it has also been applied to tasks such as music mood and instrument classifications. For instance, a study on music mood classification using audio and lyric information suggested that song mood identification can be enhanced not only through musical features but also through lyrics information [34]. A model for mood, genres, and composers detection, using an integrated approach of MFCCs and CNN, was conducted [32]. Specifically, this model captures temporal and timbre information, respectively, using two different methods — sequential and parallel structures, and then merges them. For listeners unfamiliar with the genre, a study classifying them using easily recognizable moods such as happy, angry, and sad was conducted [31]. The proposed CNN model achieved an average classification accuracy of 82% and a peak accuracy of 86% for the "happy" label. Meanwhile, in the music instrument classification domain, a study that adopts time-frequency localization features proposed a CNN model [21], where the continuous wavelet transform of the audio signals was realized through the convolutional layer of the CNN model.

While recent CNN-based methods [17], [31], [41], [46], [47] have exhibited state-of-the-art performance on MGC datasets, all except one [47] employ an early fusion strategy [66], wherein all features are combined and fed into the models as input data. Meanwhile, in the studies by [42] and [47], a different approach is used, where input data is separated into visual, spectral, and acoustic features. These features are then individually fed into the model to extract unique information, which is subsequently concatenated in the middle of the model, a process known as the late fusion strategy. However, given that spectral features may contain redundant information from each other, it is plausible that the late fusion strategy might be equally effective when applied to multiple spectral features. Therefore, this study investigates the efficacy of the late fusion strategy in handling multiple spectral features for the MGC task. Furthermore, as a CNN model is a well-established architecture, we employ it as a baseline model to compare the performances with conventional CNNs on the various MGC datasets. The performances of the baseline model can serve as a reference point for benchmarking and comparing the performances of newer neural architectures in future studies.

III. BASELINE MODEL: LATE FUSION CNN

A. PRELIMINARIES

STFT, MLS, and MFCC are the most commonly used spectral features for MGC [22], [24], [27], [30]. Specifically, the STFT is a Fourier-related transformation utilized to ascertain the sine wave frequency and phase content of specific local sections of a time-varying signal. Particularly, it furnishes time-localized frequency information for situations where signal frequency components vary over time. This is advantageous when understanding the flow of frequency within a desired time period as the signal frequency changes. From the result of the STFT, the MLS reflects the human characteristic of being more sensitive to changes in low frequencies than in high frequencies. The spectral features from STFT are scaled using a mel-scale that transforms high-frequency voice signals into low-frequency ones, to which humans are more sensitive. Furthermore, MFCC can be derived by performing a single discrete cosine transform (DCT) operation [65] on the extracted MLS features. Specifically, the DCT operation compresses the filter bank obtained from the MLS to reduce frequency correlations. Because the MLS and MFCC can provide another perspective on the spectral features, they are often used in conjunction with STFT.

Further, to integrate those spectral features for a single CNN model, feature-level fusion has exhibited effectiveness in fusing multiple spectral features [66]. In particular, there are two primary strategies for fusing input data: early fusion and late fusion. Early fusion integrates separate spectral features into a unified data source, which then undergoes the entire training process. Despite its simplicity, this approach struggles to extract the unique characteristics of each spectral feature. Specifically, when using STFT, MLS, and MFCC, the early fusion strategy may produce redundant information, given that the spectral features by MLS originate from STFT and those by MFCC originate from MLS. Consequently, the benefits of using multiple spectral features might disappear due to the overwhelming redundancy, which could potentially mislead the model.

Conversely, late fusion merges data one step before the classification stage. Specifically, input data with multi-modal features are fed into the CNN model separately, and the output from each modality is concatenated. Particularly, this method is effective in extracting the unique characteristics of each modality. Furthermore, when data sources vary in sampling rates, data dimensions, and units of measurement, it is a more straightforward and flexible approach than the early fusion method. Therefore, we construct a baseline model with a late fusion strategy for the spectral features of STFT, MLS, and MFCC and make comparisons with conventional CNN-based methods for MGC. In this paper, we specifically applied the late fusion strategy to a CNN model and compared its performance with that of conventional CNN models designed for the MGC domain.

B. LATE FUSION CNN FOR MGC

The late fusion CNN model devised in this paper is motivated by DenseNet [14] since it is robust against the gradient loss problem and adopts an effect of feature reuse, which can contribute to enhanced classification performance. In particular, DenseNet connects the feature map of the first layer to the feature map of the last layer to prevent information loss and provides a normalization effect by linking the feature maps of various layers. Specifically, the lth layer uses all of the preceding output feature maps 0,1,2…,l-1 as input feature maps, with the result being x_l. The expression of x_l is as follows:

$$x_l = H_l\left([x_0, x_1, x_2, \ldots, x_{l-1}]\right) \tag{1}$$

In Equation (1), function $H_l(\cdot)$ corresponds to batch normalization, rectified linear unit (ReLU) [71], and a 3 × 3 convolution layer.

Figure 1 illustrates the architecture of the devised model. Initially, multiple spectral features by STFT, MLS, and MFCC were extracted from the music clip. Each input data from different spectral features is resized into (256, 1296) for late fusion at the feature level. Each input data passes through a 7 × 7 group convolution layer and a max pooling layer, where the width and height of the feature maps are halved. The reduced feature maps are then forwarded to the dense block.

Dense layers in the dense blocks follow the structure of batch normalization, ReLU, 1 × 1 convolution layer, Batch Normalization, ReLU, 3 × 3 convolution layer. Precisely, four dense blocks consist of 6, 12, 24, and 16 dense layers, respectively. The ReLU function forwards the input value to the output value for numbers greater than zero and outputs zero for numbers less than zero. Particularly, a 1 × 1 convolution layer, referred to as the bottleneck layer, reduces the channel number of the feature map. Additionally, a 3 × 3 convolution layer captures important characteristics of the feature maps, and a dropout layer is added after the 3×3 convolution layer of the dense layer. A detailed structure of the dense layer is shown in Figure 2, where w, h, k_i represent the width, height, and depth of a corresponding feature map, respectively. In this study, k_0 begins from 60 and increases the total depth of the feature maps by 32 as k is concatenated.

The next component to be described is the transition layer, which reduces the horizontal and vertical dimensions of the feature map as well as the number of feature maps. Except for the last dense block, it connects to the back of the dense block and comprises batch normalization, ReLU, 1 × 1 convolution layer, and 2 × 2 adaptive-average-pooling. Here, θ, which reduces the number of feature maps via 1 × 1 convolution layer and indicates the degree of reduction, is set to 0.5 in this study. The number of feature maps is halved, and the width and height of the feature map are also halved through the 2×2 average pooling layer after passing through the transition layer. Hyperparameter θ [0, 1] aligns the output channel with

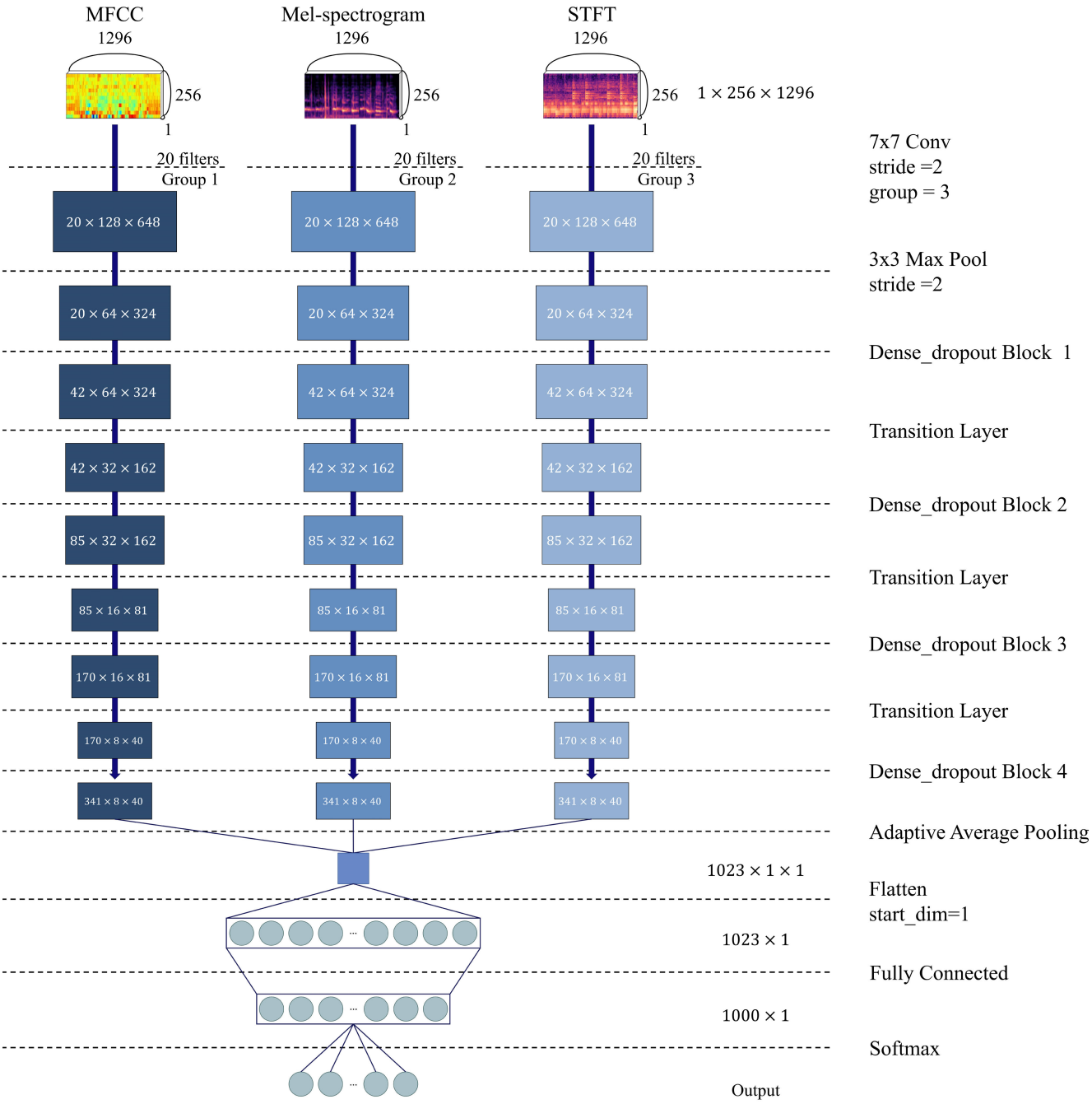


FIGURE 1. Flowchart of the devised late fusion CNN.

$m \times \theta$ as a factor adjusting the number of output feature maps by the transition layer when m feature maps are input into the dense block. If $\theta = 1$, the same number of feature maps is taken. Lastly, adaptive-average-pooling was adopted as shown in Figure 3.

After each input data passes through the four dense blocks and three transition layers, the feature maps are concatenated to fuse the information from multiple spectral features. The concatenated features then pass through a single fully connected layer. In particular, we implement only one fully connected layer with the SoftMax function to classify genres to avoid overfitting from multiple fully connected layers. For the loss function, cross-entropy [74] is used, a common metric for evaluating performance in classification models in machine learning. Assuming a predictive model trying to anticipate the q distribution without knowledge of the actual distribution q, the distribution obtained through predictive modeling is $p(x)$. When creating a p distribution that predicts

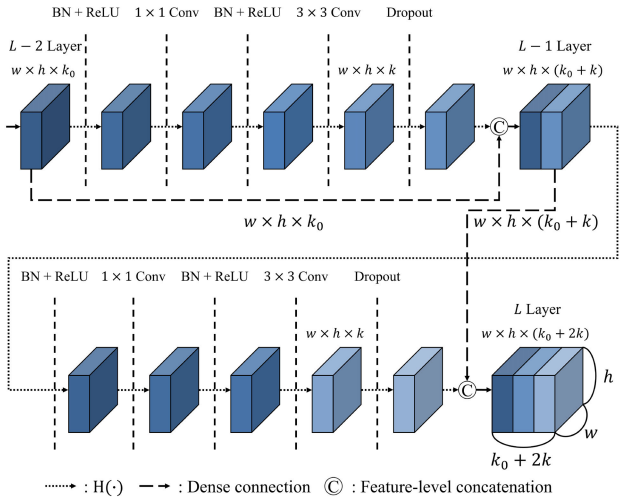


FIGURE 2. Schematic diagram of the dense layer inside the dense block where w, h, and k_l represent the width, height, and depth of a corresponding feature map. The last concatenated Lth layer is 6, 12, 24, and 16 in the dense blocks 1, 2, 3, and 4, respectively.

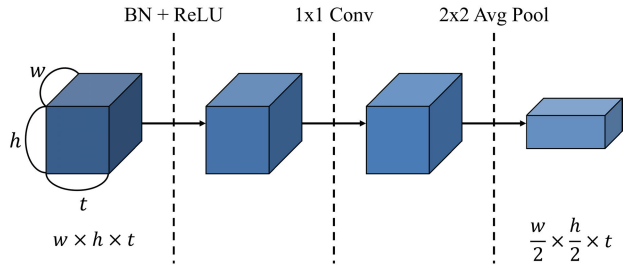


FIGURE 3. Schematic diagram of the transition layer.

the actual distribution q, cross-entropy is defined as follows:

$$H_p(q) = -\sum_{c=1}^{C} q(y_c) \log(p(y_c)) \tag{2}$$

A smaller cross-entropy value indicates the two probability distributions are closer. Minimizing entropy thus allows us to reduce the difference between actual and predicted values. Finally, the Adam optimizer [75] is used with a default learning rate of 0.001.

The application of the late fusion strategy in the field of music genre classification can be driven by its inherent compatibility with the complexities of this domain. In music genre classification, the abundance of audio features extracted from music clips results in numerous potential feature combinations for classification models, as described in Table 1. The late fusion strategy is chosen for its innate simplicity and adaptability, facilitating the management of diverse feature combinations and thus addressing the inherent variability in music data.

Furthermore, as the field of music genre classification advances, the incorporation of advanced deep learning techniques, such as attention mechanisms, becomes increasingly relevant. The structural nature of the late fusion strategy aligns with this requirement, granting users the flexibility to selectively integrate advanced techniques that align most effectively with their specific classification needs. Consequently, the late fusion strategy not only simplifies the handling of diverse feature combinations but also enables the effective integration of contemporary deep learning methodologies, positioning it as an ideal choice for music genre classification.

IV. EXPERIMENTAL RESULTS

A. EXPERIMENTAL SETTINGS

In this section, we contrast the late fusion CNN model as a baseline model with existing MGC techniques across 12 music datasets. The Ballroom dataset, comprising 698 songs from eight genres of Ballroom dancing, each with a maximum duration of 30 seconds, forms the initial benchmark [76]. The Ballroom-Extended dataset enhances the original Ballroom set with superior audio quality, a six-fold increase in tracks, the addition of five new rhythm classes, and a variety of repetitive annotations [77]. The FMA-SMALL dataset, encompassing 8,000 tracks, is a subset selected from Free Music Archive (FMA) [78]. Additionally, we utilize the GTZAN dataset, a popular choice for music classification, with 1,000 audio tracks each 30 seconds long, and the HOMBURG dataset with 1,886 songs[1] [79]. The MICM dataset for the classification of seven Dastgahs of Iranian classical music is referenced from [1]. The Seyerlehner:Unique dataset features 3,115 30-second music clips from popular songs across 14 unbalanced genres [80]. The EmoMusic-G dataset, assembled from 1,000 FMA songs for music emotion recognition, and the Emotify-G dataset, including 400 60-second song clips based on the Geneva Emotional Music Scales [81], each allows up to three tags per song. The final datasets, GiantStepsKey-G and GMD-G, encompass 604 electronic dance music pieces[2] for key estimation and 1,150 MIDI files totaling 13.6 hours and over 22,000 measures of drumming, respectively. Although the EmoMusic-G, Emotify, GiantStepsKey, and GMD datasets initially pertain to various domains, we convert them to genre domains using the provided metadata. For further details on each dataset, refer to [82].

Table 2 provides the summary statistics for all datasets, including those previously modified for the experiments. Specifically, each column represents the dataset's name, the number of patterns $|W|$, the average length of music clips in the collection, the domain as suggested by the original data providers, and the domain we applied in this study. Additionally, it enumerates the number of tags $|T|$ and the average number of patterns per genre. Upon assembling the datasets, all music files were converted into the *.wav* format. The Python library *librosa* was then used to extract the necessary features from each music file for the experiments.

[1]https://www.garageband.com
[2]https://www.beatport.com

We first conducted the STFT extraction, where the length of the windowed signal after zero-padding (*n_fft*) was set to 4,096, and each frame of audio was matched to *n_fft*. The number of audio samples between successive STFT columns (*hop_length*) was set to 512. During the MLS extraction, we used the default settings, except that we adjusted the sampling rate to 44,100 and performed normalization. For the MFCC extraction, the settings were the same as those for STFT, with the only difference being that all negative numbers were converted into positive ones. Subsequently, we zero-padded the music clips to equalize their lengths. Clips of varying original lengths were padded to 1296 units, equating to 30 seconds, to align with the length of the longest music clip in the Ballroom dataset.

The baseline model was compared to the latest conventional CNN-based methods for the MGC: ResNet50_trust [67], Pelchat and Gelowitz [68], Cheng et al. [69], and Shah et al. [70]. We detail each method as follows:

- **ResNet50_trust** is a CNN-based model utilizing the ResNet50 architecture paired with a balanced trusted loss function. The model is trained using features extracted from STFT.
- **Pelchat and Gelowitz** represent a CNN-based model comprising convolutional layers, max-pooling layers, and fully connected layers. This model is trained using spectrogram slices.
- **Cheng et al.** present a CNN-based model composed of five convolutional layers, each containing a convolutional operation, ReLU activation, max-pooling, and dropout. The model is trained using features derived from the mel-spectrogram.
- **Shah et al.** put forward a CNN-based model resembling the Cheng et al. model, trained using features from the mel-spectrogram.

For each method, the parameters were set equally for the fairness of comparison. Specifically, the initial input size of each music clip was (256, 1296), which was for the 30-second long audio, and the batch size, learning rate, and number of epochs were set to 16, 0.001, and 60, respectively. Aside from the devised baseline model, all models were trained using concatenated features from STFT, MLS, and MFCC. Additionally, hold-out cross-validation was conducted for each experiment on a given dataset; 80% of the patterns were randomly selected as training sets for model training. The remaining 20% of the patterns were used as test sets to obtain classification performance. The experiments were repeated ten times, and the average accuracy and standard deviation were calculated.

B. COMPARISON RESULTS

In Table 3, we have summarized the results of experiments conducted with five different methods, presenting each method's accuracy on different datasets. Interestingly, the ▼** and ▼* symbols indicate that the late fusion CNN (baseline) model is statistically superior to the comparison method at the significance level 99% and 95%, respectively, based on the paired t-test. The baseline model consistently outperforms the comparative methods across all datasets. Specifically, the late fusion CNN model exhibited the highest accuracy in every dataset, with an outstanding performance on the Ballroom-Extended dataset. Here, the baseline model achieved 85.49% accuracy with a standard deviation of 1.06, significantly outperforming the second-best method, Cheng et al., by 22.61 percentage points. The largest performance gap was observed on the same dataset, with the baseline model outperforming the least accurate method, Pelchat and Gelowitz's, by a striking 40.81 percentage points.

It is noteworthy that the baseline model exhibits the least variability, as indicated by relatively low standard deviations, suggesting robustness across different experimental runs. Particularly, the consistently high performance of the baseline model across various datasets indicates its adaptability to different music genres and recording qualities. A more generalized view of the method's performance is presented at the bottom of the table. The baseline model achieved a win in every dataset, resulting in a perfect win/tie/lose score of 48/0/0. The average rank of 1.00 for the baseline model further solidifies its superior performance over other methods. In addition, the baseline model exhibited statistical superiority to all comparison methods based on the paired t-test at the significance level 99%, except for the experiment of ResNet50_trust conducted using the GTZAN dataset.

Two statistical tests, Friedman and Bonferroni–Dunn tests, were employed to further demonstrate the superiority of the baseline model with quantitative evaluation. To analyze the performance of various MGC methods, different methods on different datasets were compared using the Friedman test, a widely used statistical test. Given k methods and N datasets, r_i^j represents the rank of the jth method for the ith dataset (mean ranks are shared in case of ties) and $R_j = \frac{1}{N}\sum_{i=1}^{N} r_i^j$ denotes the mean ranking for the jth method. The Friedman statistic F_F is given as

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}$$

where the χ_F^2 is defined as

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right]$$

Under the null hypothesis $H_0 : R_1 = \ldots = R_k$, statistic F_F is distributed according to the F-distribution with numerator degrees of freedom k-1 and denominator degrees of freedom (k-1)(N-1). Suppose F_F is greater than the critical value defined as the quantile of order $1-\alpha$ of the F-distribution with $k-1$ and (k-1)(N-1) degrees of freedom, where α is the assumed significance level. In that case, the null hypothesis that the performance between the comparison methods is the same at each evaluation scale is rejected. In the case of the rejection of the H_0, a specific post-hoc

TABLE 2. The standard characteristics and statistics of employed datasets.

Dataset	$\|W\|$	Average Length (sec.)	Suggested Domain	Used Domain	$\|T\|$	Clips per Genre
Ballroom	698	30	Genre	Genre	8	87.25
Ballroom-Extended	4,180	30	Genre	Genre	13	321.54
EmoMusic-G	1,000	45	Emotion	Genre	8	125.00
Emotify-G	400	43	Emotion	Genre	4	100.00
FMA-SMALL	7,996	30	Genre	Genre	8	999.50
GiantStepsKey-G	604	80	Key	Genre	23	24.91
GMD-G	1,042	300	Groove	Genre	18	57.89
GTZAN	1,000	30	Genre	Genre	10	100.00
HOMBURG	1,886	10	Genre	Genre	9	209.56
ISMIR04	2,187	120	Genre	Genre	6	121.17
MICM	1,137	360	Genre	Genre	7	162.43
Seyerlehner:Unique	3,115	30	Genre	Genre	14	311.50

TABLE 3. Experimental results of five models in terms of accuracy (▼** and ▼* indicate that the corresponding method is significantly worse than the late fusion CNN (Baseline) model based on paired t-test at 99% and 95% significance level, respectively).

Dataset	Baseline	ResNet50_trust		Pelchat and Gelowitz		Cheng et al.		Shah et al.	
Ballroom	**58.29**% ± 1.91	45.71% ± 2.86	▼**	42.43% ± 2.29	▼**	48.07% ± 4.79	▼**	47.71% ± 4.58	▼**
Ballroom-Extended	**85.49**% ± 1.06	61.82% ± 3.79	▼**	44.68% ± 1.30	▼**	62.88% ± 4.70	▼**	54.38% ± 5.33	▼**
EmoMusic-G	**36.04**% ± 0.95	27.72% ± 3.50	▼**	28.99% ± 1.78	▼**	32.95% ± 2.92	▼**	31.34% ± 3.66	▼**
Emotify-G	**74.88**% ± 2.53	62.78% ± 2.70	▼**	67.50% ± 4.17	▼**	66.00% ± 3.32	▼**	65.88% ± 5.62	▼**
FMA-SMALL	**57.33**% ± 0.94	47.03% ± 1.31	▼**	42.19% ± 1.16	▼**	46.96% ± 1.62	▼**	40.88% ± 1.73	▼**
GiantStepsKey-G	**37.27**% ± 1.80	28.43% ± 3.15	▼**	25.21% ± 2.53	▼**	27.19% ± 2.90	▼**	24.30% ± 2.11	▼**
GMD-G	**67.13**% ± 1.13	61.91% ± 3.04	▼**	63.44% ± 1.71	▼**	39.38% ± 3.36	▼**	35.36% ± 6.35	▼**
GTZAN	**82.00**% ± 2.24	77.50% ± 3.81	▼*	63.35% ± 2.08	▼**	75.10% ± 3.84	▼**	72.50% ± 2.13	▼**
HOMBURG	**58.76**% ± 1.15	53.49% ± 1.59	▼**	51.01% ± 1.51	▼**	52.54% ± 2.55	▼**	48.15% ± 2.14	▼**
ISMIR04	**80.59**% ± 1.41	70.96% ± 2.93	▼**	67.88% ± 1.70	▼**	71.32% ± 1.37	▼**	67.67% ± 1.99	▼**
MICM	**43.25**% ± 2.32	37.81% ± 3.32	▼**	39.30% ± 2.21	▼**	38.42% ± 3.16	▼**	39.34% ± 2.28	▼**
Seyerlehner:Unique	**72.57**% ± 1.19	63.62% ± 2.56	▼**	60.72% ± 2.73	▼**	63.50% ± 2.79	▼**	61.62% ± 1.70	▼**
Win/Tie/Lose	**48/0/0**	16/16/16		3/15/30		15/19/14		5/16/27	
Avg. Rank	**1.00**	3.17		3.92		2.83		4.08	

test should be conducted to analyze the relative performance between the comparison methods [83]. As indicated in Table 4, the value of the Friedman statistic ($F_F = 29.133$) exceeds the critical value of 2.583. This suggests that the null hypothesis, assuming equal performance of all methods, should be rejected. Following this, the Bonferroni–Dunn test was performed to compare the baseline model against the other methods. Particularly, the Bonferroni–Dunn post-hoc test allows us to decide whether the baseline model performs better than some other competitive method [84]. Specifically, the difference in the average rankings between the baseline model and one of the competitive methods is compared with the following critical difference (CD):

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}.$$

In the Bonferroni–Dunn test, the performance of the baseline model is deemed statistically equivalent to that of a comparative method if the average rankings for all datasets fall within the CD. This implies that if the disparity in terms of average rank between the baseline model and the comparative method exceeds the CD, we can confirm that the performance

TABLE 4. Summary of the Friedman statistic $F_F(k = 5, N = 12)$ and Critical Value in terms of the Classification Accuracy measure.

Evaluation Measure	F_F	Critical Value ($\alpha = 0.05$)
Accuracy	29.133	2.583

is significantly different, thus allowing us to discern the superior method. With $N = 12$ datasets and $k = 5$ methods under comparison, the CD at a significance level of $\alpha = 0.05$ is found to be 1.612, given that $q_\alpha = 2.498$. Figure 4 visually represents the results of the Bonferroni–Dunn test. The baseline model, which achieved the highest average rank, is depicted on the rightmost side of the figure. The CD line, which designates the range of CD, is positioned above the main line. No other comparison method falls within the CD range, allowing us to conclude that the accuracy of the baseline model significantly surpasses that of the competing methods, thereby corroborating its superiority. Additionally, this result suggests that the late fusion strategy's ability to extract unique information from each spectral feature appears more effective than the early fusion strategy.

TABLE 5. Classification accuracies and standard deviations of the baseline model for different feature combinations on all the datasets.

Dataset	STFS, MLS, MFCC	STFT	MLS	MFCC	STFT, MLS	MLS, MFCC	MFCC, STFT
Ballroom	**58.29**% ± 1.91	56.43% ± 4.12	53.36% ± 2.10	46.71% ± 3.90	57.36% ± 3.63	57.71% ± 3.12	57.71% ± 3.08
Ballroom-Extended	**85.49**% ± 1.06	83.25% ± 1.16	82.70% ± 1.19	78.29% ± 1.97	83.55% ± 0.99	85.48% ± 0.67	83.84% ± 1.89
EmoMusic-G	**36.91**% ± 0.95	34.03% ± 2.22	31.21% ± 3.04	33.83% ± 2.83	33.29% ± 2.01	34.30% ± 3.43	33.83% ± 2.31
Emotify-G	**74.88**% ± 2.53	73.63% ± 3.03	70.88% ± 2.57	68.13% ± 2.96	71.88% ± 4.80	73.13% ± 2.72	70.63% ± 3.74
FMA-SMALL	**57.33**% ± 0.94	56.00% ± 1.01	54.78% ± 0.90	45.81% ± 0.98	45.55% ± 0.88	45.11% ± 1.09	45.51% ± 0.63
GiantStepsKey-G	**37.27**% ± 1.80	35.12% ± 2.92	32.15% ± 2.87	29.50% ± 3.57	35.21% ± 3.60	33.72% ± 3.32	35.29% ± 2.53
GMD-G	**67.13**% ± 1.13	59.90% ± 2.24	64.11% ± 2.90	60.67% ± 2.37	63.78% ± 1.89	64.11% ± 2.24	62.25% ± 2.63
GTZAN	**82.00**% ± 2.24	74.75% ± 2.19	80.10% ± 1.26	60.20% ± 1.67	82.00% ± 1.73	79.15% ± 2.65	78.45% ± 2.70
HOMBURG	**58.76**% ± 1.15	57.22% ± 1.16	56.64% ± 2.31	46.72% ± 1.72	58.47% ± 1.47	55.45% ± 1.93	56.48% ± 1.90
ISMIR04	**80.59**% ± 1.41	75.07% ± 2.23	75.09% ± 1.43	74.63% ± 1.54	77.69% ± 1.00	78.97% ± 1.89	79.41% ± 2.23
MICM	**43.25**% ± 2.32	41.18% ± 2.21	40.22% ± 2.61	40.88% ± 3.19	41.36% ± 1.83	39.65% ± 2.85	41.32% ± 2.03
Seyerlehner:Unique	**72.57**% ± 1.19	71.36% ± 1.70	71.56% ± 1.10	66.57% ± 1.09	72.33% ± 1.14	72.34% ± 1.56	72.41% ± 1.37
Avg. Rank	**1.00**	4.42	4.83	6.25	3.58	3.83	3.83

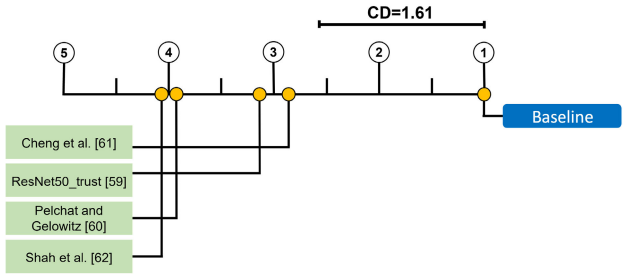


FIGURE 4. Bonferroni–Dunn test result of the five comparison methods in terms of accuracy.

TABLE 6. Comparison of the experimental results of applying early and late fusion to the baseline model in terms of accuracy (▼ indicates that the early fusion is significantly worse than the late fusion (Baseline) based on paired t-test at 95% significance level).

Dataset	Baseline	Early fusion	
Ballroom	**58.29**% ± 1.91	54.57% ± 5.22	▼
Ballroom-Extended	**85.49**% ± 1.06	82.94% ± 1.42	▼
EmoMusic-G	**36.04**% ± 0.95	35.97% ± 2.29	
Emotify-G	**74.88**% ± 2.53	72.88% ± 3.87	
FMA-SMALL	**57.33**% ± 0.94	55.63% ± 1.05	▼
GiantStepsKey-G	**37.27**% ± 1.80	32.81% ± 2.76	▼
GMD-G	**67.13**% ± 1.76	62.11% ± 2.69	▼
GTZAN	**82.00**% ± 2.24	81.70% ± 1.53	
HOMBURG	**58.76**% ± 1.15	55.37% ± 2.85	▼
ISMIR04	**80.59**% ± 1.41	79.38% ± 1.07	▼
MICM	**43.25**% ± 2.32	40.61% ± 2.06	▼
Seyerlehner:Unique	**72.57**% ± 1.19	70.48% ± 0.77	▼
Avg. Rank	**1.00**	2.00	

C. IN-DPETH ANALYSIS OF THE BASELINE MODEL

In particular, the architecture of the baseline model is designed to leverage multiple spectral features by employing a multi-head architecture and late fusion strategy. To validate this strategy, two additional experiments were conducted. The first experiment compared the impact of varying the number of spectral features utilized in the baseline model. This experiment aimed to determine whether the baseline model could effectively utilize all spectral features concurrently. The classification accuracy of the baseline model with different combinations of spectral features is shown in Table 5. Interestingly, the results suggest that the baseline model achieves its best performance when all spectral features are simultaneously employed. It is also important to note that the performance of the baseline model is at its lowest when only one of the spectral features is used across all datasets. Nevertheless, an increase in the number of spectral features used leads to a corresponding enhancement in the performance of the baseline model.

Furthermore, we conducted a comparison experiment to verify the effectiveness of the late fusion strategy. Specifically, we compared the performance of the baseline model with the early fusion strategy. For the early fusion strategy, we concatenated all spectral features and fed them into the baseline model as input data. Table 6 illustrates the classification accuracy of the baseline model with early fusion and late fusion strategies. The ▼ indicates that the performance of the early fusion strategy is significantly worse than that of the late fusion strategy based on paired t-test at 95% significance level.

Observing the results from various datasets, it is evident that the late fusion strategy consistently outperforms early fusion in terms of accuracy. In the case of the Ballroom dataset, the baseline model exhibits a statistically significant improvement with an accuracy of 58.29% over the 54.57% achieved by the early fusion. This superiority of the late fusion strategy also extends to the Ballroom-Extended, FMA-SMALL, GiantStepsKey-G, GMD-G, HOMBURG, ISMIR04, MICM, and Seyerlehner:Unique datasets. For the EmoMusic-G and GTZAN datasets, although the baseline model presents a slightly better performance, the difference does not achieve statistical significance based on the paired t-test. Meanwhile, on the Emotify-G dataset, the baseline model exhibits an improvement in accuracy, but the lack of the ▼ symbol suggests that this improvement is not statistically significant. In terms of average rank, the baseline model achieves a superior score of 1.00, indicating its consistent effectiveness across different datasets. The early fusion method, by contrast, has an average rank of 2.00. Overall, these results serve to validate the superiority of

the late fusion strategy over the early fusion, indicating the potential of the late fusion strategy for music genre classification across a broad range of datasets.

V. CONCLUSION

Music genre classification is garnering increasing interest in both practical applications and academic research. Multiple spectral features have been utilized for music genre classification, yet there are relatively few classification methods based on neural networks that can handle multiple input features concurrently. In this paper, we conducted a review of existing MGC methods and devised a late fusion CNN to set the baseline performance of conventional MGC methods.

The baseline model employs a late fusion strategy to combine features extracted from multiple spectral features into a CNN model. Through experiments and statistical tests using 12 datasets, the superior performance of the late fusion strategy for the MGC task was observed. In the future, we plan to apply the late fusion strategy to other MGC tasks, such as emotion and mood classification.

ACKNOWLEDGMENT

An earlier version of this paper was presented at the 2022 International Conference on Consumer Electronics Asia (ICCE-Asia), Yeosu, South Korea, October 2022 [DOI: 10.1109/ICCE-Asia57006.2022.9954732].

REFERENCES

[1] S.-H. Cho, Y. Park, and J. Lee, "Effective music genre classification using late fusion convolutional neural network with multiple spectral features," in *Proc. IEEE Int. Conf. Consum. Electron.-Asia (ICCE-Asia)*, Yeosu, South Korea, Oct. 2022, pp. 1–4.

[2] K. Dave and V. Varma, "Music information retrieval: Recent developments and applications," *Found. Trends Inf. Retr.*, vol. 8, nos. 4–5, pp. 263–418, 2014.

[3] B. Ferwerda, M. Tkalcic, and M. Schedl, "Personality traits and music genre preferences: How music taste varies over age groups," in *Proc. 1st Workshop Temporal Reasoning Recommender Syst. (RecTemp) 11th ACM Conf. Recommender Syst.*, Como, Italy, Aug. 2017, pp. 16–20.

[4] N. Ndou, R. Ajoodha, and A. Jadhav, "Music genre classification: A review of deep-learning and traditional machine-learning approaches," in *Proc. IEEE Int. IoT, Electron. Mechatronics Conf. (IEMTRONICS)*, Apr. 2021, pp. 239–247.

[5] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Proc. IEEE 25th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Boston, MA, USA, Sep. 2015, pp. 1–6.

[6] C. Joder, S. Essid, and G. Richard, "Temporal integration for audio classification with application to musical instrument classification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 1, pp. 174–186, Jan. 2009.

[7] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Appl. Acoust.*, vol. 158, Jan. 2020, Art. no. 107020.

[8] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-25, no. 3, pp. 235–238, Jun. 1977.

[9] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proc. Int. Symp. Music Inf. Retr.*, Oct. 2000, pp. 1–2.

[10] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3D log-mel spectrograms with deep learning network," *IEEE Access*, vol. 7, pp. 125868–125881, 2019.

[11] S. Pauws, "Musical key extraction from audio," in *Proc. 5th Int. Conf. Music Inf. Retr.*, Barcelona, Spain, Oct. 2004, pp. 1–4.

[12] A. N. Ali and M. Z. Abdullah, "One dimensional with dynamic features vector for iris classification using traditional support vector machines," *J. Theor. Appl. Inf. Technol.*, vol. 70, no. 1, pp. 1–7, 2014.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[14] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[15] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5987–5995.

[16] M. Tan and Q. Le, "EfficientNetv2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2021, pp. 10096–10106.

[17] A. Ghildiyal, K. Singh, and S. Sharma, "Music genre classification using machine learning," in *Proc. 4th Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA)*, Nov. 2020, pp. 1368–1372.

[18] I. Gallo, A. Calefati, and S. Nawaz, "Multimodal classification fusion in real-world scenarios," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 5, Kyoto, Japan, Nov. 2017, pp. 36–41.

[19] F. Castanedo, "A review of data fusion techniques," *Sci. World J.*, vol. 2013, pp. 1–19, Sep. 2013.

[20] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 303–319, Apr. 2011.

[21] A. Dutta, D. Sil, A. Chandra, and S. Palit, "CNN based musical instrument identification using time-frequency localized features," *Internet Technol. Lett.*, vol. 5, no. 1, p. e191, Jan. 2022.

[22] M. S. Rao, O. P. Kalyan, N. N. Kumar, M. T. Tabassum, and B. Srihari, "Automatic music genre classification based on linguistic frequencies using machine learning," in *Proc. Int. Conf. Recent Adv. Math. Informat. (ICRAMI)*, Tebessa, Algeria, Sep. 2021, pp. 1–5.

[23] J. Chauhan, J. Shah, E. Mundhe, and I. Jain, "Web application for machine learning based music genre classification," in *Proc. Int. Conf. Adv. Comput., Commun., Control (ICAC)*, Mumbai, India, Dec. 2021, pp. 1–6.

[24] C. Dabas, A. Agarwal, N. Gupta, V. Jain, and S. Pathak, "Machine learning evaluation for music genre classification of audio signals," *Int. J. Grid High Perform. Comput.*, vol. 12, no. 3, pp. 57–67, Jul. 2020.

[25] D. R. I. M. Setiadi, D. S. Rahardwika, E. H. Rachmawanto, C. A. Sari, A. Susanto, I. U. W. Mulyono, E. Z. Astuti, and A. Fahmi, "Effect of feature selection on the accuracy of music genre classification using SVM classifier," in *Proc. Int. Seminar Appl. Technol. Inf. Commun. (iSemantic)*, Semarang, Indonesia, Sep. 2020, pp. 7–11.

[26] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning sound representations from unlabeled video," in *Proc. Adv. Neural Inf. Process. Syst.*, Barcelona, Spain, Dec. 2016, pp. 1–9.

[27] Y. Yu, S. Luo, S. Liu, H. Qiao, Y. Liu, and L. Feng, "Deep attention based music genre classification," *Neurocomputing*, vol. 372, pp. 84–91, Jan. 2020.

[28] Y. Yi, X. Zhu, Y. Yue, and W. Wang, "Music genre classification with LSTM based on time and frequency domain features," in *Proc. IEEE 6th Int. Conf. Comput. Commun. Syst. (ICCCS)*, Chengdu, China, Apr. 2021, pp. 678–682.

[29] S. Prabavathy, V. Rathikarani, and P. Dhanalakshmi, "An enhanced musical instrument classification using deep convolutional neural network," *Int. J. Recent Technol. Eng.*, vol. 8, no. 4, pp. 8772–8774, 2019.

[30] D. Bisharad and R. H. Laskar, "Music genre recognition using residual neural networks," in *Proc. TENCON - IEEE Region 10 Conf. (TENCON)*, Kerala, India, Oct. 2019, pp. 2063–2068.

[31] F. U. Widowati, F. S. Nugroho, and G. F. Shidik, "Classification of music moods based on CNN," in *Proc. Int. Seminar Appl. Technol. Inf. Commun.*, Semarang, Indonesia, Sep. 2018, pp. 318–321.

[32] T. Lidy and A. Schindler, "Parallel convolutional neural networks for music genre and mood classification," in *Proc. 17th Int. Conf. Music Inf. Retr.*, New York, NY, USA, Aug. 2016, pp. 1–4.

[33] I. Panagakis, E. Benetos, and C. Kotropoulos, "Music genre classification: A multilinear approach," in *Proc. 9th Int. Conf. Music Inf. Retr.*, Sep. 2008, pp. 583–588.

[34] C. Laurier, J. Grivolla, and P. Herrera, "Multimodal music mood classification using audio and lyrics," in *Proc. 2008 7th Int. Conf. Mach. Learn. Appl.*, San Diego, CA, USA, Dec. 2008, pp. 688–693.

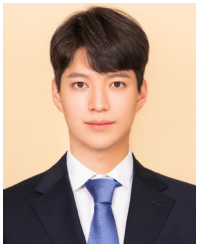
[35] A. I. Al-Shoshan, "Speech and music classification and separation: A review," *J. King Saud Univ. Eng. Sci.*, vol. 19, no. 1, pp. 95–132, 2006.

[36] R. Basili, A. Serafini, and A. Stellato, "Classification of musical genre: A machine learning approach," in *Proc. Int. Conf. Music Inf. Retr.*, Barcelona, Spain, Oct. 2004, pp. 1–4.

[37] T. Li, M. Ogihara, and Q. Li, "A comparative study on content-based music genre classification," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Informaion Retr.*, Toronto, ON, Canada, Jul. 2003, pp. 282–289.
[38] B. Liang and M. Gu, "Music genre classification using transfer learning," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Shenzhen, China, Aug. 2020, pp. 392–393.
[39] J. Mehta, D. Gandhi, G. Thakur, and P. Kanani, "Music genre classification using transfer learning on log-based MEL spectrogram," in *Proc. 5th Int. Conf. Comput. Methodologies Commun. (ICCMC)*, Erode, India, Apr. 2021, pp. 1101–1107.
[40] Y. Zhuang, Y. Chen, and J. Zheng, "Music genre classification with transformer classifier," in *Proc. 4th Int. Conf. Digit. Signal Process.*, Chengdu, China, Jun. 2020, pp. 155–159.
[41] J. Liu, C. Wang, and L. Zha, "A middle-level learning feature interaction method with deep learning for multi-feature music genre classification," *Electronics*, vol. 10, no. 18, p. 2206, Sep. 2021.
[42] C.-H. Lee, J.-L. Shih, K.-M. Yu, and H.-S. Lin, "Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features," *IEEE Trans. Multimedia*, vol. 11, no. 4, pp. 670–682, Jun. 2009.
[43] C. Liu, L. Feng, G. Liu, H. Wang, and S. Liu, "Bottom-up broadcast neural network for music genre classification," *Multimedia Tools Appl.*, vol. 80, no. 5, pp. 7313–7331, Feb. 2021.
[44] Y. Song and C. Zhang, "Content-based information fusion for semi-supervised music genre classification," *IEEE Trans. Multimedia*, vol. 10, no. 1, pp. 145–152, Jan. 2008.
[45] T. Pérez-García, C. Pérez-Sancho, and J. M. Iñesta, "Harmonic and instrumental information fusion for musical genre classification," in *Proc. 3rd Int. Workshop Mach. Learn. Music*, Firenze, Italy, Oct. 2010, pp. 49–52.
[46] W. W. Y. Ng, W. Zeng, and T. Wang, "Multi-level local feature coding fusion for music genre recognition," *IEEE Access*, vol. 8, pp. 152713–152727, 2020.
[47] X. Cai and H. Zhang, "Music genre classification based on auditory image, spectral and acoustic features," *Multimedia Syst.*, vol. 28, no. 3, pp. 779–791, Jun. 2022.
[48] R. Yang, L. Feng, H. Wang, J. Yao, and S. Luo, "Parallel recurrent convolutional neural networks-based music genre classification method for mobile devices," *IEEE Access*, vol. 8, pp. 19629–19637, 2020.
[49] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Jul. 2002.
[50] H. Abdi and L. J. Williams, "Principal component analysis," *WIREs Comput. Statistic*, vol. 2, no. 4, pp. 433–459, Jul./Aug. 2010.
[51] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.
[52] Z. Feng, L. Mo, and M. Li, "A random forest-based ensemble method for activity recognition," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Milan, Italy, Aug. 2015, pp. 5074–5077.
[53] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.
[54] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Trans. Syst. Man, Cybern.*, vol. 21, no. 3, pp. 660–674, May 1991.
[55] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," in *Proc. OTM Confederated Int. Conf. Move Meaningful Internet Syst.*, Catania, Italy, Nov. 2003, pp. 986–996.
[56] M. S. Alam, K.-C. Kwon, M. A. Alam, M. Y. Abbass, S. M. Imtiaz, and N. Kim, "Trajectory-based air-writing recognition using deep neural network and depth sensor," *Sensors*, vol. 20, no. 2, p. 376, Jan. 2020.
[57] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *Proc. Int. Conf. Eng. Technol. (ICET)*, Antalya, Turkey, Aug. 2017, pp. 1–6.
[58] L. R. Medsker and L. C. Jain, "Recurrent neural networks," *Design Appl.*, vol. 5, pp. 64–67, Dec. 2001.
[59] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomed. Signal Process. Control*, vol. 47, pp. 312–323, Jan. 2019.
[60] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd ACM Int. Conf. Multimedia*, Brisbane, QLD, Australia, Oct. 2015, pp. 1015–1018.
[61] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 1041–1044.
[62] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, "Evaluation of algorithms using games: The case of music tagging," in *Proc. 10th Int. Soc. Music Inf. Retr. Conf.*, Kobe, Japan, Oct. 2009, pp. 387–392.
[63] B. McFee, T. Bertin-Mahieux, D. P. W. Ellis, and G. R. G. Lanckriet, "The million song dataset challenge," in *Proc. 21st Int. Conf. World Wide Web*, Lyon, France, Apr. 2012, pp. 909–916.
[64] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The MTG-Jamendo dataset for automatic music tagging," in *Proc. 36th Int. Conf. Mach. Learn.*, Jun. 2019, pp. 1–3.
[65] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Trans. Comput.*, vol. COM-100, no. 1, pp. 90–93, Jan. 1974.
[66] S. Y. Boulahia, A. Amamra, M. R. Madi, and S. Daikh, "Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition," *Mach. Vis. Appl.*, vol. 32, no. 6, pp. 1–18, Nov. 2021.
[67] 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 Appl.*, vol. 81, no. 4, pp. 4621–4647, Feb. 2022.
[68] N. Pelchat and C. M. Gelowitz, "Neural network music genre classification," *Can. J. Electr. Comput. Eng.*, vol. 43, no. 3, pp. 170–173, Summer. 2020.
[69] Y.-H. Cheng, P.-C. Chang, and C.-N. Kuo, "Convolutional neural networks approach for music genre classification," in *Proc. Int. Symp. Comput., Consum. Control (IS3C)*, Taichung, Taiwan, Nov. 2020, pp. 399–403.
[70] K. S. Mounika, S. Deyaradevi, K. Swetha, Vanitha, and V, "Music genre classification using deep learning," in *Proc. Int. Conf. Advancements Electr., Electron., Commun., Comput. Autom. (ICAECA)*, Erod, India, Oct. 2021, pp. 1–7.
[71] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, Haifa, Israel, Jun. 2010, pp. .807–814.
[72] Y. Cai and W. Xu, "The best input feature when using convolutional neural network for cough recognition," *J. Phys., Conf. Ser.*, vol. 1865, no. 4, Apr. 2021, Art. no. 042111.
[73] J.-H. Yang, S. Park, S. Kim, Y. Cho, and J. J. Yoh, "Accurate real-time monitoring of fine dust using a densely connected convolutional networks with measured plasma emissions," *Chemosphere*, vol. 293, Apr. 2022, Art. no. 133604.
[74] P.-T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Ann. Oper. Res.*, vol. 134, no. 1, pp. 19–67, Feb. 2005.
[75] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[76] F. Gouyon, S. Dixon, E. Pampalk, and G. Widmer, "Evaluating rhythmic descriptors for musical genre classification," in *Proc. AES 25th Int. Conf. Metadata Audio*, London, U.K., Jun. 2004, pp. 1–9.
[77] T. Pavlín, J. Cech, and J. Matas, "Ballroom dance recognition from audio recordings," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Taichung, Taiwan, Jan. 2021, pp. 2142–2149.
[78] S. Chillara, A. S. Kavitha, S. A. Neginhal, S. Haldia, and K. S. Vidyullatha, "Music genre classification using machine learning algorithms: A comparison," *Int. Res. J. Eng. Technol.*, vol. 6, no. 5, pp. 851–858, 2018.
[79] H. Homburg, I. Mierswa, B. Möller, K. Morik, and M. Wurst, "A benchmark dataset for audio classification and clustering," in *Proc. 6th Int. Conf. Music Inf. Retr.*, London, U.K., Sep. 2005, pp. 528–531.
[80] K. Seyerlehner, G. Widmer, and T. Pohle, "Fusing block-level features for music similarity estimation," in *Proc. 13th Int. Conf. Digital Audio Eff.*, Graz, Austria, Sep. 2010, pp. 225–232.
[81] M. Chełkowska-Zacharewicz and M. Janowski, "Polish adaptation of the Geneva emotional music scale: Factor structure and reliability," *Psychol. Music*, vol. 49, no. 5, pp. 1117–1131, Sep. 2021.
[82] J. Chae, S.-H. Cho, J. Park, D.-W. Kim, and J. Lee, "Toward a fair evaluation and analysis of feature selection for music tag classification," *IEEE Access*, vol. 9, pp. 147717–147731, 2021.
[83] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Dec. 2006.
[84] O. J. Dunn, "Multiple comparisons among means," *J. Amer. Stat. Assoc.*, vol. 56, no. 293, pp. 52–64, Mar. 1961.
[85] I. Sa, Z. Ge, F. Dayoub, B. Upcroft, T. Perez, and C. McCool, "DeepFruits: A fruit detection system using deep neural networks," *Sensors*, vol. 16, no. 8, p. 1222, Aug. 2016.

[86] Z. Xing, S. Zhao, W. Guo, F. Meng, X. Guo, S. Wang, and H. He, "Coal resources under carbon peak: Segmentation of massive laser point clouds for coal mining in underground dusty environments using integrated graph deep learning model," *Energy*, vol. 285, Dec. 2023, Art. no. 128771.
[87] C. Xie, H. Song, H. Zhu, K. Mi, Z. Li, Y. Zhang, J. Cheng, H. Zhou, R. Li, and H. Cai, "Music genre classification based on res-gated CNN and attention mechanism," *Multimedia Tools Appl.*, pp. 1–16, 2023.
[88] Y.-N. Hung, C. H. Yang, P.-Y. Chen, and A. Lerch, "Low-resource music genre classification with cross-modal neural model reprogramming," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Rhodes Island, Greece, Jun. 2023, pp. 1–5.
[89] Z. Wen, A. Chen, G. Zhou, J. Yi, and W. Peng, "Parallel attention of representation global time–frequency correlation for music genre classification," *Multimedia Tools Appl.*, to be published.
[90] H. Zhao, C. Zhang, B. Zhu, Z. Ma, and K. Zhang, "S3T: Self-supervised pre-training with Swin transformer for music classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Singapore, May 2022, pp. 606–610.
[91] Z. Liu, T. Bian, and M. Yang, "Locally activated gated neural network for automatic music genre classification," *Appl. Sci.*, vol. 13, no. 8, pp. 1–11, Apr. 2023.

WANGDUK SEO received the B.S., M.S., and Ph.D. degrees in computer science from Chung-Ang University, Seoul, South Korea, in 2017, 2019, and 2023, respectively. He is currently a full-time Researcher with the AI/ML Innovation Research Center, Chung-Ang University. His research interests include meta-heuristic optimization, multi-label learning, and feature selection.



SUNG-HYUN CHO received the B.S. degree from Korea University, Sejong-si, South Korea, and the M.S. degree from the Department of Artificial Intelligence, Chung-Ang University, Seoul, South Korea. His research interests include recommendation systems, multimodal learning methods, and feature selection.

PAWEŁ TEISSEYRE received the Ph.D. degree from the Institute of Computer Science, Polish Academy of Sciences, in 2013. He is currently an Assistant Professor with the Institute of Computer Science, Polish Academy of Sciences, and the Faculty of Mathematics and Information Sciences, Warsaw University of Technology. His research interests include feature selection in high-dimensional supervised problems, multi-label classification, learning from partially labeled data, and applications of machine learning methods in medicine and genetics.

JAESUNG LEE received the B.S., M.S., and Ph.D. degrees in computer science from Chung-Ang University, Seoul, Republic of Korea, in 2007, 2009, and 2013, respectively. He also studies classification and feature selection, especially multilabel learning with information theory. He is currently the Head and an Associate Professor with the Department of Artificial Intelligence, Chung-Ang University, where he is also the Chief of the AI/ML Innovation Research Center. His research interests include machine learning, multilabel learning, model selection, and neural architecture search.