Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Toward a Fair Evaluation and Analysis of Feature Selection for Music Tag Classification

JONGHOON CHAE[1], SUNG-HYUN CHO[2], JAEGYUN PARK[1], DAE-WON KIM[1], (Member, IEEE), AND JAESUNG LEE[2]

[1]School of Computer Science and Engineering, Chung-Ang University, 221, Heukseok-Dong, Dongjak-Gu, Seoul 06974, Republic of Korea
[2]Department of Artificial Intelligence, Chung-Ang University, 221, Heukseok-Dong, Dongjak-Gu, Seoul 06974, Republic of Korea

Corresponding author: Dae-Won Kim (dwkim@cau.ac.kr) and Jaesung Lee(jslee.cau@gmail.com).

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2021-0-01341,Artificial Intelligence Graduate School Program(Chung-Ang University)) and the Chung-Ang University Research Grants in 2020.

ABSTRACT Accurate music tag classification of music clips has been attracting great attention recently, because it allows one to provide various music excerpts, including unpopular ones, to users based on the clips' acoustic similarities. Given a user's preferred music, acoustic features are extracted and then fed into the classifier, which outputs the related tag to recommend new music. Furthermore, the accuracy of the tag classifiers can be improved by selecting the best feature subset based on the domain to which the tag belongs. However, recent studies have struggled to evaluate the superiority of various classifiers because they utilize different feature extractors. In this study, to conduct a direct comparison of existing methods of classification, we create 20 music datasets with the same acoustic feature structure. In addition, we propose an effective evolutionary feature selection algorithm to evaluate the effectiveness of feature selection for tag classification. Our experiments demonstrate that the proposed method improves the accuracy of tag classification, and the analysis with multiple datasets provides valuable insights, such as the important features for general music tag classification in target domains.

INDEX TERMS Music tag classification, feature selection, machine learning, evolutionary algorithm.

I. INTRODUCTION

Automatic music tag classification (MTC) is used to find a relevant music tag, such as emotion or genre, for a music excerpt based on its music signal or extracted acoustic features [1]–[3]. This task can be achieved by training a classifier using music excerpts with relevant tags annotated by a human being. After training the classifier, relevant tags for undiscovered or newly released music excerpts can be identified without human intervention by feeding the excerpts as input data to the trained classifier. Since automatic MTC has high learning potential, improving the accuracy of the classifier is an essential task.

In the past decades, various researchers considered acoustic feature selection (FS) as a key pre-processing step for improving the learning accuracy of MTC. The detailed procedure is as follows:

1) A collection of digital music clips or excerpts and the relevant tags are gathered. Regarding application constraints such as response time, the music excerpts can be divided into clips with a specific duration.
2) An acoustic feature extractor is selected to transform the signal information of each music excerpt into a series of statistical values that form the formal dataset for training the classifier.
3) Since relevant acoustic features may vary according to the domain of the tags, an FS algorithm can be employed to improve classification accuracy.
4) The classifier learns the music excerpts based on the selected acoustic features with improved accuracy because noisy features can be removed through the FS process.

MTC performance varies according to the domain of the music collection, acoustic feature extractor, feature selector, and classifier. However, recent studies report the performance of MTC based on different settings, making it impossible to identify and compare different aspects of performance

[4]–[6]. Therefore, we cannot judge the impact of FS on improving MTC performance.

In this work, we report the performance of automatic music classification in a fair setting. The strategy and the contributions of this study can be summarized as follows:

- We gathered as many music collections as possible to prevent a biased conclusion—20 datasets were included in our experiments.
- To guarantee a fair basis for the comparisons, we selected MIRtoolbox [7], one of the most popular toolkits in the field, as the acoustic feature extractor.
- To examine the true potential of FS, we devised a novel evolutionary FS algorithm based on an evolutionary search process enhanced by a feature filter.
- Our experiments, based on the same acoustic feature structure, allow a feature analysis of multiple datasets to identify important acoustic features across various MTC settings.

Our experimental results indicate that an effective FS method can improve the performance of MTC consistently regardless of the underlying setting.

II. RELATED WORK

Using different combinations of the music collection, acoustic feature extractor, FS algorithm, and classifier, various MTC systems can be built with different performance levels. In this section, we review some notable studies related to MTC based on these concepts.

First, regarding the music collection, several datasets have been used in MTC studies. One well-known dataset is the Latin Music Database [6], [14], [24], [25]. It contains 3,227 music pieces belonging to 10 different Latin musical genres: Axe, Bachata, Bolero, Forro, Gaucha, Merengue, Pagode, Salsa, Sertaneja, and Tango. The ISMIR'2004 dataset consists of 1,458 music pieces assigned to six Western genres: classical, electronic, jazz and blues, metal and punk, rock and pop, and world music [6]. Another popular dataset is GTZAN [3], [18], [22]. It is widely used in studies on music classification, and it consists of 1,000 songs from 10 popular genres: blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, and rock. Each song is 30s long with a sample rate of 22,050kHz. The Seyerlehner:1517-Artists dataset contains 3,180 original tracks from 23 music genres. From the original tracks, only tracks from different artists in each category were selected in [1]. MagnaTagAtune comprises 4,476 songs belonging to 24 musical genres, making it the most diverse collection [29]. The MIR-1K dataset, a collection of Tollywood and Bollywood songs from the Indian film industry, has been used for segmentation of vocal and non-vocal clips [11]. Other well-known datasets are EMO-DB, eNTEFACE05, EMVO, SAVEE [9], Traditional Malay Music [26], AllMusicGuide [21], and Thai music collection [2]. Many studies have formed a novel dataset by combining several datasets or even creating a new one. For example, Carnatic, Hindustani, and Bollywood musical datasets were joined together to create a single dataset [12]; in another study, 574 vehicle sounds were selected from three instrumental datasets—the McGill University collection, RWC database, and University of Iowa instrument samples [15], [20]. The datasets in the works of [10], [13], [16], [17], [23], [28], [31] were newly collected for the respective studies. Some were collected from music CDs and the Internet [31], while some were from individuals who were asked to classify respectively. These datasets were used to validate the performance of music classification algorithms. For example, in the work of [4], 12 datasets were used—Audionautix, Bugs2664, BugsEmo, CAL500, China3004, Emotions, Genre3, Highlight, KOCCA40, MusicEmo-A, MusicEmo-B, and Style812.

Second, regarding the acoustic feature extractors, various extractors have been developed over time as MTC research continues and diversifies. The two popular frameworks are Music Analysis, Retrieval and Synthesis for Audio Signals (MARSYAS) [2], [6], [24]–[26] and Advanced MUSic Explorer (AMUSE) [5], [20], [21]. Feature extractors such as jAudio [18], [29], MIRtoolbox [4], [12], [18], and the Chordino Vamp Plugin [1] can be used both independently or based on the aforementioned two frameworks. VOICEBOX [17], MPEG-7 [18], OpenSMILE [9], CLAM [19], [23], and librosa from Python [3] are also used for feature extraction.

Third, regarding FS algorithms, previous studies frequently use a genetic algorithm (GA) to select the relevant acoustic features for given tags. This method can be combined with other methods, and the converted GAs or other FS algorithms were used contextually in each study. For instance, the One-Against-All (OAA) and Round Robin (RR) space decomposition approaches have been combined with GA [6], [24], [25]. S-Metric Selection Evolutionary Multi-objective Algorithms (SMS-EMOA) [1], [5], [20], [21] are often selected as an optimization heuristic in MTC studies. They sort the population into several solution fronts through fast non-dominated sorting [20]. The interactive genetic algorithm (IGA) is used both in [23] and [19]. In addition, GA with the m-features operator was introduced in [28]. Sequential forward selection (SFS) has been used both individually [29], [30] and in combination with ReliefF [17]. Particle swarm optimization [2], the self-adaptive particle swarm optimization (SaPSO) [32], the self-adaptive harmony search (SAHS) algorithm [18], the interactive feature selection (IFS) method [30], and improved binary global harmony search (IBGHS) [8] are the other FS algorithms used in related research. [30] and [26] conducted studies using more than two algorithms: three algorithms in [30] and eight algorithms in [26].

Fourth, regarding the classifiers, we will mainly discuss five popular ones used in MTC studies. The decision tree (DT) is often employed as a classifier to decide the tag for a given music excerpt. Each node in a DT has two child nodes, except for the leaf nodes, which are used to decide the relevant tag for a given music excerpt [2], [5], [6], [12], [20], [21], [24], [25]. Closely related to the DT, the random forest (RF) is an ensemble learning method that randomly

TABLE 1: Summary of related papers.

Author	Year	# of employed Datasets	Acoustic Feature Extractor	Feature Selection	Classifier	Ref.
Ginsel et al.	2020	1	Chordino Vamp Plugin	Multi-objective Evolutionary FS using SMS-EMOA	kNN, RF	[1]
Campobello et al.	2020	1	librosa	GA	Novel Classifier	[3]
Gholami et al.	2020	18	-	IBGHS	kNN	[8]
Lee et al.	2019	10	MIRtoolbox	GA	Multilabel NB	[4]
Leartpantulak and Kitjaidure	2019	1	-	-	kNN, DT, RF, SVM, NB, Stacking Ensemble	[2]
Özseven	2019	4	OpenSMILE v2.1.0	FS based on Emotion changes	SVM, MLP, kNN	[9]
Ma	2019	1	-	GA	-	[10]
Murthy et al.	2018	2	-	GA	SVM, NFC, RF, NN	[11]
Kalapatapu et al.	2016	3	MIRtoolbox	GA, SFS, IG, CBFS	C4.5, kNN, ANN, SVM	[12]
Serwach et al.	2016	2	-	GA	ANN, kNN	[13]
Zottesso et al.	2016	1	-	GA	SVM	[14]
Alexandre et al.	2015	1	-	GA	ELM	[15]
Kour and Mehan	2015	1	-	-	SVM, BPNN	[16]
Wu and Wang	2015	1	VOICEBOX	ReliefF-SFS	SVM	[17]
Huang et al.	2014	1	MIRtoolbox, jAudio, MPEG-7	SAHS	SVM	[18]
Athani et al.	2014	-	CLAM	IGA	kNN	[19]
Vatolkin	2013	1	AMUSE	SMS-EMOA	C4.5, RF, NB, SVM	[5]
Vatolkin et al.	2012	3	AMUSE	SMS-EMOA	C4.5, RF, NB, SVM	[20]
Vatolkin et al.	2011	1	AMUSE	SMS-EMOA	C4.5, RF, NB, SVM	[21]
Karkavitsas et al.	2011	1	-	GA	kNN	[22]
Kim et al.	2010	-	CLAM	IGA	kNN	[23]
Silla et al.	2009	2	MARSYAS	GA, OAA + GA, RR + GA	C4.5, kNN, NB, MLP, SVM	[6]
Silla et al.	2008	1	MARSYAS	GA, OAA + GA, RR + GA	NB, DT, SVM, MLP	[24]
Silla et al.	2008	1	MARSYAS	GA, OAA + GA, RR + GA	C4.5, kNN, MLP, NB, SVM	[25]
Doraisamy et al.	2008	1	MARSYAS	CBFS, PCA, χ^2-statistics Gain Ratio, Feature Wrapper(SVM)	20 classifiers in WEKA Toolkit	[26]
Rho et al.	2007	-	-	GA	-	[27]
Alexandre et al.	2007	1	-	GA	-	[28]
Fiebrink and Fujinaga	2006	1	jAudio	SFS	kNN	[29]
Kim et al.	2006	-	-	GA, SFS, IFS	-	[30]
Xu et al.	2003	1	-	-	SVM, ANN, GMM, HMM	[31]

selects acoustic features from several different sets [1], [2], [11]. The k-nearest neighbor (kNN) classifier calculates the distance between the test music pieces and training music pieces in terms of the acoustic features [1], [8], [9]. The naïve Bayes (NB) classifier uses Bayesian theory for deciding the relevant tag of music pieces [2], [5], [20]. It estimates the output tag by its highest probability based on feature distribution. Support vector machine (SVM) is a supervised learning model used for MTC [2], [9], [11]. In addition to the classifiers introduced thus far, many classifiers—such as the multi-layer perceptron (MLP) [6], [9], [24], neural network (NN) [11]–[13], stacking ensemble [2], neuro-fuzzy classifier (NFC) [11], and back propagation neural network (BPNN) [16]—have been used in MTC studies. Specifically, one study developed new kinds of classifiers [3]. [9] and [26] chose 18 classifiers.

Table 1 summarizes the number of music datasets considered in each corresponding study, the chosen acoustic feature extractor, the FS algorithm employed, and the classifier used for identifying the tags. Table 8 lists the abbreviations used in Table 1. Our brief review shows that the choices in each study varied, indicating that it is difficult to directly compare the impact of FS on the MTC.

III. MATERIALS AND METHODS

A. PREPROCESSING

We used the MIRtoolbox to extract the acoustic features from music excerpts. The basic operators of MIRtoolbox are related to the management of audio waveforms, frame-based analysis, periodicity estimation, auditory modeling, peak selection, and sonification of results. During preprocessing, we first input the music excerpts in the collection

TABLE 2: Summary of public music collections considered in this study.

Collection	Description	Domain	Tags	Refs.
Ballroom	698 songs of maximum 30s in eight genres of ballroom dancing (BallroomDancers.com)	Genre	Cha Cha, Jive, Quickstep, Rumba, Samba, Tango, Viennese Waltz, Waltz	[33]
Ballroom-Extended	Improved version of the Ballroom dataset (better audio quality, six times more tracks, five new rhythm classes and annotations of different types of repetitions)	Genre	Cha Cha, Jive, Quickstep, Rumba, Samba, Tango, Viennese Waltz, Waltz, Foxtrot, Pasodoble, Salsa, Slow Waltz, Wcswing	[34]
emoMusic	1,000 songs of 44,100Hz selected from FMA for music emotion recognition	Emotion	Arousal, Valence	[35]
Emotify	400 song excerpts of 60s Tags from 64 annotators based on the Geneva Emotional Music Scales [36] with a maximum of three tags for each song	Emotion	Amazement, Solemnity, Tenderness, Nostalgia, Calmness, Power, Joyful Activation, Tension, Sadness	[37]
FMA-MEDIUM	Free Music Archive (FMA) consists of Creative Commons audio from 105,547 tracks from 16,341 artists and 14,854 albums of 161 genres FMA-MEDIUM consists of 25,000 tracks each 30 seconds long, 16 unbalanced genres selected from FMA	Genre	Blues, Classical, Country, Easy Listening, Electronic, Experimental, Folk, Hiphop, Instrumental, International, Jazz, Old-Time/Historic, Pop, Rock, Soul-RnB, Spoken	[38]
FMA-SMALL	FMA-SMALL consists of 8,000 tracks, each 30s long, belonging to eight balanced genres selected from FMA	Genre	Hiphop, Pop, Folk, Experimental, Rock, International, Electronic, Instrumental	[38]
GiantStepsKey	604 electronic dance music excerpts from www.beatport.com for key estimation	Key	Key	[39]
GMD	13.6 hours, 1,150 MIDI files, and over 22,000 measures of drumming	Groove	Hits, Offsets, Velocities	[40]
GoodSounds	12 different instruments recorded by one or up to four different microphones Each note is recorded in a mono .flac audio file (48,100Hz, 32 bits)	Instrument	Flute, Cello, Violin, Clarinet, Trumpet, Saxophone(Alto), Saxophone(Tenor), Saxophone(Baritone), Saxophone(Soprano), Oboe, Piccolo, Bass	[41]
GTZAN	1,000 audio tracks, 30s long, 10 genres each represented by 100 tracks 22,050Hz mono 16-bit audio files from personal CDs, radio, microphone recordings	Genre	Blues, Classical, Country, Disco, Hiphop, Jazz, Metal, Pop, Reggae, Rock	[42]

to the `mirfeatures` function in MIRtoolbox. Then, a total of 888 numeric features were computed and returned, which were further categorized into dynamic (6), rhythm (37), timbre (739), pitch (0), and tonal (106). Then the outputs were extracted to standard CSV files via the `mirexport` function of MIRtoolbox, which exports the statistical information of each music excerpt.

Next, due to unexpected processes such as the divided-by-zero error, the statistical information may include Not-a-Number (NaN) values. In this study, we used kNN imputation [53], where $k = 3$, to deal with the NaN values. Thus, all the NaN values were replaced with a value estimated by referencing the nearest three music excerpts. However, if an acoustic feature contains too many NaN values, the kNN imputation cannot be applied because the algorithm is unable to identify the nearest neighbors; in such cases, we removed those features. We encoded the tag information using the one-hot encoding scheme [54] because this allows for multiple tags to be assigned to a single music excerpt.

Finally, for the case in which classifiers are used that are known for being effective with categorical features, we used the LAIM discretization scheme [55] for discretizing the original numeric features into the categorical features.

B. PROPOSED FEATURE SELECTION ALGORITHM

1) Initialization and Evaluation

To search for the optimal feature subset, the proposed method first initializes a population composed of chromosomes. Each

TABLE 2: Summary of public music collections considered in this study (Continued.)

Collection	Description	Domain	Tags	Refs.
HOMBURG	1,886 songs from www.garageband.com 44,100Hz frequency and 128kb bit MP3 files	Genre	Alternative, Blues, Electronic, Jazz, Folk/Country, Pop, Funk/Soul/RnB, Rap/Hiphop, Rock	[43]
ISMIR04	2187 audio files in MP3 format from http://magnatune.com/	Genre	Classical, Electronic, Jazz/Blues, Metal/Punk, Rock/Pop, World	[44]
Medley-solos-DB	Cross-collection dataset selected from two different music collections for music instrument recognition (3s long)	Instrument	Clarinet, Distorted Electric Guitar, Female Singer, Flute, Piano, Saxophone(Tenor), Trumpet, Violin	[45]
MER500	Hindi film music dataset 100 audio files of 10s each	Emotion	Romantic, Happy, Sad, Devotional, Party	[46]
MICM	Maryam Iranian Classical Music Dataset for the classification of the seven Dastgahs of Iranian classical music	Genre	Shour, Homayoun Mahour, Segah, Chahargah, Rastpanjgah, Nava	[47]
MIREX-like_mood	Multi-modal source consisting of 903 audio clips (22,050Hz, 16 bit, Mono), 764 lyrics, and 193 MIDIs	Mood	Five Mood Clusters	[48]
PCMIR	Persian Classical Instrument Recognition 2,410 samples that are each 5–10s long	Instrument	Ney, Tar, Santur, Kamancheh, Tonbak, Ud, Setar	[49]
Seyerlehner:Unique	3,115 music excerpts of popular songs in 14 unbalanced genres, with each excerpt being 30s long	Genre	Blues, Classical, Country, Dance, Electronica, Hiphop, Jazz, Reggae, Rock, Schlager, Soul/RnB, Folk, World, Spoken	[50]
Soundtracks	Short MP3 excerpts, each of 15s, from film soundtracks	Emotion	Happy, Sad, Tender, Fearful, Angry, Surprising, Valence(Pos./Neg.), Energy(Pos./Neg.), Tension(Pos./Neg.)	[51]
Tropical Genres	A new collection for Columbia classical music genre classification	Genre	Bachata, Cumbia Merengue, Salsa, Vallenato	[52]

chromosome is represented as a binary vector consisting of ones and zeros depending on whether a feature is selected. The chromosomes are initialized randomly such that they are evenly distributed across the entire search space. To verify that removing noisy features can improve accuracy, the initial population includes one chromosome that selects all features. After that, each chromosome is evaluated as a fitness value by the classifier. Specifically, the classifier is trained with a feature subset represented by each chromosome and then predicts the tag for each test music excerpt. Given the correct tags and predicted tags, a fitness value is obtained by calculating the accuracy of the tagging. Therefore, a better feature subset has a higher fitness value.

2) Parent Selection

Parent selection is the process of selecting parents to generate offspring for the next generation. To pass on the relevant features to the offspring, parents should be chosen from the population based on their fitness values. To this end, we adopted a tournament selection algorithm [56]. Given chromosome groups sampled at random from the population, the tournaments are conducted by comparing fitness values. The winner of each tournament is selected as a parent. In an MTS problem, a feature subset with a high fitness value does not ensure high accuracy for each tag because the dependencies between the tags are different. For example, two parents with high fitness values may consist of similar features related to the same tags. To generate superior offspring, a pair of two parents must have a complementary relationship, with each containing features related to different tags. If one parent has features that depend on one part of the tags, another parent that has features relevant to the remaining tags can complement it.

To deal with the issue, we introduce a new parent-matching process. Let c be a chromosome chosen by a tournament and T be a set of tags. A tag-specific accuracy vector $a^c = [a_1, ..., a_{|T|}]$ is computed by the trained classifiers, each used to predict each tag; here, a_i is the accuracy of the ith tag predicted by c. To divide T into a strong tag subset T_s and weak tag subset T_w, the elements of a^c first are sorted in ascending order. After that, T_s and T_w are identified based on the point where the difference between the two consecutive values in sorted a^c is the highest because the optimal point is unknown. Given T_w on c, its spouse c' is chosen as follows.

$$c' \leftarrow \arg\max_{c'} \sum_{i \in T_w} a_i^{c'} \tag{1}$$

3) Offspring Generation via Feature Filter

Traditionally, the offspring are generated by randomly recombining pairs of parents because it is difficult to assess the importance of each feature only based on the fitness value of a specific chromosome. If such information can be estimated, further improved offspring can be generated by utilizing the benefits of our parent-matching process. Given a pair of parents c and c', the offspring must inherit features closely related to T_s and T_w from c and c', respectively. To this end, we employ a feature filter that enables us to compute the relationship between features and tags based on information theory.

Let S_c and S'_c be feature subsets represented by c and c'. Given an offspring $S_n = \varnothing$, S_c can pass a feature $f^+ \in \{S_c \setminus S_n\}$ sequentially to S_n. To maximize the dependency between features and T_s, f^+ is selected as follows.

$$\arg\max_{f^+ \in \{S_c \setminus S_n\}} I(S_n, f^+; T_s) \tag{2}$$

where $I(X;Y) = H(X) + H(Y) - H(X,Y)$ is Shannon's mutual information between the variable set X and Y. To calculate $H(T_s)$, we transform a categorical variable T into a binary one-hot encoding for calculating Eq. (2). Therefore, we adopt a generalized information-theoretic criterion [57], which is a recent feature filter. It calculates f^+ as follows.

$$\arg\max_{f^+ \in \{S_c \setminus S_n\}} |T_s| \sum_{f \in S_n} H(f^+, f) - |S_n| \sum_{l \in T_s} H(f^+, l) \tag{3}$$

where S'_c also can pass features to S_n by replacing S_c and T_s with S'_c and T_w, respectively. To explore feature subsets of various sizes, we randomly set the number of features that S_c and S'_c pass into S_n at each iteration.

The offspring are generated as a combination of only the features that the population has. Therefore, some features in the original set are neglected, resulting in local optima. A simple approach to solve this issue is to extend the aforementioned offspring generation process to the population level. Let F and S_p be the original feature set and union of feature subsets within the population, respectively. Given T_s and T_w on S_p, new features to add into the population are selected as follows.

$$\arg\max_{f^+ \in \{F \setminus S_p\}} |T_w| \sum_{f \in S_p} H(f^+, f) - |S_p| \sum_{l \in T_w} H(f^+, l) \tag{4}$$

Similarly, features to delete in the population are selected by replacing F, S_p, and T_w with S_p, $\varnothing$, and T_s, respectively. After that, the modified S_p is added to the population as a new chromosome. Finally, the proposed method repeats these processes until the termination condition is met.

4) Termination

The termination condition is based on the number of fitness function calls (FFCs), that is the number of evaluations of individuals by the classifier. The algorithm terminates its search process if the number of remaining FFCs is zero. The number of allowed FFCs is a user-defined parameter. Algorithm 1 represents the pseudocode of the proposed method.

Algorithm 1 Proposed feature selection algorithm

```
Input: $F$;                          ▷ the original feature set $F$
Output: $S$;                         ▷ the final feature subset $S$
$u \leftarrow 0$;                    ▷ the number of spent FFCs $u$
initializing $P$;                    ▷ the population $P$
evaluating $P$;                      ▷ compute fitness values via classifiers
$u \leftarrow u + m$;                ▷ the population size $m$
while $u < v$ do                     ▷ the maximum number of FFCs $v$
    $N \leftarrow \emptyset$;
    for each tournament do
        $c \leftarrow$ run a tournament selection;
        $c' \leftarrow$ compute Eq. (1);
        $o \leftarrow$ offspring of $c$ and $c'$;    ▷ use Eq. (3)
        $N \leftarrow \{N \cup o\}$;
    end for
    $o \leftarrow$ run mutation;                    ▷ use Eq. (4)
    $N \leftarrow \{N \cup o\}$;
    evaluating $N$;
    $u \leftarrow u + (m/$tournament size$)$;
    $P \leftarrow \{P \cup N\}$;
    $P \leftarrow$ keep $m$ chromosomes with best fitness values;
end while
$S \leftarrow$ the best feature subset so far;
```

IV. EXPERIMENTAL RESULTS

We created 20 music datasets from different domains to validate the effect of FS on MTC. Table 2 provides short descriptions of the music collections employed, where each music collection consists of audio files and music tags annotated by users or through a specific system. As shown in Table 2, the domains of the collections cover 12 genres, three emotions, three instruments, one mood, one groove, and one key. Since the main objective of this study was to conduct a fair experiment on FS, the acoustic features were extracted using the same feature extractor—MIRtoolbox [7]—which provides a set of integrated features for each music excerpt. The extracted features can be organized along five main

TABLE 3: The standard characteristics and statistics of employed datasets

Dataset	$\lvert W\rvert$	Avg. Length	$\lvert F\rvert$	Suggested Domain	Used Domain	$\lvert T\rvert$	Avg. Size of Tag		
							Mean	±	Std.
Ballroom	698	30	888	Genre	Genre	8	87.25	±	17.56
Ballroom-Extended	4,180	30	888	Genre	Genre	13	321.54	±	195.70
emoMusic-G	1,000	45	888	Emotion	Genre	8	125.00	±	0.00
Emotify-G	400	43	888	Emotion	Genre	4	100.00	±	0.00
FMA-MEDIUM	24,984	30	888	Genre	Genre	16	1561.50	±	2069.36
FMA-SMALL	7,996	30	888	Genre	Genre	8	999.50	±	0.50
GiantStepsKey-G	604	80	888	Key	Genre	23	24.9130	±	28.0274
GMD-G	1,042	300	888	Groove	Genre	18	57.89	±	70.07
GoodSounds	8,399	9	865	Instrument	Instrument	12	549.40	±	198.70
GTZAN	1,000	30	888	Genre	Genre	10	100.00	±	0.00
HOMBURG	1,886	10	888	Genre	Genre	9	209.56	±	134.87
ISMIR04	2187	120	888	Genre	Genre	6	121.17	±	95.60
Medley-solos-DB	21,571	3	873	Instrument	Instrument	8	2537.88	±	2016.08
MER500	493	30	888	Emotion	Emotion	5	98.60	±	2.33
MICM	1,137	360	888	Genre	Genre	7	162.43	±	119.72
MIREX-like_mood	903	30	884	Mood	Mood	28	32.25	±	4.78
PCMIR	2,410	5	886	Instrument	Instrument	6	401.67	±	46.24
Seyerlehner:Unique	3,115	30	888	Genre	Genre	14	311.50	±	248.35
Soundtracks	470	24	884	Emotion	Emotion	12	52.22	±	19.88
Tropical Genres	1,500	30	888	Genre	Genre	5	300.00	±	0.00

TABLE 4: List of modified datasets and considered tags

Dataset	Tags
emoMusic-G	Blues, Classical, Country, Folk, Electronic, Jazz, Pop, Rock
Emotify-G	Rock, Classical, Pop, Electronic
GiantStepsKey-G	Reggae-dub, Chill-out, Indie-dance-nu-dsc., Hiphop, Glitch-hop, Deep-house, House, Tech-house, Techno, Minimal, Funk-r-and-b, Pop-rock, Drum-and-bass, Hardcore-hard-tech. Electronica, Dj-tools, Electro-house, Hard-dance, Dubstep, Breaks, Trance, Psy-trance, Progressive-house
GMD-G	Afrobeat, Afrocuban, Blues, Country, Dance, Funk, Gospel, Highlife, Hiphop, Jazz, Latin, Middleeastern, Neworleans, Pop, Punk, Reggae, Rock, Soul

musical dimensions: dynamics, rhythm, timbre, pitch, and tonality.

Table 3 provides the statistics of the datasets created by applying the acoustic feature extractor to the music excerpts in the corresponding collection. Although a single domain that its relevant music tags can be included is suggested by the creator of collection in most cases, a few collections such as emoMusic, Emotify, GiantStepsKey, and GMD may cover multiple domains. For these collections, because the tags originally suggested by the creator are real-valued and consequently unsuitable for the classification task, we created a corresponding music dataset by using the genre tags shown in Table 4. Datasets are created using the genre tags emoMusic-G, Emotify-G, GiantStepsKey-G, and GMD-G. In Table 3, the terms $|W|$, Avg. Length, $|F|$, Suggested Domain, Used Domain, $|T|$, and Avg. Size of Tag, indicate the number of patterns, the average length of music excerpts in the collection, the number of extracted acoustic features, the domain suggested originally by the creator of the collection, the domain used for creating the corresponding dataset, the number of tags, and the average number of patterns per tag and its standard deviation, respectively.

In this study, we used the NB classifier for comparing the performance of the proposed and conventional FS methods because of its popularity and effectiveness. Specifically, we considered the NB classifier to obtain the accuracy values of the final feature subsets given by each FS method. To simulate performance in the real world, we used conventional hold-out cross-validation for each dataset. Of the patterns, 80% were randomly chosen as a training set, and the remaining 20% were randomly chosen as a test set. We compared the proposed method to three conventional methods: MPGA-LCC, MLACO, and CBFS.

- MPGA-LCC [58] is a Multi-population Genetic Algorithm for FS based on a new communication process among sub-populations. This communication process generates better offspring compared to the parent generation by employing the complementary discriminating powers of sub-populations.
- MLACO [59] is a new relevance-redundancy FS method based on Ant Colony Optimization. This algorithm tries to find the most promising features with the lowest redundancy and the highest relevancy.
- CBFS [60] ranks features according to a heuristic evaluation function based on Pearson's correlation coefficient. To conduct a fair comparison considering the 300 FFCs of EA-based comparison methods, we allowed the CBFS to validate the performance of 300 candidate feature subsets created by adding high-ranked features one-by-one to the feature subset. Among the 300 candidate feature subsets, the feature subset that yields the best accuracy value in the training phase is chosen as the final feature subset.

For the parameter settings of the evolutionary algorithms, 20% of the training patterns are used for evaluating the fitness of each chromosome. In short, 64%, 16%, and 20%

TABLE 5: Comparison results of four methods in terms of accuracy (▼/△ indicates that the corresponding method is significantly worse than the proposed method based on a paired t-test at the 95% significance level.)

Dataset	Proposed	MPGA-LCC	MLACO	CBFS
Ballroom	**83.67%** ± 4.64	73.31% ± 4.81 ▼	67.41% ± 6.30 ▼	64.82% ± 2.68 ▼
Ballroom-Extended	**63.49%** ± 1.41	51.82% ± 4.00 ▼	43.89% ± 4.29 ▼	51.71% ± 3.87 ▼
emoMusic-G	**28.80%** ± 2.80	23.10% ± 4.71 ▼	25.25% ± 2.24 ▼	26.85% ± 1.89 ▼
Emotify-G	**74.25%** ± 4.38	67.63% ± 7.35 ▼	71.62% ± 5.62 ▼	68.87% ± 5.73 ▼
FMA-MEDIUM	**51.98%** ± 0.88	50.47% ± 0.85 ▼	46.30% ± 1.04 ▼	43.76% ± 0.81 ▼
FMA-SMALL	**20.14%** ± 1.56	16.92% ± 0.92 ▼	16.47% ± 1.00 ▼	17.33% ± 1.19 ▼
GiantStepsKey-G	**17.83%** ± 3.79	15.17% ± 3.23	17.25% ± 0.12	10.67% ± 3.16 ▼
GMD-G	**26.25%** ± 2.63	24.52% ± 2.47	21.68% ± 3.17 ▼	26.06% ± 3.42
GoodSounds	**19.99%** ± 0.61	17.68% ± 1.15 ▼	19.92% ± 0.97	19.78% ± 0.97
GTZAN	**75.70%** ± 3.03	56.10% ± 4.95 ▼	59.00% ± 4.59 ▼	66.55% ± 2.22 ▼
HOMBURG	**31.17%** ± 2.88	29.07% ± 2.17 ▼	26.21% ± 1.94 ▼	23.37% ± 2.97 ▼
ISMIR04	89.89% ± 1.97	84.87% ± 10.78	61.36% ± 4.01 ▼	**91.22%** ± 1.19
Medley-solos-DB	**78.70%** ± 0.65	74.56% ± 2.01 ▼	72.33% ± 1.32 ▼	76.61% ± 0.40 ▼
MER500	**76.94%** ± 3.76	76.84% ± 4.79	65.41% ± 1.46 ▼	74.08% ± 4.79
MICM	7.80% ± 3.20	6.61% ± 3.74	**9.60%** ± 2.26	3.04% ± 5.29 ▼
MIREX-like_mood	**17.00%** ± 2.57	9.50% ± 2.29 ▼	12.22% ± 2.46 ▼	13.94% ± 3.11 ▼
PCMIR	**82.01%** ± 1.59	72.95% ± 2.93 ▼	70.68% ± 2.98 ▼	78.46% ± 1.47 ▼
Seyerlehner:Unique	**65.59%** ± 1.92	59.45% ± 2.08 ▼	57.23% ± 2.05 ▼	62.17% ± 1.67 ▼
Soundtracks	**45.42%** ± 5.16	34.44% ± 4.57 ▼	28.19% ± 7.86 ▼	32.36% ± 3.93 ▼
Tropical Genres	**92.97%** ± 1.09	88.63% ± 1.07 ▼	86.27% ± 3.61 ▼	92.50% ± 1.06
Avg. Rank	**1.10**✓	2.95	3.25	2.70

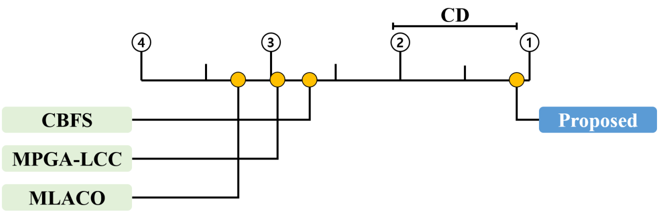


FIGURE 1: Bonferroni-Dunn test result of the four comparison methods in terms of accuracy.

of the patterns are used for training, validation, and testing, respectively. We set the size of the population to 50, and the maximum number of FFCs was limited to 300. To measure the performance of the four FS methods, we employed an accuracy measure in the range of [0,100]. Given the number of test patterns and the number of correct classifications, the accuracy is calculated as:

$$\text{Accuracy (\%)} = \frac{\text{Number of correct classifications}}{\text{Total number of classifications}} \times 100$$

where the total number of classifications is $|W| \times 0.2$ because we employed the 8:2 hold-out cross-validation. All experiments were repeated 10 times, and the average of the measured values was used to compare the performance of the different FS methods. To validate the superiority of proposed method, we performed additional experiments using a paired t-test [61] at the 95% significance level for comparing the proposed method with the other methods, and the Bonferroni-Dunn test [62] for comparing the average ranks of the proposed method with those of the other methods. In the Bonferroni-Dunn test, if the difference between the proposed method and the conventional method in terms of the average rank is larger than the value of the critical difference (CD), their performance is confirmed to be different and the superior method can be determined. In our experiments, we set the significance level α to 0.05, and thus, $CD = 0.9773$.

Table 5 shows the comparison results of the four methods in terms of average accuracy with standard deviation from ten repeated experiments. The best accuracy value among the four methods for each dataset is in bold. In the table, ▼/△ indicates that the corresponding method is significantly worse compared to the proposed method at the 95% significance level. The experimental result indicates that the proposed method outperforms the comparison methods for 13 datasets. For example, the proposed method yields an accuracy of 83.67%, 63.49%, 75.70%, and 45.42% for the Ballroom, Ballroom-Extended, GTZAN, and Soundtracks datasets, respectively. By contrast, the accuracy values of the second-best method in those datasets are 73.31%, 51.82%, 66.55%, and 34.44% respectively. Thus, the difference between two methods is 10.36%, 11.67%, 9.15%, and 10.98%, respectively, indicating that the proposed method significantly outperforms the second-best method for these four datasets. A similar tendency can be observed from the experiments with other datasets, resulting in an average rank of 1.10. Thus, the experimental results indicate that our proposed method can find a feature subset closely related to tags by introducing a new complementary parent matching process and offspring generation process based on information theory. Based on the experimental results of Table 5, we conducted the Bonferroni-Dunn test as shown in Figure 1. Since the proposed method achieved the best average rank, it is placed in the right most side in the figure. Above the middle line, a line CD runs from the location of the proposed method on the line. Since there are no other comparison methods within CD, the superiority of the proposed method is confirmed.

To validate the effect of FS on MTC, we compared the classification performance based on the feature subset se-

TABLE 6: Comparison of classification performance of feature subset selected by the proposed method and original feature set (▼/△ indicates that the classification performance on the original feature set is significantly worse/better compared to the proposed method based on a paired t-test at the 95% significance level.)

Dataset	Proposed	Original	Improvement
Ballroom	**83.67**% ± 4.64	75.25% ± 4.51 ▼	8.42%
Ballroom-Extended	**63.49**% ± 1.41	53.11% ± 1.99 ▼	10.38%
emoMusic-G	28.80% ± 2.80	**29.70**% ± 1.93	-0.9%
Emotify-G	**74.25**% ± 4.38	72.75% ± 3.81 ▼	1.5%
FMA-MEDIUM	**51.98**% ± 0.88	46.53% ± 0.75 ▼	5.45%
FMA-SMALL	**20.14**% ± 1.56	18.82% ± 1.10 ▼	1.32%
GiantStepsKey-G	17.83% ± 3.79	18.58% ± 4.08	-0.75%
GMD	**26.25**% ± 2.63	18.51% ± 2.61 ▼	7.74%
GoodSounds	**19.99**% ± 0.61	8.33% ± 0.73 ▼	11.66%
GTZAN	**75.70**% ± 3.03	71.20% ± 2.18 ▼	4.5%
HOMBURG	**31.17**% ± 2.88	27.61% ± 1.55 ▼	3.56%
ISMIR04	**89.89**% ± 1.97	77.49% ± 2.32 ▼	12.4%
Medley-solos-DB	**78.70**% ± 0.65	74.84% ± 0.30 ▼	3.86%
MER500	**76.94**% ± 3.76	72.14% ± 3.63 ▼	4.8%
MICM	**7.80**% ± 3.20	7.58% ± 2.05	0.22%
MIREX-like_mood	17.00% ± 2.57	**17.28**% ± 3.35	-0.28%
PCMIR	**82.01**% ± 1.59	77.72% ± 1.63 ▼	4.29%
Seyerlehner:Unique	**65.59**% ± 1.92	60.31% ± 1.52 ▼	5.28%
Soundtracks	**45.42**% ± 5.16	43.06% ± 4.77	2.36%
Tropical Genres	**92.97**% ± 1.09	91.60% ± 1.02 ▼	1.37%
Average	**52.48**%	48.12%	4.36%

lected by the proposed method and the original set composed of about 900 features as shown in Table 6. In the last column, the improvement in terms of accuracy of the feature subset selected by the proposed method over the original feature set is provided. Although the improvement of accuracy varies with the dataset, Table 6 indicates that the accuracy generally improved as a result of FS under the proposed method. Specifically, the proposed method improves the accuracy for the Ballroom, Ballroom-Extended, and GoodSounds datasets by 10% on average. As shown in the last row of Table 6, the FS process yields a 4.36% improvement on average. It is interesting to note that the feature subset selected by the other methods may yield a classification performance worse than that based on the original feature set. For example, the classification accuracy for the Ballroom dataset of the original feature subset is 75.25% whereas that of the feature subset selected by CBFS is 64.82%. Table 6 indicates that the original feature set includes some noisy features that may confuse classifiers. Thus, our experiments indicate that the FS process can improve the performance of MTC, but the choice of FS method is important.

V. ANALYSIS AND REMARKS

In this study, we created music datasets using the same acoustic feature extractor and identified important features for the given music tags. Since all the datasets in our study share a common feature structure, a subsequent analysis was conducted over multiple datasets for clues such as the important features for general MTC tasks, specific domains, and nationalities. In this section, we analyze the selected features over multiple music datasets.

A. OVERALL DESCRIPTION OF TOP 50 FEATURES SELECTED

Of the features in the fields of dynamics, rhythm, timbre, pitch, and tonality given by MIRtoolbox, features selected through the proposed method differed based on the dataset. Figure 2 shows the frequently selected 50 features across the entire experiment. The full names are detailed in Table 9 in the Appendix. In Figure 2, each bar represents the different feature fields of MIRtoolbox. Since we ran ten iterations for each dataset and we have 20 datasets, each feature can be selected for a maximum of 200 times. The dynamics and pitch features are excluded from the figure because they were not selected across all the datasets, indicating that the top 50 selected features are from the rhythm, timbre, and tonality fields. Timbre, in particular, with 27 features among the top 50, is clearly important for MTC. Among the 50 features, 16 are related to $mfcc$, which are widely used in the study of traditional audio signal processing such as voice and music, accounting for the largest proportion. These 16 features consist of five $mfcc$ features, six $delta-mfcc(dmfcc)$ features representing the audio rate, and five $delta-delta-mfcc(ddmfcc)$ features representing the acceleration of audio. The next most popular feature set within the timbre field is that of spectrum-related features. Seven spectrum-related features are extracted based on the Fast Fourier Transform, representing spectral distance, the coefficient of skewness, entropy, and so on. The timbre field also includes $rolloff$ and $brightness$, which estimate the number of high frequencies, and $spectral\_irregularity$, which shows fluctuations in the successive peaks of the spectrum. For the tonality field, 15 features were selected. The features related to $keystrength$ that distinguish the major and minor keys account for the largest proportion, with six features. The $chromagram$, which comprises four features, shows the distribution of energy along with the pitches or pitch classes. The rest are: three modal-related features, one $HarmonicChangeDetectionFunction$ feature, and one $keyclarity$ feature. Finally, in the rhythm field, a total of eight features were selected, including six tempo-related features and one each for $peak$, and $attack$. The six tempo-related features estimate the tempo by detecting periodicities from the event detection curve.

B. ANALYSIS BASED ON CORRELATION COEFFICIENT FOR EACH DOMAIN

For a more detailed analysis, we calculated the correlation coefficient (CC) between the selection frequencies of features from each dataset, where the selection frequency of each feature from a given dataset is the number of times that each feature is selected during ten iterative experiments. An array representing the selection frequency of each feature can be obtained for each dataset, and CCs between the arrays, which represent the similarity of FS between two datasets, can be calculated. Since there can be differences in tendency

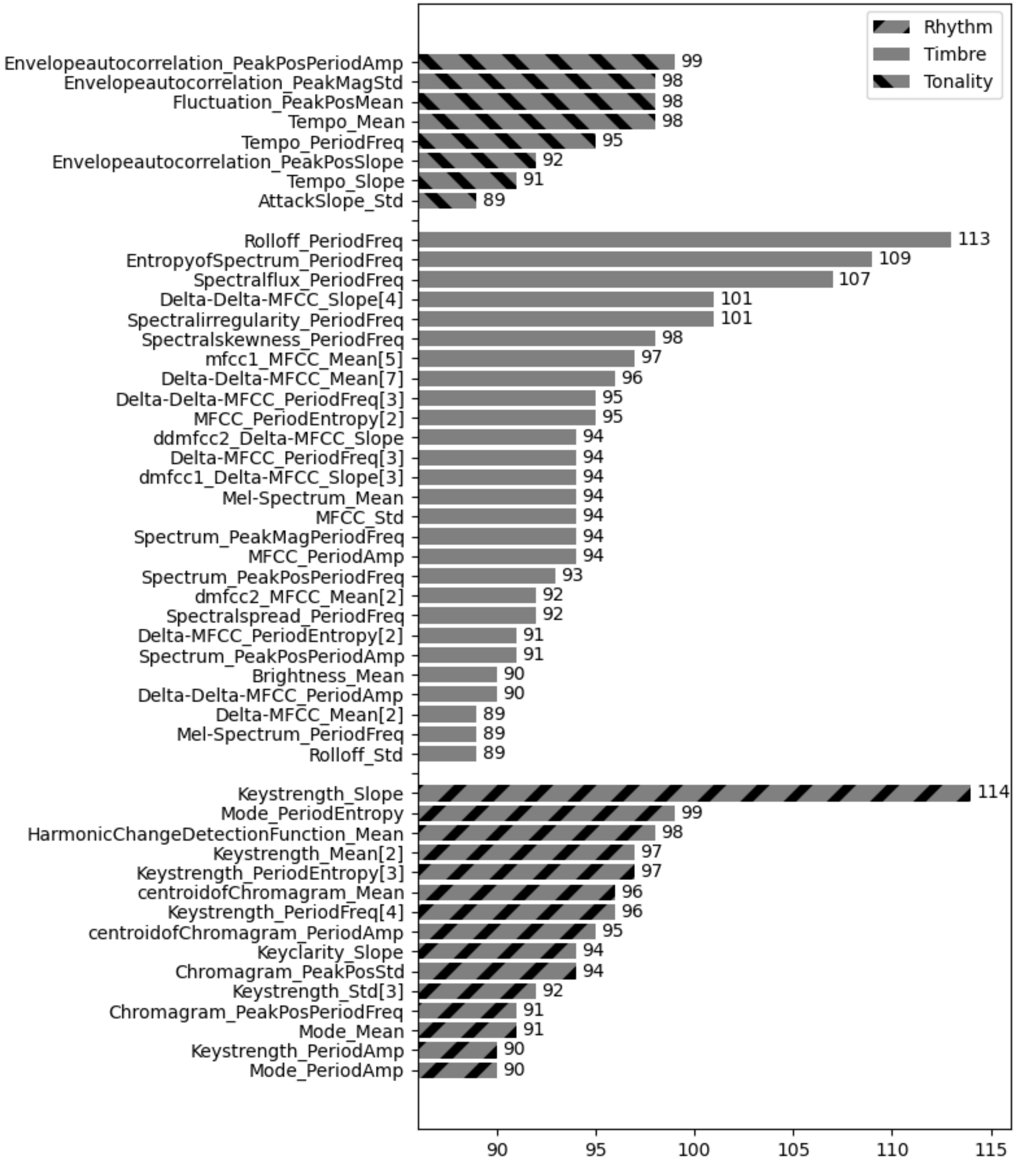
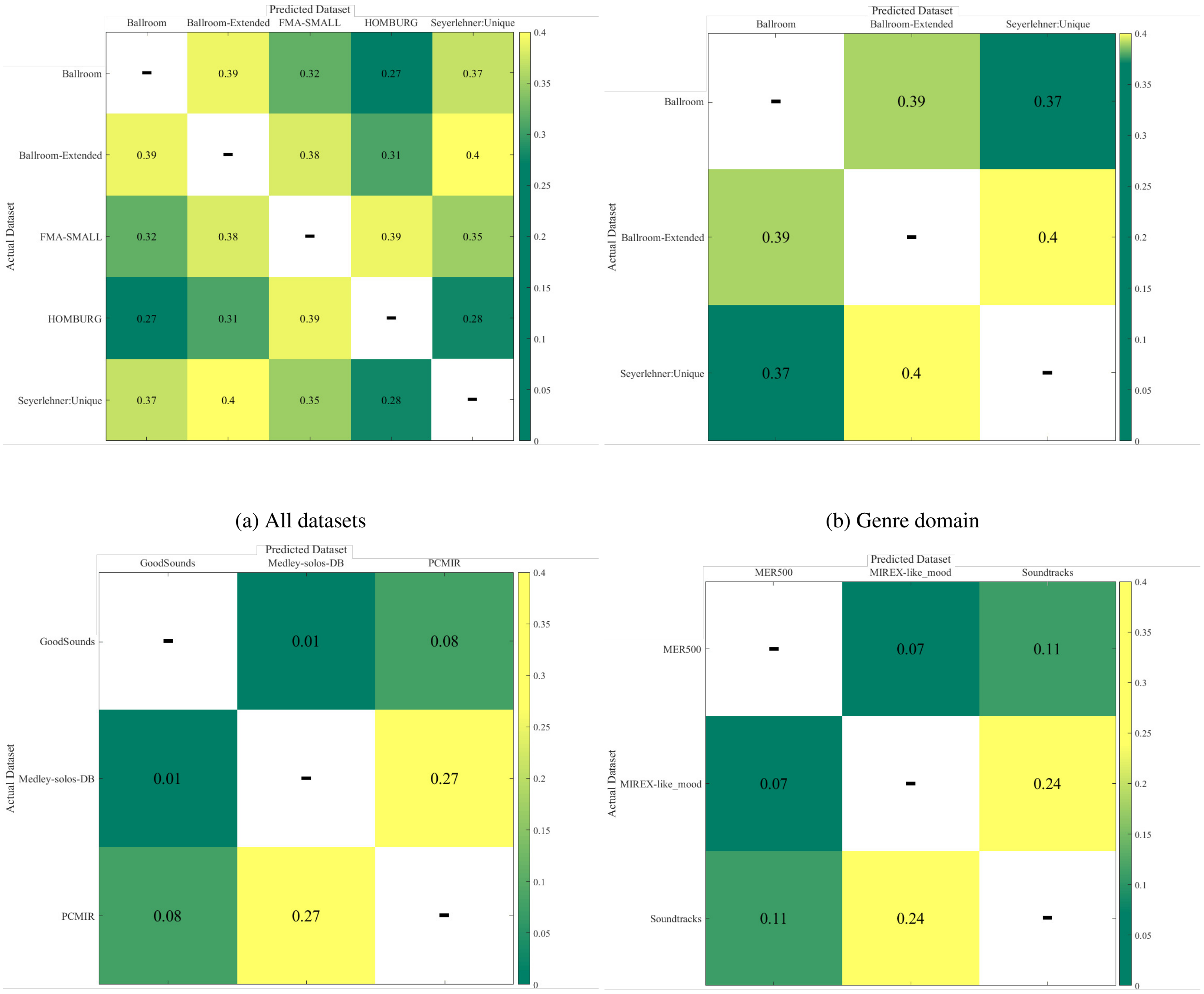


FIGURE 2: Top 50 frequently selected features and the number of times selected across the entire experiment (Features in Rhythm, Timber, Tonality field given by MIRtoolbox represented in different style.)



(a) All datasets

(b) Genre domain

(c) Instrument domain

(d) Emotion (Mood) domain

FIGURE 3: Dataset pairs with the highest correlation coefficient values.

TABLE 7: Correlation coefficients of selection frequency between dataset-pairs and the number of commonly-selected features among the top 50 frequently selected features for each dataset.

Coverage	Dataset 1	Dataset 2	Correlation Coefficient	# of commonly-selected features Rhythm	Timbre	Tonality
Overall	Ballroom-Extended	Seyerlehner:Unique	0.40	2	15	6
	Ballroom	Ballroom-Extended	0.39	3	7	1
	FMA-SMALL	HOMBURG	0.39	2	14	4
	Ballroom-Extended	FMA-SMALL	0.38	1	6	5
	Ballroom	Seyerlehner:Unique	0.37	4	14	2
Genre	Ballroom-Extended	Seyerlehner:Unique	0.40	2	15	6
	Ballroom	Ballroom-Extended	0.39	3	7	1
	Ballroom	Seyerlehner:Unique	0.37	4	14	2
Instrument	Medley-solos-DB	PCMIR	0.27	1	7	0
	PCMIR	Goodsounds	0.08	0	2	3
	GoodSounds	Medley-solos-DB	0.01	0	0	0
Emotion (Mood)	MIREX-like_mood	Soundtracks	0.24	0	7	3
	Soundtracks	MER500	0.11	1	3	2
	MER500	MIREX-like_mood	0.07	0	4	0

according to the domain of the datasets, we showed the CCs at an overall level (considering all 20 datasets), and by genre, instrument, and emotion (mood). Figure 3 shows the CC matrices between dataset pairs with the highest CCs, while Table 7 details our experiments. In the table, Coverage, Dataset, # of commonly selected features refer to the datasets that the experiment was conducted on, the dataset pair for which the CC value is calculated, and the common features in the top 50 features for each dataset. The dynamics and pitch fields are excluded from the column as they do not feature among the commonly selected features.

Our experiment showed that the top five dataset pairs with the highest CCs from the perspective of all datasets are: (Ballroom-Extended, Seyerlehner:Unique), (Ballroom, Ballroom-Extended), (FMA-SMALL, HOMBURG), (Ballroom-Extended, FMA-SMALL), and (Ballroom, Seyerlehner:Unique). The (Ballroom-Extended, Seyerlehner:Unique) dataset pair had the highest CC of 0.4, with 23 commonly selected features among the top 50 features (46%), the highest ratio compared to other dataset pairs. The 23 features consist of two in the rhythm field, 15 in the timbre field, and six in the tonality field. Of the 15 feature in the timbre field, 12 were $mfcc$-related features (two $mfcc$, three $dmfcc$, and seven $ddmfcc$). The six features of the tonality field consist of one $Mode\_Periodentropy$ and five $keystrength$-related features. The five overlapping $keystrength$-related features indicate that the two datasets consist of songs similar in major and minor. Next, we observe that the value of CCs for the (Ballroom, Ballroom-Extended) dataset pair is 0.39, which is the second highest CC in the experiment. However, only 11 features (three rhythm, seven timbre, and one tonality-related features) are common to both datasets. This pair contains one each of $mfcc$, $dmfcc$, and $ddmfcc$. Only $Keystrength\_Slope$, the most popular feature in the tonality field, was included in the commonly selected features. The (FMA-SMALL, HOMBURG) dataset pair are unrelated datasets from the viewpoint of the considered music excerpts, but ranked high with a CC value of 0.39. In this pair, 20 features are commonly selected (two rhythm, 14 timbre, and four tonality-related features), and 11 out of the 14 features in the timbre field are $mfcc$-related (three $mfcc$, three $dmfcc$, and five $ddmfcc$). The tonality field includes $Keystrength\_Slope$ and $Mode\_periodEntropy$ features. The (Ballroom-Extended, FMA-SMALL) dataset pair had a CC value of 0.38, with 12 commonly selected features (one rhythm, six timbre, and five tonality-related features). In addition to $Rolloff\_PeriodFreq$ and $Entropy\_PeriodFreq$, which were selected in all dataset pairs, this pair contains two $spectral\_irregularity$ and two $mfcc$ features. Regarding the tonality field, this pair contains the most diverse features among the pairs in the genre domain, such as $chromagram$, $mode$, and $keystrength$. Finally, the (Ballroom, Seyerlehner:Unique) dataset pair had the fifth-highest CC of 0.37, and the second most 20 features overlapped. It consists of four rhythm, 14 timbre, and two tonality fields. Two $mfcc$, three $dmfcc$, and four $ddmfcc$ features that are categorized into the timbre field are commonly selected.

In the genre domain, three dataset pairs—(Ballroom-Extended, Seyerlehner:Unique), (Ballroom, Ballroom-Extended), and (Ballroom, Seyerlehner:Unique)—yield the largest CC values. Since these three dataset pairs are a subset of the overall dataset pairs, we omitted a detailed analysis of the dataset pairs from the viewpoint of the genre domain. Next, in the instrument domain, we compared all three datasets simultaneously. The GoodSounds dataset selected only 27 features from the 888 features, so we analyzed the overlapping features from the Medley-solos-DB and PCMIR datasets. The (Medley-solos-DB, PCMIR) dataset pair had a CC value of 0.27 with only eight commonly selected features (one rhythm and seven timbre features). Except for the one $spectral\_irregularity$ and four $mfcc$-related features (two $mfcc$ and two $dmfcc$), the other commonly selected features were different compared to those for the genre domain, indicating differences in the tendency of selected features from the genre and instrument domains. Finally, we conducted our experiment on datasets in the emotion (mood) domain. The values of CC for (MIREX-like_mood, Soundtracks), (Soundtracks, MER500), and (MER500, MIREX-like_mood) are 0.24, 0.11, and 0.07, respectively. Since the (MIREX-like_mood, Soundtracks) dataset pair yields the best CC value—with the remaining two dataset pairs having very small values—we conducted our analysis on the (MIREX-like_mood, Soundtracks) dataset pair. In these two datasets, seven features of the timbre field are selected commonly, all of which were $mfcc$-related (one $mfcc$, one $dmfcc$, and five $ddmfcc$). Two of the three features in the tonality field were $keystrength$-related features.

C. REMARKS

In our experiment with multiple datasets, we observed which features were frequently selected by domain and what musical characteristics those features represent. First, $spectral\_irregularity$, $mfcc$, and $keystrength$-related features are important not only in the genre domain but also in the instrument and emotion (mood) domains. Features such as $Rolloff\_PeriodFreq$, Entropyofspectrum_PeriodFreq, $spectral\_irregularity$, $mfcc$, and $keystrength$ are commonly selected throughout the entire dataset. These common features mean that they play an important role in feature selection, and it is essential to include them when performing feature selection in a limited number. Second, the top five correlation coefficients are all from the genre domain. This implies that genre domain datasets have more overlapping features than the other two domains do. It is easier to know which features are important when classifying the genre domain.

Third, the instrument and emotion (mood) domains lack a rhythm or tonality field, and the overall number of overlaps is also small. Whereas in the genre domain, many similar features were selected in addition to frequently overlapping features, completely irrelevant features except for essential

TABLE 8: List of abbreviations used in Table 1.

Abbreviation	Description
ANN	Artificial Neural Networks
BPNN	Backpropagation Neural Networks
C4.5	One of the most popular decision tree variant
CBFS	Correlation-based Feature Selection
DT	Decision Tree
ELM	Extreme Learning Machine
FS	Feature Selection
GA	Genetic Algorithm
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
IBGHS	Improved Binary Global Harmony Search
IFS	Interactive Feature Selection
IG	Information Gain
IGA	Interactive Genetic Algorithm
kNN	k-Nearest Neighbor
LIBSVM	A Library for Support Vector Machines
MLP	Multi Layer Perceptron
NB	Naïve Bayes
NFC	Neuro-Fuzzy classifier
OAA	One-Against-All
PCA	Principal Component Analysis
RF	Random Forest
RR	Round Robin
SAHS	Self-Adaptive Harmony Search
SFS	Sequential Forward Selection
SMS-EMOA	S-Metric Selection Evolutionary Multi-Objective Algorithm
SVM	Support Vector Machines

TABLE 9: List of abbreviation used in Figure 2.

Abbreviation	Description
PeakPos	The abscissae position of the detected peaks, in sample index
PeakMag	Magnitude associated with each bin
Mean	The average along frames
Std	The standard deviation along frames
Slope	The linear slope of the trend along frames (the derivative of the line that would best fit the curve)
PeriodFreq	The frequency of the maximal periodicity detected in the frame-by-frame evolution of the values, estimated through the computation of the autocorrelation sequence. The first descending slope starting from zero lag is removed from this analysis, as it is not related to the periodicity of the curve.
PeriodAmp	The normalized amplitude of that main periodicity

features were extracted in the instrument and emotion (mood) domains. The results of the correlation coefficients are relatively lower than in the genre domain. Although they share essential features, it is important to share other similar features as well. Lastly, six features from the dynamics field are all related to root-mean-square. These features appear useless when classifying, even though the features were extracted through MIRtoolbox.

VI. CONCLUSIONS

In this study, to resolve existing difficulties in comparing the performance of various MTC methods, we created and evaluated 20 music datasets using the same feature extractor. As a result, all the 20 datasets have the same acoustic feature structure. To verify the true effect of FS on MTC tasks, we devised a novel evolutionary FS algorithm. Owing to the same acoustic feature structure of the 20 datasets, we were able to conduct additional experiments with multiple datasets, leading to valuable insights.

In the future, we plan on conducting an experiment based on convolutional neural network (CNN), which is which has been in the spotlight in music classification. We intend to utilize the findings of this study in the CNN research. In previous studies using CNN, the spatial information obtained through an analysis of music signals such as a melspectrogram, was considered an input to the CNN. Considering the important features found in this study, performance improvement is expected by providing higher quality input information to the CNN.

APPENDIX

Table 8 lists all of the abbreviations used in Table 1. Table 9 provides a detailed description of the y-axis in Figure 2.

REFERENCES

[1] P. Ginsel, I. Vatolkin, and G. Rudolph, “Analysis of structural complexity features for music genre recognition,” in *2020 IEEE Congress on Evolutionary Computation*. IEEE, 2020, pp. 1–8.

[2] K. Leartpantulak and Y. Kitjaidure, “Music genre classification of audio signals using particle swarm optimization and stacking ensemble,” in *2019 7th International Electrical Engineering Congress*. IEEE, 2019, pp. 1–4.

[3] G. Campobello, D. Dell’Aquila, M. Russo, and A. Segreto, “Neuro-genetic programming for multigenre classification of music content,” *Applied Soft Computing*, vol. 94, p. 106488, 2020.

[4] J. Lee, W. Seo, J.-H. Park, and D.-W. Kim, “Compact feature subset-based multi-label music categorization for mobile devices,” *Multimedia Tools and Applications*, vol. 78, no. 4, pp. 4869–4883, 2019.

[5] I. Vatolkin, “Improving supervised music classification by means of multi-objective evolutionary feature selection,” Ph.D. dissertation, 2013.

[6] C. N. Silla Jr, A. L. Koerich, and C. A. Kaestner, “A feature selection approach for automatic music genre classification,” *International Journal of Semantic Computing*, vol. 3, no. 02, pp. 183–208, 2009.

[7] O. Lartillot and P. Toiviainen, “A matlab toolbox for musical feature extraction from audio,” in *International conference on digital audio effects*, vol. 237. Bordeaux, 2007, p. 244.

[8] J. Gholami, F. Pourpanah, and X. Wang, “Feature selection based on improved binary global harmony search for data classification,” *Applied Soft Computing*, vol. 93, p. 106402, 2020.

[9] T. Özseven, “A novel feature selection method for speech emotion recognition,” *Applied Acoustics*, vol. 146, pp. 320–326, 2019.

[10] Z. Ma, “Detecting music genres from music data set: Select features by genetic algorithm, implement an artificial neural network, simplify it and compare with other works.”

[11] Y. S. Murthy and S. G. Koolagudi, “Classification of vocal and non-vocal segments in audio clips using genetic algorithm based feature selection (gafs),” *Expert Systems with Applications*, vol. 106, pp. 77–91, 2018.

[12] P. Kalapatapu, S. Goli, P. Arthum, and A. Malapati, “A study on feature selection and classification techniques of indian music,” *Procedia Computer Science*, vol. 98, pp. 125–131, 2016.

[13] M. Serwach and B. Stasiak, “Ga-based parameterization and feature selection for automatic music genre recognition,” in *2016 17th International Conference Computational Problems of Electrical Engineering*. IEEE, 2016, pp. 1–5.

[14] R. H. Zottesso, Y. M. Costa, and D. Bertolini, “Music genre classification using visual features with feature selection,” in *2016 35th international conference of the Chilean computer science society*. IEEE, 2016, pp. 1–6.

[15] E. Alexandre, L. Cuadra, S. Salcedo-Sanz, A. Pastor-Sánchez, and C. Casanova-Mateo, “Hybridizing extreme learning machines and genetic algorithms to select acoustic features in vehicle classification applications,” *Neurocomputing*, vol. 152, pp. 58–68, 2015.

[16] G. Kour and N. Mehan, “Music genre classification using mfcc, svm and bpnn,” *International Journal of Computer Applications*, vol. 112, no. 6, 2015.

[17] M. Wu and Y. Wang, "A feature selection algorithm of music genre classification based on relieff and sfs," in *2015 IEEE/ACIS 14th International Conference on Computer and Information Science*. IEEE, 2015, pp. 539–544.

[18] Y.-F. Huang, S.-M. Lin, H.-Y. Wu, and Y.-S. Li, "Music genre classification based on local feature selection using a self-adaptive harmony search algorithm," *Data & Knowledge Engineering*, vol. 92, pp. 60–76, 2014.

[19] M. Athani, N. Pathak, and A. U. Khan, "Dynamic music recommender system using genetic algorithm," *Int J Eng Adv Technol*, vol. 3, no. 4, pp. 230–232, 2014.

[20] I. Vatolkin, M. Preuß, G. Rudolph, M. Eichhoff, and C. Weihs, "Multi-objective evolutionary feature selection for instrument recognition in polyphonic audio mixtures," *Soft Computing*, vol. 16, no. 12, pp. 2027–2047, 2012.

[21] I. Vatolkin, M. Preuß, and G. Rudolph, "Multi-objective feature selection in music genre and style recognition tasks," in *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, 2011, pp. 411–418.

[22] G. V. Karkavitsas and G. A. Tsihrintzis, "Automatic music genre classification using hybrid genetic algorithms," in *Intelligent Interactive Multimedia Systems and Services*. Springer, 2011, pp. 323–335.

[23] H.-T. Kim, E. Kim, J.-H. Lee, and C. W. Ahn, "A recommender system based on genetic algorithm for music data," in *2010 2nd International Conference on Computer Engineering and Technology*, vol. 6. IEEE, 2010, pp. V6–414.

[24] C. N. Silla Jr, A. L. Koerich, and C. A. Kaestner, "Feature selection in automatic music genre classification," in *2008 Tenth IEEE International Symposium on Multimedia*. IEEE, 2008, pp. 39–44.

[25] C. N. Silla, A. L. Koerich, and C. A. Kaestner, "A machine learning approach to automatic music genre classification," *Journal of the Brazilian Computer Society*, vol. 14, no. 3, pp. 7–18, 2008.

[26] S. Doraisamy, S. Golzari, N. Mohd, M. N. Sulaiman, and N. I. Udzir, "A study on feature selection and classification techniques for automatic genre classification of traditional malay music." in *ISMIR*, 2008, pp. 331–336.

[27] S. Rho, E. Hwang, and M. Kim, "Music information retrieval using a ga-based relevance feedback," in *2007 International Conference on Multimedia and Ubiquitous Engineering*. IEEE, 2007, pp. 739–744.

[28] E. Alexandre, L. Cuadra, M. Rosa, and F. Lopez-Ferreras, "Feature selection for sound classification in hearing aids through restricted search driven by genetic algorithms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2249–2256, 2007.

[29] R. Fiebrink and I. Fujinaga, "Feature selection pitfalls and music classification." in *ISMIR*, 2006, pp. 340–341.

[30] H.-D. Kim, C.-H. Park, H.-C. Yang, and K.-B. Sim, "Genetic algorithm based feature selection method development for pattern recognition," in *2006 SICE-ICASE International Joint Conference*. IEEE, 2006, pp. 1020–1025.

[31] C. Xu, N. C. Maddage, X. Shao, F. Cao, and Q. Tian, "Musical genre classification using support vector machines," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings*, vol. 5. IEEE, 2003, pp. V–429.

[32] Y. Xue, B. Xue, and M. Zhang, "Self-adaptive particle swarm optimization for large-scale feature selection in classification," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 13, no. 5, pp. 1–27, 2019.

[33] F. Gouyon, S. Dixon, E. Pampalk, and G. Widmer, "Evaluating rhythmic descriptors for musical genre classification," in *Proceedings of the AES 25th International Conference*, vol. 196, 2004, p. 204.

[34] U. Marchand and G. Peeters, "The extended ballroom dataset," 2016.

[35] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, "1000 songs for emotional analysis of music," in *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*, 2013, pp. 1–6.

[36] M. Zentner, D. Grandjean, and K. R. Scherer, "Emotions evoked by the sound of music: characterization, classification, and measurement." *Emotion*, vol. 8, no. 4, p. 494, 2008.

[37] A. Aljanaki, F. Wiering, and R. C. Veltkamp, "Studying emotion induced by music through a crowdsourcing game," *Information Processing & Management*, vol. 52, no. 1, pp. 115–128, 2016.

[38] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "Fma: A dataset for music analysis," *arXiv preprint arXiv:1612.01840*, 2016.

[39] P. Knees, Á. Faraldo Pérez, H. Boyer, R. Vogl, S. Böck, F. Hörschläger, M. Le Goff *et al.*, "Two data sets for tempo estimation and key detection in electronic dance music annotated from user corrections," in *Proceedings of the 16th International Society for Music Information Retrieval Conference; 2015 Oct 26-30; Málaga, Spain.[Málaga]: International Society for Music Information Retrieval, 2015. p. 364-70.* International Society for Music Information Retrieval, 2015.

[40] J. Gillick, A. Roberts, J. Engel, D. Eck, and D. Bamman, "Learning to groove with inverse sequence transformations," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2269–2279.

[41] O. Romani Picas, H. Parra Rodriguez, D. Dabiri, H. Tokuda, W. Hariya, K. Oishi, and X. Serra, "A real-time system for measuring sound goodness in instrumental sounds," in *Audio Engineering Society Convention 138*. Audio Engineering Society, 2015.

[42] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, 2002.

[43] H. Homburg, I. Mierswa, B. Möller, K. Morik, and M. Wurst, "A benchmark dataset for audio classification and clustering." in *ISMIR*, vol. 2005, 2005, pp. 528–31.

[44] P. Cano, E. Gómez, F. Gouyon, P. Herrera, M. Koppenberger, B. Ong, X. Serra, S. Streich, and N. Wack, "Ismir 2004 audio description contest," *Music Technology Group of the Universitat Pompeu Fabra, Tech. Rep*, 2006.

[45] V. Lostanlen and C.-E. Cella, "Deep convolutional networks on the pitch spiral for musical instrument recognition," *arXiv preprint arXiv:1605.06644*, 2016.

[46] S. Malekzadeh, "Mer500." Kaggle, 2020. [Online]. Available: https://www.kaggle.com/makvel/mer500

[47] ——, "Micm music dataset." Kaggle, 2018. [Online]. Available: https://www.kaggle.com/dsv/193325

[48] X. Downie, C. Laurier, and M. Ehmann, "The 2007 mirex audio mood classification task: Lessons learned," in *Proc. 9th Int. Conf. Music Inf. Retrieval*, 2008, pp. 462–467.

[49] S. M. H. Mousavi, V. S. Prasath, and S. M. H. Mousavi, "Persian classical music instrument recognition (pcmir) using a novel persian music database," in *2019 9th International Conference on Computer and Knowledge Engineering*. IEEE, 2019, pp. 122–130.

[50] K. Seyerlehner, G. Widmer, and T. Pohle, "Fusing block-level features for music similarity estimation," in *Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx-10)*, 2010, pp. 225–232.

[51] T. Eerola and J. K. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music," *Psychology of Music*, vol. 39, no. 1, pp. 18–49, 2011.

[52] C. Salazar, "Tropical genres dataset." Kaggle, 2020. [Online]. Available: https://www.kaggle.com/carlossalazar65/tropical-genres-dataset

[53] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[54] J. Li, Y. Si, T. Xu, and S. Jiang, "Deep convolutional neural network based ecg classification system using information fusion and one-hot encoding techniques," *Mathematical Problems in Engineering*, vol. 2018, 2018.

[55] A. Cano, J. M. Luna, E. L. Gibaja, and S. Ventura, "Laim discretization for multi-label data," *Information Sciences*, vol. 330, pp. 370–384, 2016.

[56] B. L. Miller, D. E. Goldberg *et al.*, "Genetic algorithms, tournament selection, and the effects of noise," *Complex systems*, vol. 9, no. 3, pp. 193–212, 1995.

[57] W. Seo, D.-W. Kim, and J. Lee, "Generalized information-theoretic criterion for multi-label feature selection," *IEEE Access*, vol. 7, pp. 122 854–122 863, 2019.

[58] J. Park, M.-W. Park, D.-W. Kim, and J. Lee, "Multi-population genetic algorithm for multilabel feature selection based on label complementary communication," *Entropy*, vol. 22, no. 8, p. 876, 2020.

[59] M. Paniri, M. B. Dowlatshahi, and H. Nezamabadi-pour, "Mlaco: A multi-label feature selection algorithm based on ant colony optimization," *Knowledge-Based Systems*, vol. 192, p. 105285, 2020.

[60] M. A. Hall, "Correlation-based feature selection for machine learning," 1999.

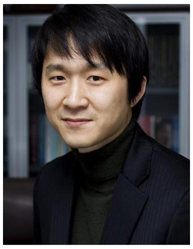
[61] T. K. Kim, "T test as a parametric statistic," *Korean journal of anesthesiology*, vol. 68, no. 6, p. 540, 2015.

[62] O. J. Dunn, "Multiple comparisons among means," *Journal of the American statistical association*, vol. 56, no. 293, pp. 52–64, 1961.

JONGHOON CHAE received the B.S. degree from Catholic University and the M.S. degree from Chung-Ang University, Seoul, South Korea. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering at Chung-Ang University, Seoul, South Korea. His research interests include artificial intelligence, data mining, feature selection and music classification.

SUNG-HYUN CHO received the B.S. degree from Korea University, Sejong, South Korea. He is currently pursuing the M.S. degree with the Department of Artificial Intelligence at Chung-Ang University, Seoul, South Korea. His research interests include recommendation system, evolutionary search and feature selection.

JAEGYUN PARK received the B.S. degree from Eulji University, Seongnam and the M.S. degree from Chung-Ang University, Seoul, South Korea. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering at Chung-Ang University, Seoul, South Korea. His research interests include continual learning, sensor-based activity recognition and feature selection.



DAE-WON KIM is currently a professor in the School of Computer Science and Engineering, Chung-Ang University, Seoul, South Korea. Prior to coming to Chung-Ang University, he did his Postdoctoral researcher, Ph.D., M.S. at Korea Advanced Institute of Science and Technology, and the B.S. at Kyungpook National University, Daegu, South Korea. His research interest includes advanced data mining algorithms with innovative applications to bioinformatics, music emotion recognition, educational data mining, affective computing, and robot interaction.

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 is currently an associate professor in the Department of Artificial Intelligence, Chung-Ang University. His research interests include machine learning, multilabel learning, model selection, and neural architecture search. In theoretical domain, he also studies classification, feature selection, and especially multilabel learning with information theory.