Six fature source types for multi modal music classification

Sunday, 13 November 2022 17:21

TISMIR

RESEARCH

Multi-Objective Investigation of Six Feature Source Types for Multi-Modal Music Classification

IGOR VATOLKIN
CORY MCKAY
*Author affiliations can be found in the back matter of this article

]u[ubiquity pres

ABSTRACT

CORRESPONDING AUTHOR:

ABSTRACT

Every type of musical data (audio, symbolic, lyrics, etc.) has its limitations, and cannot always capture all relevant properties of a particular musical category. In contrast to more typical MIR setups where supervised classification models are trained on only one or two types of data, we propose a more diversified approach to music classification and analysis based on six modalities: audio signals, semantic tags inferred from the audio, symbolic MIDI representations, album cover images, playlist co-occurrences, and lyric texts. Some of the descriptors we extract from these data are low-level, while others encapsulate interpretable semantic knowledge that describes melodic, rhythmic, instrumental, and other properties of music. With the intent of measuring the individual impact of different feature groups on different categories, we propose two evaluation criteria based on "non-dominated hypervolumes": multi-group feature "importance" and "redundancy". Both of these are calculated after the application of a multi-objective feature selection strategy using evolutionary algorithms, with a novel approach to optimizing trade-offs between both "pure" and "mixed" feature subsets. These techniques permit an exploration of how different modalities and feature types contribute to class discrimination. We use genre classification as a sample research domain to which these techniques can be applied, and present exploratory experiments on two disjoint datasets of different sizes, involving three genre ontologies of varied class similarity. Our results highlight the potential of combining features extracted from different modalities, and can provide insight on the relative significance of different modalities and features in different contexts.

CORRESPONDING AUTHOR:
Igor Vatolkin
TU Dortmund University, Department of Computer Science, Germany
igor.vatolkin@udo.edu

KEYWORDS:
Multi-modal data; multi-objective feature selection; supervised music classificatio

TO CITE THIS ARTICLE:
Vatolkin, I., & McKay, C. (2022 Multi-Objective Investigation of Six Feature Source Types for Multi-Modal Music Classification. *Transactions of the International Society for Music Information Retrieval*, 5(1), pp. 1–19. DOI: *https://do org/10.5334/tismir.67*

1. INTRODUCTION

Musical information can manifest in a variety of different modalities,[1] each of which can be of greater or lesser interest to different types of domain experts, and with respect to different purposes. For example, a musician specializing in orally transmitted (as opposed to written) musics will likely be particularly interested in audio representations, while a musicologist specializing in Western Renaissance music will more probably focus on symbolic representations, and textual representations of vocal content could be important to both. A historian or sociologist, on the other hand, might be particularly interested in images of illuminations on ancient manuscripts, or in publicity videos used in contemporary popular music.

Some modalities encapsulate information that cannot be found in certain other modalities, and some overlap at least partially in what they can reveal. Even in the latter case, certain kinds of information can be more easily extracted from certain modalities than others, such as the segmentation of individual melodies found in polyphonic music represented in symbolic compared to audio formats. This suggests significant potential gains in combining different modalities in a variety of MIR salience, and some not, some of which may be of use in improving automatic classification performance, and some not; either way, it is our hope that such patterns can provide directed motivation for further multidisciplinary investigations. Features extracted from diverse types of data may both improve classifier performance and reveal insights on the music itself.

Supervised genre recognition in particular is chosen as the sample subject of our experiments, as it has been a long-standing area of MIR interest. Multi-modal research is also of particular relevance to genre, since it allows researchers to explore intersections between different dimensions of the musical experience, both directly musical (e.g., symbolic music features) and not (e.g., album art features). It is important to acknowledge, however, that genre classification is a complex, problematic area. Genre labels can be fuzzily defined, and the practice of limiting pieces to single labels can be problematic (McKay and Fujinaga, 2006). Furthermore, genre can be difficult to isolate unambiguously in classification experiments from confounding characteristics (Sturm, 2013a), which has implications for the meaningfulness of measures like classification accuracy. Sturm further notes that the errors machine classifiers make are identifiably different from misclassifications a human would make

gains in combining different modalities in a variety of MIR research areas.

It would therefore be useful to have a framework for quantitatively exploring the relative extent to which various modalities, and the feature types that can be extracted from them, are meaningful to given musical research problems. Making progress towards developing a general framework of this kind is the core goal of this article. Unfortunately, testing and deploying such a framework is currently quite difficult, given that, as discussed below, there are very few existing datasets combining reliably matched sources belonging to more than a few different modalities and, furthermore, very little work is available in the MIR literature discussing methodologies for meaningfully comparing the relative significance of different musical modalities in general terms.

This article seeks to contribute to the ultimate long-term development of this kind of framework as follows: 1) presenting novel non-dominated hypervolume-based approaches for measuring feature importance and redundancy; 2) assembling two combined datasets, each involving six modalities; and 3) using these to explore the potential of multi-modal MIR research and highlight the need for improved multi-modal datasets. Automatic music classification provides a good domain for performing such explorations, as there is MIR interest in it, and a range of different types of musical information are available. We focus in particular on exploring the ability of our novel methodologies to reveal statistical patterns associated with modalities and feature types, some of which may have musicological or psychoacoustic

different from misclassifications a human would make, and presents this as an indication that they are failing to properly model genre. We are certainly not claiming in this paper to solve such deep issues with genre; our goal is simply to use genre classification as an interesting MIR domain to illustrate the multi-modal optimization techniques we propose.

In our study, we combined six different source types in each of two multi-modal datasets we assembled. To compare the influences of these six sources, we introduced a multi-objective approach to selecting features extracted from them: our first function seeks to minimize balanced classification error, and the second attempts to maximize (or minimize) the proportion of features selected from particular feature groups.

This methodology allows us not only to compare source types by identifying high-performing "pure" features subsets for each of the six data types (i.e. with no features from any of the other data types), but also to observe the extent to which classification error can be reduced further when features from other source types are also allowed. We also experimented with minimizing both the classification error and the proportion of features from each individual group, so as to explore whether good classification performance can be achieved *without* features of the group. We propose *normalized multi-group feature importance* and *normalized multi-group feature redundancy* as two formal measures for comparing and analyzing information extracted from different modalities (e.g., audio vs. lyrics), as well as from feature sub-groups extracted from individual modalities (e.g., timbre vs. pitch).

The novel non-dominated hypervolume approach and two associated new general measures (importance and redundancy) for exploring the relative significance of and interactions between different feature groups in arbitrary problem domains are key original contributions of this work, as are the two datasets we expanded to each now consist of six modalities (more than any existing MIR dataset, and substantially more than almost all). The results and analyses of the exploratory experiments we performed using these measures and datasets to examine the potential of comparative multi-modal research are also novel contributions, and they demonstrate the important need for more diverse shared multi-modal datasets.

This paper is structured as follows: Section 2 identifies

(2002) distinguish between features associated with timbral texture, rhythm, and pitch, and Saari et al. (2011) consider features based on dynamics, rhythm, pitch, harmony, timbre, and structure. In both cases, however, the audio signal served as the only source type.

Other studies have been conducted on features extracted from two different source types, especially audio and lyrics, with varying outcomes. Dhanaraj and Logan (2005) reported that the combination of these sources did not improve hit song prediction, and Zangerle et al. (2018) found that audio features alone were better at playlist prediction than the two combined. Other studies, however, showed improvements in classification performance for both genre (Neumayer and Rauber, 2007; Mayer and Rauber, 2010) and mood (Laurier et al.,

related work on multi-modal music classification and feature selection. Section 3 describes the multi-objective feature selection and binary classification methodologies we employed, and Section 4 focuses on the measures we devised to derive meaning from our experiments. Section 5 provides an overview of the datasets, modalities, features, and partitioning methodology used in our study. Experimental results are discussed in Section 6, and concluding remarks are presented in Section 7.

2. RELATED WORK

2.1 MULTI-MODAL MUSIC CLASSIFICATION

A common MIR classification approach is to start with the audio signal and extract hand-crafted features from it, as proposed by Tzanetakis and Cook (2002),[2] or, more recently, to let neural networks model the audio descriptors themselves (Costa et al., 2017; Sigtia and Dixon, 2014). Other studies have used features extracted from symbolic representations (Dannenberg et al., 1997), lyrics (Logan et al., 2004), or tags (Lamere, 2008). Each musical data source type has its advantages and limitations. Audio features can be extracted from a track independently of its popularity. User tags, however, may be noisy for less popular or new music, due to the "cold start" problem (Celma, 2010). It is still hard to reliably extract interpretable semantic information like chord progressions from polyphonic audio signals. While high-level features can be reliably extracted from symbolic data, such representations often exclude properties describing individual performance interpretations or studio processing. Furthermore, neither audio nor symbolic data specify cultural information explicitly, such as a piece's country of origin or language; mining lyrics and cultural data can reveal important properties.

For these and other reasons, many publications on music classification have combined features of different types. However, in much of the literature these features reflect semantic properties based on music theory or auditory perception that are all extracted from the same type of musical data. For instance, Tzanetakis and Cook 2008; Hu et al., 2017) when features from both source types were combined. Combining audio and symbolic features also improved genre classification (Cataltepe et al., 2007), as did combining audio features with visual descriptors from music videos (Schindler, 2019). Grouping audio and tag features improved mood classification (Bischoff et al., 2009). Audio and image features have also been combined for mood prediction (Dunker et al., 2008).

Far fewer publications have combined three or more feature sources. McKay and Fujinaga (2008) and McKay et al. (2010) observed a common, but not universal, improvement in genre classification performance when features extracted from audio, symbolic, cultural, and lyric data were combined. The combination of audio, symbolic, and lyric features improved emotion detection (Panda et al., 2013). A more recent approach improved genre classification by learning features from audio, text, and images using deep neural networks (Oramas et al., 2017; 2018). McFee and Lanckriet (2012) generated playlists by combining audio, lyrics, social tags, collaborative filtering, and additional metadata.

Other works broadly discuss the incorporation of various modalities to music classification and provide overviews of related studies (Mayer and Rauber, 2010; Jannach et al., 2017). Knees and Schedl (2013) present an overview of contextual modalities beyond audio and the score, distinguishing between text retrieval, co-occurrences, and user ratings. Simonetta et al. (2019) provide a partial survey and discussion of multi-modal MIR research.

General concern about evaluation of music classification is an important theme in MIR, and Sturm has been particularly influential in this area (Sturm, 2012a; b; 2013a; b). Not only is it important to employ meaningful measures and considerations of statistical significance, one should also consider essential issues with ground truth and underlying ontological complexity, as well as unique elements of each problem domain. As emphasized by Sturm (2012a), experiments should ideally go beyond focusing on simple classification

measures to also considering aspects like generalizability to diverse datasets or robustness to data transformations that would not alter human classifications.

There are a few important publicly accessible multi-modal datasets, but much work remains to be done.

Mayer et al. (2010) applied FS to features extracted from both audio and MIDI files within a Cartesian ensemble of classifiers, and the number of feature dimensions was reduced to less than 4% of the original, while maintaining good performance for two of four

Copyright restrictions pose a problem, especially but not only with respect to audio, which is typically only available as short previews accessible via third-party APIs or as pre-extracted features. Orio et al. (2011) introduced a benchmark with audio features, user tags, web pages, and expert labels. The MuMu dataset (Oramas et al., 2017) compiles audio, review texts, and album cover images. DALI (Meseguer-Brocal et al., 2018) contains audio files, lyrics, and vocal note annotations. Bogdanov et al. (2019) presented the AcousticBrainz dataset, which contains low-level audio features together with high-level descriptors (moods, vocals, etc.) generated by pre-trained classifiers.

2.2 MUSICAL FEATURE SELECTION

The number of features involved increases as modalities are added, some of which may be irrelevant or redundant. Feature selection (FS) can remove unnecessary descriptors and achieve better classification performance, while also reducing computation and storage demands. Ideally, one desires FS methods that consider feature combinations (not just features individually), or that can identify especially relevant or interpretable properties of categories.

Guyon et al. (2006) provide a general overview of feature selection, and Fujinaga (1998) published early work applying FS to music classification, where a genetic algorithm was used to estimate weights for a k-nearest neighbor instrument identifier. A later contribution emphasized the requirement of using an independent test set to avoid FS overfitting (Fiebrink and Fujinaga, 2006).

Doraisamy et al. (2008) compared 7 feature selection methods in combination with 18 classifiers, and found that classification could be improved in traditional Malay music genre recognition involving a relatively small dataset of 191 pieces. However, the general success of FS is not always evident. For example, Huang et al. (2014) reported genre classification accuracy increases with all tested FS approaches, but for a larger Latin genre dataset with more than 3,000 pieces Silla Jr. et al. (2009) observed that the success of FS was dependent on the choice of classification method, and in some cases accuracy did not increase. Two other studies on mood recognition (Saari et al., 2011) and genre recognition (Lim et al., 2012) found an increase in performance after FS, and estimated curves modeling the impact of increasing feature set size: after an initial strong growth in accuracy involving small numbers of high-performing features, the accuracy then stagnated or decreased for larger feature sets.

datasets. Panda et al. (2013) identified relevant mood recognition features from audio, scores, and lyrics by applying a Relief algorithm for feature weight estimation.

Multi-objective feature selection (MOFS) is another promising direction, as several important and less correlated evaluation criteria may be in conflict (precision, recall, computing costs, robustness, interpretability, etc.). Vatolkin et al. (2011) were the first to apply MOFS to supervised music classification, by minimizing the classification error and number of features for genre and style recognition by means of an evolutionary algorithm. This setup was later extended to the optimization of all pairs between seven different evaluation criteria (Vatolkin, 2015) and the measurement of "album effect" (modelling albums rather than genres) (Vatolkin et al., 2015).

Our broader literature survey indicates that the substantial majority of MIR FS focuses on applying FS to audio data only, and optimizes feature sets with respect to only one measure. In Section 4, we propose two novel optimization scenarios for MOFS, which are designed to measure the classification impact of features, and compare different modalities or feature subgroups.

3. FEATURE SELECTION METHODOLOGY

3.1 MULTI-OBJECTIVE EVOLUTIONARY ALGORITHM

An important step in FS is the evaluation of feature subsets. A single validation measure like classification error or F-score is typically calculated, but in practice several potentially conflicting measures can be relevant. Multi-objective feature selection is formally defined by Vatolkin et al. (2015) as the search for an optimal feature vector $\mathbf{q}^*$:

$$\mathbf{q}^* = \underset{\mathbf{q}}{\operatorname{argmin}}[m_1(\mathbf{y}_L, \mathbf{y}_P(\mathbf{x},\mathbf{q}), \Phi(\mathbf{x},\mathbf{q})), \ldots, m_O(\mathbf{y}_L, \mathbf{y}_P(\mathbf{x},\mathbf{q}), \Phi(\mathbf{x},\mathbf{q}))], \quad (1)$$

where $\mathbf{x}$ is the original complete feature vector; $\mathbf{q}$ is a binary vector containing an entry for every feature in $\mathbf{x}$, indicating which features are and are not selected; $\Phi(\mathbf{x},\mathbf{q})$ is a given set of selected features; $\mathbf{y}_L$ is a vector indicating ground truth class labels; $\mathbf{y}_P$ is the vector of class labels predicted using a given $\Phi(\mathbf{x},\mathbf{q})$, and $m_1,\ldots,m_O$ are the objective functions to minimize.[3]

This study focuses on investigating the relative classification impact of feature subsets drawn from different types of data. Although single-objective FS could be used to just minimize classification error for subsets drawn from each group independently, we are

also interested in the potential effects of combining features drawn from different groups to ultimately improve performance. This can be addressed via the simultaneous optimization of two criteria, repeated separately for each of the six feature type groups. First, we aim to minimize the balanced relative binary classification error m_e (mean of the relative error on pieces belonging to and not belonging to a given class). Second, we aim to maximize[4] the proportion g_k of the features from the source type k currently being considered, $k \in \{1,...,6\}$. In other words, we try to simultaneously keep a given feature set as "pure" as possible (i.e. excluding as many features from the other five groups as we can), while at the same time keeping the classification error as small as possible. As explained in Section 4 with respect to feature group "importance," this optimization strategy provides a framework for comparing the relative classification efficacy of feature types (each potentially extracted from a different source type) not only in terms of how effective each feature type is in isolation, but also in terms of the extent to which the addition of feature types drawn from additional modalities might improve classification performance, among other things. To our best knowledge, this approach has not previously been applied to multi-modal music classification or feature evaluation.

Feature selection is an NP-hard optimization task (Amaldi and Kann, 1998), and resolving competing solutions in multi-objective selection can be harder still (e.g., a pure feature set with larger error vs. a mixed set with smaller error). Evolutionary algorithms (Zitzler, 2012) are a good choice for this task, particularly for larger feature sets (Kudo and Sklansky, 2000). The stochastic aspect of the evolutionary process helps to overcome local optima, and a population of diverse co-existing solutions permits complex explorations of trade-offs between the objectives. Also, many other FS methods rank features individually, thereby under-emphasizing situations where several individually irrelevant or correlated features become relevant in combination (Weihs et al., 2017, pp. 391–392).

This study uses a variant of the multi-objective evolutionary algorithm described by Vatolkin et al. (2015), which sought to minimize both classification error and the number of selected features, with a few adjustments necessitated by the different goals here. We used the results of preliminary experiments to adjust hyperparameters to start with more varied initial solutions. The initial feature rate hyperparameter I_{FR}, which controls the expected proportion of selected features before evolution begins, was set to 0.5. For each of the $p = 50$ initial feature sets, there is a 10% chance of 50% chance of being selected (because $I_{FR} = 0.5$). In the third case, we use another random hyperparameter $D \in [0,1]$ to create different proportions of features in the initial feature subset. Later, during each evolutionary iteration, we generate new feature sets from randomly selected parent solutions with mutation, which randomly switches some features on and off. The probability of a bit flip for each feature is set to γ/N, where N is the overall number of features and the mutation strength $\gamma = 64$ (based on initial experiments).

3.2 FITNESS FUNCTION AND CLASSIFICATION

Feature sets are evaluated by a fitness function that considers both m_e and g_k, based on individual hypervolume contributions, as explained in Section 4. The most fit μ solutions are selected from the parent and offspring populations, and are carried over to comprise the next generation. The number of generations was set to 2500, as a good compromise between runtime demands and convergence behavior.

Binary classification[5] experiments were performed for each genre using random forests (Breiman, 2001), as implemented by the WEKA library (Witten et al., 2016). Fortunately, random forests are robust to overfitting, an important concern given our large number of features and relatively small datasets (Reunanen (2003) and Fiebrink and Fujinaga (2006) provide good discussions of overfitting in FS). As observed by Hastie et al. (2009, p.596): "when the number of relevant variables increases, the performance of random forests is surprisingly robust to an increase in the number of noise variables." We used random forest decision tree ensembles (100 trees per forest) set up so that each tree only considered $\log_2|\Phi(\mathbf{x},\mathbf{q})| + 1$ randomly selected activated features in a given feature set. So, even if an individual feature is disproportionately effective due to bias in a small training set, the danger of that feature significantly impacting the overall classification model is reduced. As more multi-modal data become available in the future, it will permit experimentation with alternative classifier types, such as deep neural networks.

3.3 CONTEXT

It should be emphasized at this point that the purpose of the methodology described above is not to demonstrate how one might find optimal or near-optimal feature sets for specific applied classification problems. This is why results are not compared to alternate dimensionality reduction methodologies, such as forward-backward selection or factor analysis. Rather, the goal here is to investigate the extent to which information drawn from different types of musical data or feature types can be

selecting only "pure" features from the current feature type group, a 10% chance of selecting only "other" features, and an 80% chance of selecting a mix of the two. In the first and second cases, each feature has a usefully combined.

The optimization scenarios designed and implemented in this article (see Section 4 for details) are largely aligned with the "Classify" and "Features" approaches identified

in Sturm's experimental design ontology (Sturm, 2012a p.4). Additional approaches from this ontology could be usefully implemented as MOFS evaluation measures in future research; for example, the ability to generalize to diverse datasets could be maximized ("Generalize"), as could robustness to data tranformations ("Robust").

4. FEATURE TYPE EVALUATION METHODOLOGY

4.1 HYPERVOLUME-BASED COMPARISON OF FEATURE GROUPS

A feature set $\mathbf{q}_1$ can be said to *dominate* feature set $\mathbf{q}_2$ ($\mathbf{q}_1 \prec \mathbf{q}_2$) when it is not worse w.r.t. all O evaluation measures, and is better for at least one measure:[6]

$$\begin{aligned} &\forall i \in \{1,\ldots,O\} : m_i(\mathbf{q}_1) \leq m_i(\mathbf{q}_2) \text{ and} \\ &\exists j \in \{1,\ldots,O\} : m_j(\mathbf{q}_1) < m_j(\mathbf{q}_2). \end{aligned} \tag{2}$$

The ϕ best compromise feature sets $\mathbf{q}_1,\cdots,\mathbf{q}_\phi$, which are not dominated by any other feature sets, comprise the *non-dominated front*, and are characterized by their *dominated hypervolume* (Weihs et al., 2017, p. 278):

$$H(\mathbf{q}_1,\ldots,\mathbf{q}_\phi;\mathbf{r}) = \Lambda_d\left(\bigcup_{i=1}^{\phi}[\mathbf{q}_i,\mathbf{r}]\right), \tag{3}$$

where $\mathbf{r} \in \mathbb{R}^o$ is a reference point corresponding to a worst possible feature set and Λ_d is the volume of a set in $\mathbb{R}^o$. Put simply, the hypervolume contains all possible feature sets that are dominated by feature sets in the non-dominated front.

Given a theoretical *ideal* feature set $\mathbf{q}_{ID}$ with the best individual values of $m_1,\ldots,m_o$ from all non-dominated feature sets, we define h as:

$$h = H(\mathbf{q}_{ID};\mathbf{r}) - H(\mathbf{q}_1,\ldots,\mathbf{q}_\phi;\mathbf{r}). \tag{4}$$

The meaning of h is represented by the shaded areas in ***Figures 1*** and ***3***, each of which correspond to different m_i optimization approaches, as explained below.

4.2 MULTI-GROUP IMPORTANCE AND REDUNDANCY

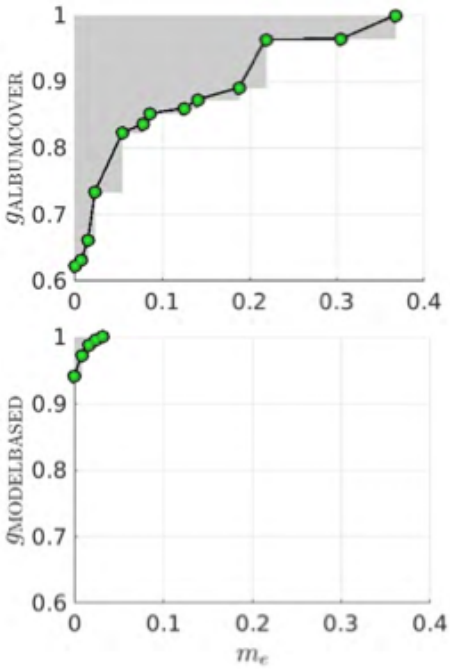


Figure 1: Examples of non-dominated feature sets (connected circles) after feature selection in an experiment on Rock music (see Section 5) using two criteria: the first is binary classification error m_e, which is minimized, and the second is the proportion g_k of the features from the k-th group, which is maximized. The share of album cover features is maximized in the upper sub-figure, and the share of model-predicted semantic tags is maximized in the lower sub-figure.

of "other" features is required to achieve the very best m_e in this optimization scenario: $m_e = 0$ is achieved with a feature set for which 38% (100%–62%) of features do not belong to the album cover group,[7] compared to 6% (100%–94%) for the model-based features.

So, a smaller h value (as represented by the shaded area in ***Figure 1***) resulting from this optimization strategy suggests that the feature type in question is better at identifying the given class. We can say that such a feature type is more "important" by defining its *multi-group importance* i_h as:

$$i_h = 1 - h(m_e \downarrow, g_k \uparrow). \tag{5}$$

REDUNDANCY

Recall from Section 3 the strategy of minimizing m_e and maximizing g_k, for the purpose of obtaining feature sets that reduce classification error while at the same time keeping the feature group as "pure" as possible. Here the reference point $\boldsymbol{r}$ has $m_e = 1$ and $g_k = 0$. An example of the results of applying this optimization strategy is shown in ***Figure 1***. It can be seen from this figure that album cover features (the upper sub-figure) perform much worse than model-based features (the lower sub-figure) with respect to Rock music. First, the minimum m_e^* for a pure album cover feature subset is only 0.37, compared to 0.03 for a pure model-based subset (pure feature subsets, where $g_k = 1$, are at the upper right). Second, a larger proportion

other features does not reduce classification error, e.g., ***Figure 2(a)***. This will happen in practice if a feature group encapsulates all accessible relevant information. However, the opposite case, where $i_h = 0.0$ and $h = 1.0$, is very unlikely to occur in practice, as are other cases with very low i_h. To illustrate this, consider the example shown in ***Figure 2(b)***, with a very low importance of $i_h = 0.0615$ and $m_e = 1$ for the pure feature set; for this to happen, the binary classifier would have to misclassify every track using the pure feature group, despite having a 0.5 chance of correctly classifying an instance simply by guessing. m_e would also need to decrease substantially when other features types are added. An example of a more realistic worst case, what we call a "completely non-important" feature group, is shown in ***Figure 2(c)***: here the pure feature group effectively consists of random noise, since $m_e = 0.5$ for it, and m_e decreases linearly as other features are added, with $m_e = 0.0$ when the pure group is excluded entirely. This more realistic worst case (for a binary classifier) leads to $\lim_{\phi\to\infty} h = 0.25$ and $\lim_{\phi\to\infty} i_h = 0.75$ (recall that ϕ is the number of non-dominated solutions). So, since we expect no values of i_h below 0.75, we can normalize i_h between 0.75 and 1, setting any values below 0.75 (i.e. worse than random) to 0; this results in the *normalized multi-group importance*:

$$I_h = max\left\{\frac{i_h - 0.75}{0.25}, 0\right\}. \quad (6)$$

An interesting alternative approach is to minimize both m_e and g_k: this allows one to investigate the extent to which it is possible to achieve high classification performance *without* the feature group in question. ***Figure 3*** demonstrates an example of this approach, with

$h(m_e \downarrow, g_k \uparrow)$ refers to h after minimization of m_e and maximization of g_k.

It is important not to confuse this notion of multi-group "importance" with the concept of feature "relevance" during feature selection; as noted by Kohavi and John (1997), a feature is "relevant" only if its removal decreases classification performance. Since the focus of this work is on developing methodologies for measuring the relative significance of different types of data and features to various classes, and not just on optimizing classification performance, this notion of "importance" is better suited to our needs.

A best-possible importance of $i_h = 1$ is achieved when $h = 0$ and extending the pure feature group with classification performance. We can thus define a feature group's *multi-group redundancy* r_h as:

$$r_h = 1 - h(m_e \downarrow, g_k \downarrow). \quad (7)$$

$h(m_e \downarrow, g_k \downarrow)$ refers to h after minimization of both m_e and g_k. Just as i_h can be normalized to I_h, r_h can be scaled to the *normalized multi-group redundancy*:

$$R_h = max\left\{\frac{r_h - 0.75}{0.25}, 0\right\}. \quad (8)$$

Different goals are involved in identifying feature groups with high I_h or low R_h values. Multi-group importance helps identify features whose classification performance cannot be significantly improved by adding features from other groups. For example, if a music class is characterized by distorted guitars or other spectrally noisy instruments, but all other classes under consideration involve only spectrally clean harmonic instruments, then

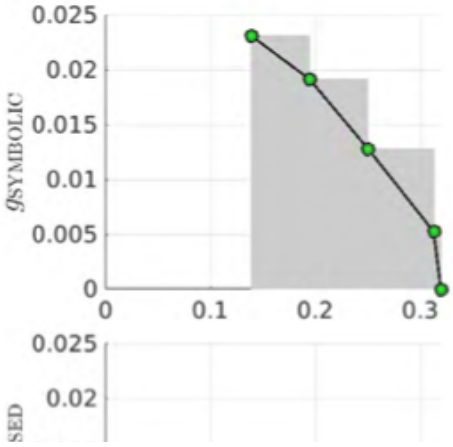


Figure 3 demonstrates an example of this approach, with respect to Traditional Blues music. The upper sub-figure shows that the best $m_e = 0.32$ achieved without symbolic features can be reduced to $m_e = 0.14$ when the feature set consists of 2.31% symbolic features. The bottom sub-figure shows that the impact of model-based features is less, since m_e is only reduced from 0.13 to 0.09 by allowing them.

So, under this optimization strategy a lower h-value suggests that a feature group is more redundant, since allowing its features has a reduced impact on

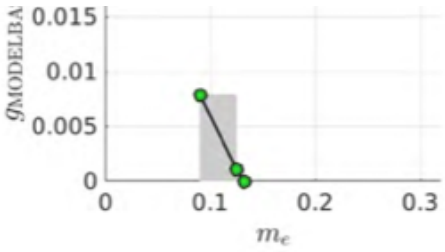


Figure 3: Binary classification performance of symbolic (top) and model-based (bottom) features on Traditional Blues music (see Section 5), based on minimization of both m_e and g_k.

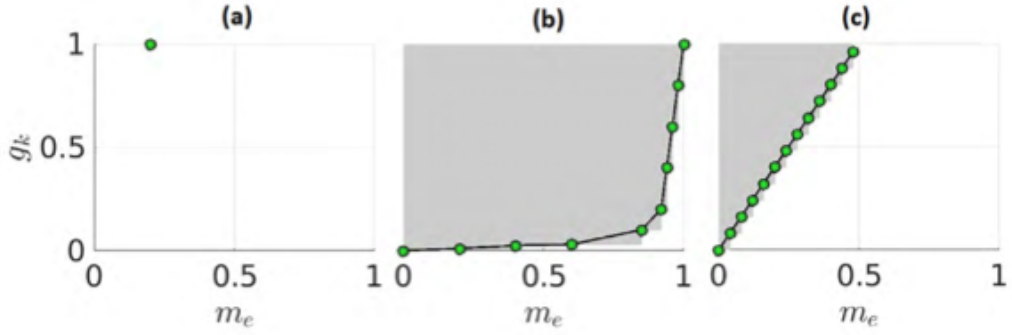


Figure 2: Theoretically possible non-dominated fronts for the minimization of m_e and maximization of g_k.

both features based on instrumentation and features based on spectral properties of the audio signal will be "important" (and potentially musically meaningful) in describing this class. Multi-group redundancy, on the other hand, helps identify feature groups that can be omitted in order to create more robust and efficient classification models that use as few features as possible. In the aforementioned example, a model that classifies tracks based only on instrumentation or based only on spectral properties might be entirely sufficient.

5 DATASETS AND MODAL FEATURE GROUPS

Now that we have introduced our feature selection methodology and optimization strategies for comparing different feature types in general terms, we can turn our attention to the particular datasets, modalities, and features we used to conduct our experiments, as well as our data partitioning approach. As an overview, we combined six different source types to create one dataset consisting of 2.803 features extracted from 1.575 pieces of expression, like the number of segment changes. Finally, the ground truth user-submitted genre labels are noisily and inconsistently annotated.

SLAC (McKay et al., 2010) consists of 250 pieces of music, and includes separately acquired MP3 files, MIDI files, lyrics, and cultural data[11] for each piece.[12] It is divided into five broad genres (Blues, Classical, Jazz, Rap, and Rock), which can be expanded into five pairs of more closely-related sub-genres: Modern Blues (BluesMod), Traditional Blues (BluesTra), Baroque (ClassBar), Romantic (ClassRom), Bop (JazzBop), Swing (JazzSwi), Hardcore Rap (RapHar), Pop Rap (RapPop), Alternative Rock (RockAlt), and Metal (RockMet). This arrangement permits experiments examining performance on both broad and more similar genres. Although SLAC is relatively small, it has the advantage of allowing audio, symbolic, and other features to be extracted independently (the MP3, MIDI, lyric, and cultural data were each acquired independently). Furthermore, its music was carefully selected and hand-labeled by genre based on expert knowledge, with the specific SLAC genre ontology in mind.

Overall, LMD-aligned and SLAC are complementary for

selected from a modified version of the LMD-aligned dataset (Raffel, 2016), and 2,671 features extracted from the 250 pieces in the SLAC dataset (McKay et al., 2010).[8]

5.1 DATA USED

LMD-aligned (Raffel, 2016) is a part of the Lakh dataset, which contains 31,034 MIDI files aligned to publicly available 30-second 7digital previews[9] associated with the Million Song Dataset (MSD) (Bertin-Mahieux et al., 2011). The number of matching tracks declines significantly when modalities are expanded, however. First, only 12,827 tracks have lyrics with a musiXmatch ID and copyright permission.[10] Then, after eliminating tracks without tagtraum genre annotations (Schreiber, 2015) and without album covers available on either of the two APIs we accessed (see Section 5.5), only 4,579 tracks remained. 26 MIDI files and 20 audio previews then had to be removed because of errors. The result was unbalanced, with Rock (2282 tracks) and Pop (1828) dominating, and genres like Punk (26) or Reggae (38) underrepresented. We therefore removed all but the five genres with at least 300 tracks: Rock, Pop, Country (415), Electronic (323), and RnB (316). Balanced experimental partitioning required further reduction of the data, leading to the final number of 1,575 tracks (3 partitions of 525 tracks equally distributed across 5 genres, see ***Table 3*** in Section 5.8). These data also have limitations with respect to the features that can be be extracted from them: the musiXmatch data have incomplete lyrics for certain tracks and, since the audio clips are only 30s, some long-framed audio features had a diminished force this research: LMD-aligned is larger with almost all data publicly available, and SLAC has high-quality labels, fully accessible features, and a balanced genre distribution.

Table 1 provides a summary of the primary feature groups we extracted, each corresponding to a different modality. We employed a large feature catalogue with a high number of bespoke features so that we could explore as broad a range of musical qualities as possible, in hope of revealing unexpected meaningful patterns. Using too few features would risk limiting results due to the effects of our pre-existing biases when selecting features, and would reduce the potential for serendipitous discoveries. Of course, this approach comes with disadvantages, as a large feature catalogue increases complexity, which can be particularly problematic with small datasets.

Some features describe similar musical concepts extracted from different modalities, like the tempo or instrument presence features, which are estimated both from the audio and symbolic data. This reflects the reality that such features may sometimes be more reliably extracted from one modality than another, something that should be taken into account when measuring the "importance" and "redundancy" of the corresponding modalities.

Not all features are relevant to all music. The features used here are designed such that an absence of a relevant musical quality can still result in a meaningful feature, so that, for example, an instrumental piece will have a *NumberOfWords* value of 0, which could usefully reflect a higher prevalence of instrumental music in the piece's genre.

Group	Sub-Groups	Sample Features	Dim.
Audio signal	Timbre, pitch + harmony, tempo + rhythm + structure, structural complexity	MFCCs and delta MFCCs (Lartillot and Toiviainen, 2007), CMRARE modulation features (Martin and Nagathil, 2009), chroma DCT-reduced log pitch (Müller and Ewert, 2011), structural complexity (Mauch and Levy, 2011) for chroma, chords, harmony, tempo/rhythm, timbre	908
Model-based	Instruments, instrumental complexity, moods, various semantic descriptors	Share of guitar, piano, wind, and strings, semantic descriptors annotated by music experts: orchestra occurence, clear or rough vocals, melodic range, dynamics, digital effects, level of activation	494
Symbolic	Pitch, melodic, chords, rhythm, tempo, instrument presence, instruments, texture, dynamics	Pitch class histogram, amount of arpeggiation, tempo, number of instruments, dynamic range and variation	789
Album covers	–	SIFT descriptors (Lowe, 2004)	100

Playlists	–	Co-occurrences of artists (Vatolkin et al., 2014)	293
Lyrics	–	Average number of syllables per word, rate of misspelling, vocabulary size, bag-of-words, Doc2Vec	87/219

Table 1: Summary of feature groups associated with each of the six modalities. The complete list of features is provided in the supplementary material, Sections A.1 to A.6.

5.2 AUDIO SIGNAL FEATURES

Audio features were extracted with AMUSE (Vatolkin et al., 2010). This group of features includes both low-level audio features and "semantic audio features" from (Vatolkin et al., 2015). The dimensionality of this set was reduced by removing similar features (e.g., calculated using different chroma and MFCC implementations). Features extracted from frames below 1s in length were stored only for time frames between onset times previously estimated with MIRtoolbox (Lartillot and Toiviainen, 2007). The list of the audio signal features is provided in the supplementary material (Section A.1), and has 908 dimensions.

5.3 FEATURES BASED ON SEMANTIC TAGS PREDICTED FROM AUDIO BY PRE-TRAINED MODELS

These features are based on high-level, interpretable annotations relating to areas like instrumentation, vocal characteristics, or moods inferred from the audio signal by an ensemble of classifiers pre-trained on expert-annotated audio data disjoint from SLAC and LMD-aligned, as described by Vatolkin et al. (2015). This approach transfers modeled expert tag predictions to genre recognition; although the models operate as black boxes, they can provide interpretable insights on genre, as they predict semantically meaningful characteristics. Choi et al. (2017) provide another musical example of transfer learning, in their case using neuron activation weights as a feature for further classification tasks. Section A.2 in the supplementary material summarizes these 494 features.

5.4 SYMBOLIC FEATURES

These features were extracted from MIDI files with jSymbolic 2.2 (McKay et al., 2018). Certain histogram features were omitted to reduce dimensionality, but summary features derived from these histograms were kept. Section A.3 in the supplementary material lists the 789 jSymbolic features we used.

5.5 ALBUM COVER DATA AND FEATURES

We mined images of album covers and extracted features from them using the methodology described by Wilkes (2019). The covers were automatically downloaded using the Discogs[13] and MusicBrainz[14] APIs (the first available covers on Discogs were saved manually when automatic download failed). We then extracted scale-invariant feature transform (SIFT) descriptors, as described by Lowe (2004), which estimate "key points" in pictures and highlight relevant statistical properties. Bag-of-keypoint features (Csurka et al., 2004) were then extracted in order to reduce dimensionality; the original SIFT descriptors were mapped to 100 clusters, which together comprised a visual vocabulary. Finally, we calculated the relative frequencies of these visual words for each album cover, which led to a 100-dimensional album cover feature vector.

5.6 PLAYLIST DATA AND CO-OCCURRENCE FEATURES

Playlist features were extracted as by Vatolkin et al. (2014). This involved creating a list of representative music tracks for six genres and eight styles based on AllMusicGuide[15] annotations from the training sets used by Vatolkin et al. (2015). For each class, ten "positive" and ten "negative" artists were stored (e.g., Beethoven and Haydn as positive Classical examples, and Ray Charles and Madonna as negative ones). As in Vatolkin et al. (2014), relevant artists for each class were selected based on the top ten co-occurring artists for 280 genre- and style-representative artists in playlists from 8tracks and Last.fm datasets (Bonnin and Jannach, 2014). We then calculated normalized playlist co-occurrences between tracks by these class-relevant artists and the genres were also represented. For example, a training set for the BluesMod sub-genre contained 8 positive tracks of this class, 8 negative tracks of the BluesTra sub-genre,

SLAC and LMD-aligned music; after removing identical artists, this produced a playlist co-occurrence feature vector with 293 dimensions, shown in Section A.5 in the supplementary material.[16]

5.7 LYRIC TEXT FEATURES

For SLAC, lyrics for all non-instrumental pieces were accessed and had features extracted from them using jLyrics, lyricFetcher, and related tools described by McKay et al. (2010). This resulted in 87 feature values per piece. As some of these tools are not publicly available and it was not possible to extract all features from LMD-aligned, we supplemented the jLyrics features for LMD-aligned with a 100-dimensional Bag-of-Words vector (Harris, 1954) based on the Term-Frequency/Inverse Document Frequency (TF-IDF) implemented in scikit-learn (Pedregosa et al., 2011) and the 100-dimensional Doc2Vec (Le and Mikolov, 2014) function implemented in gensim (Řehůřek and Sojka, 2010). Section A.6 in the supplementary material provides more details.

5.8 DATA PARTITIONING

The datasets were each divided into three equal non-overlapping partitions, or folds, in our experiments, and these were used in a 3-fold cross-validation scheme: in each of the three corresponding splits, each fold served as either training data, validation data to measure feature set fitness during FS, or testing data for final independent evaluation (see *Table 2*).

For a better balance, the training set for a given genre's binary experiment was compiled from subsets of the corresponding fold using the maximum possible same number of "positive" and "negative" tracks with respect to the genre to predict. The negative tracks were sampled equally from each remaining genre. The validation and test sets were not balanced, however. So, for LMD-aligned, balancing required constraining the fold sizes based on the genre with the smallest number of tracks (RnB, with 316 tracks), with the result that for a given genre 105 tracks were available as positives in each of the three folds, and 104 as training negatives (26 for each remaining genre); this left 420 negatives each for validation and testing (105 for each remaining genre). In the exceptional case of the SLAC sub-genre experiments, instances belonging to the paired similar sub-genre were emphasized in each negative group, but all other sub- and 8 negative tracks from all 8 remaining sub-genres (one track per sub-genre). *Table 3* shows the distribution of tracks for a given genre and a given split.

Split	Training	Validation	Test
1	Fold 1	Fold 2	Fold 3
2	Fold 2	Fold 3	Fold 1
3	Fold 3	Fold 1	Fold 2

Table 2: Fold assignments in cross-validation splits.

6. RESULTS AND DISCUSSION

6.1 CLASSIFICATION ERROR AND MULTI-GROUP IMPORTANCE

Table 4 summarizes the multi-objective $h(m_e \downarrow, g_k \uparrow)$ FS results for each feature type. Each experiment was repeated ten times, with different randomly initialized feature subsets every time, according to a "multi-start" procedure (Martí et al., 2018); this permitted a better exploration of the search space and reduced the potential influence of local optima. Evaluation was performed in all cases on the reserved test set, which was not involved in the optimization process. We performed single-tailed Wilcoxon rank tests in order to get a sense of statistical certainty and significance comparing modalities for each (sub-)genre. The detailed p-values are shown in the supplementary material (Sections B.1 to B.6)

Playlist co-occurrence features tended to have the best individual (pure) m_e^*: they had the lowest mean m_e^* for 13 of the 20 classes, and the lowest m_e^* for 40 of the 60 individual genre/fold combinations (see Section C.1 in the supplementary material), as well as the highest mean I_h^* for 9 classes and 41 combinations. This might be expected, as listeners often create playlists based on genre preferences. Album cover and lyrics features, in contrast, were the worst performers, with one or the other corresponding to the highest mean error for all 20 classes (17 and 3, respectively, with the lyrics alone performing worst for LMD-aligned Rock and the SLAC ClassBar and JazzSwi sub-genres).

Despite the clear effectiveness of pure playlist features, still better performance was almost always obtained when they were combined with other feature types. This is demonstrated by the fact that for no class on average and in only 4 of the 60 genre/fold

Tracks	Training	Validation	Test
LMD-aligned genres			
Positives	105	105	105
Negatives	104	420	420
SLAC genres			
Positives	16	16	16
Negatives	16	64	64
SLAC sub-genres			
Positives	8	8	8
Negatives	16	72	72

Table 3: Numbers of positive and negative tracks in the training, validation, and test sets for a split.

		LMD-aligned genres					SLAC genres				
Group		**Country**	**Electronic**	**Pop**	**RnB**	**Rock**	**Blues**	**Classical**	**Rock**	**Jazz**	**Rap**
Audio Signal	m_e^*	0.241±0.01	0.212±0.02	0.439±0.02	0.319±0.03	0.423±0.02	0.181±0.06	0.051±0.02	0.030±0.02	0.091±0.03	0.022±0.02
	I_h^*	0.968±0.01	0.966±0.02	0.875±0.11	0.939±0.04	0.877±0.03	**0.989±0.00**	0.998±0.00	0.999±0.00	0.997±0.00	1.000±0.00
Model-Based	m_e^*	0.294±0.00	0.259±0.01	0.433±0.02	0.333±0.03	0.402±0.02	0.185±0.01	0.074±0.04	0.052±0.02	0.129±0.04	0.024±0.01
	I_h^*	0.928±0.02	0.944±0.02	0.894±0.08	0.949±0.01	0.899±0.01	0.983±0.01	0.991±0.00	0.996±0.00	0.989±0.01	1.000±0.00
Symbolic	m_e^*	0.254±0.02	0.256±0.01	0.456±0.01	0.329±0.03	0.457±0.02	0.130±0.03	0.016±0.01	0.048±0.03	0.047±0.01	0.126±0.04
	I_h^*	0.927±0.03	0.933±0.03	0.846±0.03	0.956±0.01	0.821±0.06	0.987±0.01	**0.999±0.00**	0.996±0.00	**0.999±0.00**	0.990±0.01
Album Cover	m_e^*	0.403±0.01	0.429±0.00	0.479±0.01	0.453±0.02	0.471±0.01	0.375±0.05	0.366±0.03	0.400±0.04	0.448±0.06	0.435±0.05
	I_h^*	0.830±0.01	0.844±0.01	0.753±0.08	0.839±0.05	0.765±0.03	0.856±0.02	0.901±0.01	0.865±0.04	0.890±0.01	0.913±0.01
Playlists	m_e^*	0.060±0.01	0.109±0.00	0.223±0.01	0.073±0.01	0.225±0.01	0.102±0.08	0.109±0.15	0.026±0.04	0.040±0.01	0.015±0.03
	I_h^*	**1.000±0.00**	**0.999±0.00**	**0.998±0.00**	**0.998±0.00**	**0.996±0.01**	0.874±0.22	0.889±0.15	**1.000±0.00**	0.950±0.06	**1.000±0.00**
Lyrics	m_e^*	0.304±0.00	0.387±0.02	0.446±0.01	0.370±0.02	0.473±0.01	0.180±0.05	0.150±0.03	0.140±0.06	0.273±0.08	0.051±0.01
	I_h^*	0.912±0.02	0.899±0.01	0.857±0.03	0.922±0.02	0.811±0.03	0.908±0.04	0.973±0.00	0.959±0.02	0.850±0.10	0.992±0.01
Combined	m_e^{**}	0.052±0.01	0.099±0.00	0.212±0.02	0.067±0.01	0.211±0.01	0.006±0.00	0.002±0.00	0.000±0.00	0.006±0.01	0.000±0.00

		SLAC sub-genres									
Group		**BluesMod**	**BluesTra**	**ClassBar**	**ClassRom**	**JazzBop**	**JazzSwi**	**RapHar**	**RapPop**	**RockAlt**	**RockMet**
Audio Signal	m_e^*	0.287±0.07	0.296±0.17	0.417±0.07	0.073±0.07	0.167±0.02	0.387±0.13	0.218±0.07	0.336±0.01	0.213±0.16	0.094±0.05
	I_h^*	0.984±0.01	0.976±0.04	0.946±0.03	0.989±0.01	**0.996±0.00**	0.975±0.01	0.987±0.01	0.957±0.04	0.976±0.03	**0.999±0.00**
Model-Based	m_e^*	0.267±0.07	0.317±0.11	0.416±0.01	0.149±0.17	0.178±0.08	0.323±0.09	0.202±0.08	0.314±0.01	0.348±0.09	0.159±0.08
	I_h^*	0.974±0.01	0.937±0.06	0.854±0.08	0.946±0.09	0.975±0.01	0.947±0.04	0.988±0.01	0.959±0.03	0.966±0.01	0.993±0.01
Symbolic	m_e^*	0.271±0.09	0.242±0.07	0.248±0.11	0.116±0.05	0.154±0.09	0.118±0.04	0.343±0.09	0.321±0.03	0.266±0.08	0.213±0.11
	I_h^*	0.946±0.05	**0.999±0.00**	**0.963±0.04**	**0.999±0.00**	0.850±0.14	**0.991±0.01**	0.988±0.01	**0.998±0.00**	0.967±0.02	0.911±0.15
Album Cover	m_e^*	0.453±0.03	0.492±0.11	0.366±0.17	0.458±0.08	0.437±0.08	0.428±0.07	0.522±0.04	0.435±0.01	0.382±0.17	0.483±0.07
	I_h^*	0.829±0.13	0.781±0.08	0.838±0.11	0.606±0.20	0.764±0.09	0.806±0.03	0.666±0.11	0.797±0.07	0.657±0.26	0.769±0.04
Playlists	m_e^*	0.135±0.12	0.389±0.17	0.257±0.15	0.246±0.12	0.057±0.02	0.159±0.16	0.258±0.15	0.241±0.02	0.070±0.06	0.098±0.08
	I_h^*	**1.000±0.00**	0.943±0.08	0.747±0.42	0.969±0.03	0.993±0.01	0.988±0.02	0.987±0.01	0.997±0.00	**1.000±0.00**	0.999±0.00
Lyrics	m_e^*	0.315±0.06	0.403±0.09	0.477±0.05	0.320±0.06	0.374±0.16	0.479±0.04	0.158±0.01	0.290±0.02	0.270±0.10	0.372±0.05
	I_h^*	0.937±0.01	0.806±0.11	0.919±0.09	0.868±0.07	0.888±0.04	0.808±0.10	**0.992±0.01**	0.883±0.06	0.881±0.07	0.861±0.08
Combined	m_e^{**}	0.073±0.07	0.158±0.04	0.113±0.05	0.036±0.05	0.023±0.03	0.046±0.05	0.086±0.03	0.169±0.01	0.028±0.03	0.022±0.03

Table 4: Comparison of the six feature types based on h ($m_e \downarrow$, $g_k \uparrow$) FS optimization. Mean and standard deviations are estimated for the three folds in the splits in which they respectively played a test role (see Section 5.8), and across all ten repetitions of each experiment. All rows but the two starting with "Combined" indicate mean best test classification errors m_e^* for pure feature groups only (lower values are better) and mean normalized multi-group feature importance I_h^* (higher values are better). m_e^* values are averaged across the ten repetitions, and each I_h^* value specifies the highest-importance non-dominated solution among all ten experimental trials. The mean best m_e^* and I_h^* for each class is in bold, and cell background color indicates sorted mean I_h^* values: deep red indicates highest importance and deep blue corresponds to lowest importance for a given column and its folds. Finally, the values in the rows starting with "Combined" indicate the smallest mean test error m_e^{**} obtained across all non-dominated solutions for each class, including (in this row only) mixed feature sets. The following procedure was used to estimate m_e^{**}: first, the smallest error from all non-dominated solutions for each individual experimental run is noted, this is then averaged across the ten experimental trials, and the minimum is taken across all six feature groups.

combinations (SLAC Rock fold 1; Rap folds 2 and 3; and RockAlt fold 1, cf. Section C.1 in the supplementary material) did playlist features alone match the best m_e^{**} achieved when mixed feature types were allowed (as shown in the rows starting with "Combined" in ***Table 4*** or the last three columns in Table 13, Section C.1 in the supplementary material). In fact, all mean m_e^{**}-values achieved with combined feature sets in ***Table 4*** are lower than the mean smallest errors of all the pure feature groups for each corresponding genre. This supports the potential utility of a multi-modal approach for both attempting to improve classification performance and for research on learning more about complex class traits, even when a single highly discriminative source data type is available.

Turning our attention to I_h^*, it should be noted that a value of 1.000 on ***Table 4*** does not typically correspond carry some useful information. By far the smallest mean I_h^* in ***Table 4*** is 0.606 (ClassRom, album cover features).

Symbolic features have the highest mean I_h^* for 7 of the 20 classes, audio signal features for 3, and lyrics features for 1. This relative success of symbolic features does make some intuitive sense, as many harmonic, melodic, rhythmic, and other important musical properties can be derived from the score. However, other characteristics like applied digital effects cannot typically be extracted from symbolic data.

Certain feature types are better suited to certain genres than others. For instance, audio signal and model-based features seem to be more important for Rap than symbolic features. However, symbolic features have higher I_h^* values for the RapPop sub-genre, which suggests they are better at separating RapPop tracks from RapHar (as well as other genres). Lyrics are the most

to an h of absolute 0, but rather to very small h values rounded to zero. So, for entries with high I_h^* values, improvements do exist when features are added from other groups, but they are very small. It is also notable that none of the feature groups resulted in an I_h^* anywhere near 0, which suggests that all the feature groups do important for RapHar and the least important for Jazz. Playlist features have a lower importance for the SLAC Blues and Classical genres, but have a higher importance for the sub-genres BluesMod and ClassRom, which may be explained by a poor balance of specific sub-genres in the playlist data taken into account.

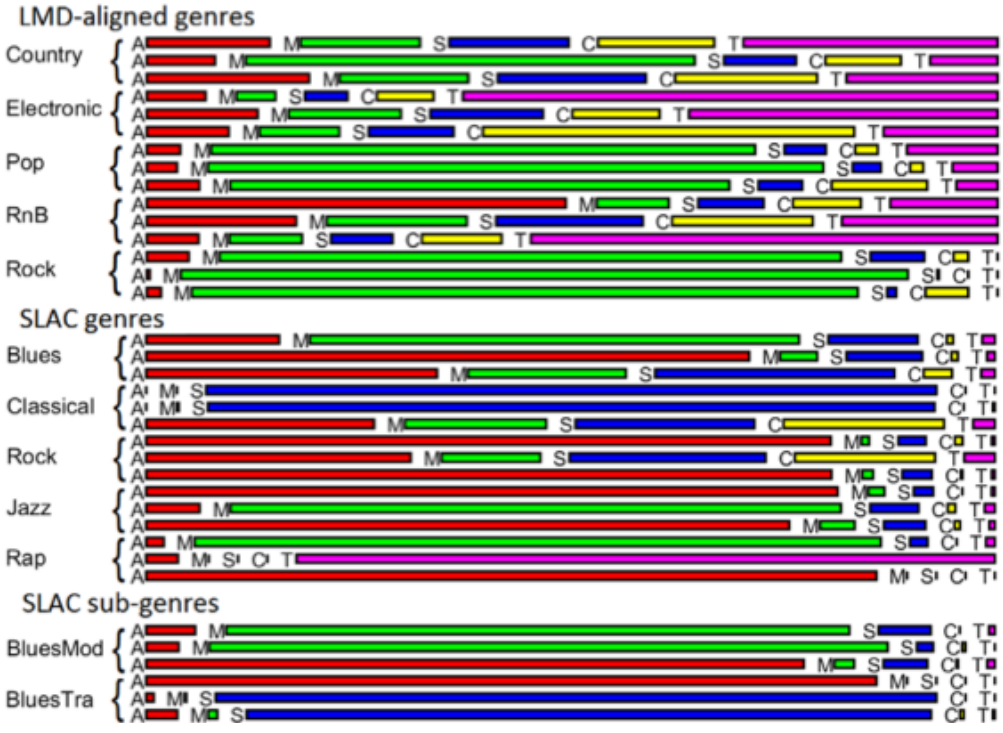
As one might expect, differences between folds are lower in LMD-aligned than SLAC, as the former contains more music. Also, differences between the "strongest" and "weakest" modalities are clearer in LMD-aligned: playlists achieve the smallest m_e^* and largest I_h^* in all 15 genre/fold combinations, and album covers the largest m_e^* and smallest I_h^*.

6.2 EVALUATION WITHOUT PLAYLISTS

Although playlist features are highly predictive of genre, they have two important disadvantages: they are less effective in revealing semantic properties of genre, and they suffer from the cold-start problem. We therefore repeated the experiments described above omitting playlist features, to see how the other feature groups would perform without them.

Figure 4 shows the extent to which each feature group is represented in the mixed feature subset with the smallest test error for each genre. The most-represented feature type depends on the particular genre; for example, symbolic features are good at identifying Classical, BluesTra, and JazzSwi music, but are less important than audio or model-based features for other genres (or lyrics, for some of the LMD-aligned genres). Interestingly, album covers also play a more prominent role in LMD-aligned than in SLAC. Of course, a single feature from a given group could potentially have high discriminative power, even if no other features from its group do.

Large differences often exist across the three folds, particularly for SLAC genres. For instance, the best feature sets for two of the three RockMet folds are mainly audio, but for the third fold model-predicted semantic features dominate. Two possible explanations for this come to mind. First, SLAC is a relatively small dataset, and it also contains a few tracks by the same artists, especially in



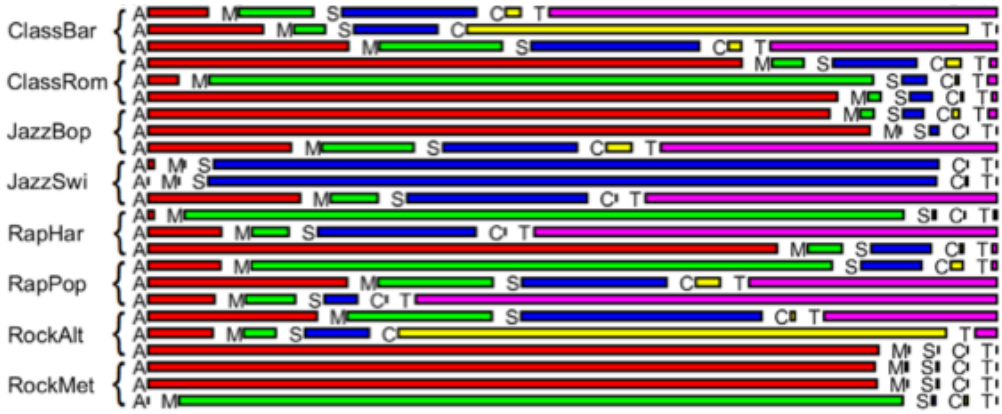


Figure 4: Share of each non-playlist feature group in the feature subsets with the smallest test errors for each genre. A: audio; M: model-predicted tag; S: symbolic; C: album cover; T: lyrics. Results are based on $h(m_e \downarrow, g_k \uparrow)$ FS optimization, and are shown for each of the three folds separately, for the splits in which they played a test role.

the case of RockMet. This necessitated a difficult decision during fold partitioning: including tracks from the same artist in both the training and validation partitions leads to higher fold interdependency, but restricting all the tracks by the same artist to a single fold (as done in this study) could lead to the learning of artists rather than genres; this may be responsible for some of the inter-fold variance. Second, the feature space is huge and the exploration method has its limits, so one cannot guarantee that optimal features are always found. Having more distinct features from the same group and, in particular, different groups would help to overcome local optima. As many application scenarios may involve small or overlapping datasets (e.g., if a listener defines some personal music category with just a few representative tracks), this strengthens our recommendation to combine features from different sources.

6.3 MULTI-GROUP REDUNDANCY

Table 5 reports multi-group feature redundancy (R_h^*) results after $h(m_e \downarrow, g_k \downarrow)$ optimization. The research question in this optimization is to find whether it is possible to achieve a good classification performance using as few features from the given group as possible. Note that very low values of mean R_h^* are rare: the smallest one is 0.8036.

Symbolic features appear to be the least redundant, as they contribute to the smallest mean R_h^* values for 8 of 20 classes and for 24 of the 60 genre/fold combinations (Section C.2 in the supplementary material lists individual values). This means that classification performance BluesTra and ClassBar. Jazz-Swi and Classical are the only classes for which all feature groups but symbolic have a fully redundant mean $R_h = 1.0$. Lyrics features are least redundant for the Country and Rap genres.

Overall, we see once again that it can be useful to have a variety of feature types available when dealing with varied classes. Different modalities can differ widely in both their importance and redundancy with respect to different classes.

6.4 SPECIFIC FEATURE GROUPS

The comparative strategies discussed above can be applied not only to comparing features drawn from different modalities like audio or lyrics, but also to specialized feature groups from the same modality. For instance, one can explore which kinds of symbolic features (harmonic, rhythmic, etc.) are most helpful in recognizing a particular genre. Other potential approaches might focus on examining feature groups in terms of extraction cost, availability in open-source vs. closed-source software frameworks, methods for aggregating over time frames, etc. To delve into this experimentally, we selected 15 feature sub-groups. This is illustrated in *Table 6*, which compares mean I_h^* values for these sub-groups across all folds after $h(m_e \downarrow, g_k \uparrow)$ FS optimization.

The particular grouping of audio signal features is based on Vatolkin et al. (2015). However, we have created a separate group for structural complexity features based on Mauch and Levy (2011), as their calculation principle is distinct from the others: the changes of original feature

values). This means that classification performance suffers more when symbolic features are omitted from the feature set, supporting the notion that it may be hard to identify all important genre properties from the audio signal only, or from other modalities other than the score. Lowest mean R_h^* values are achieved for 6 classes using audio signal features, for 4 classes using model-based features, and for 2 classes using lyrics. Album cover features are completely redundant for 14 classes, but are second-best for the SLAC sub-genres

is distinct from the others: the changes of original feature values are measured over long time frames. Model-based features are also grouped after Vatolkin et al. (2015); the "various" sub-group is a subset of binary descriptors of personal music categories defined by musicologists in a study by Rötter et al. (2013), like "presence of drums" or "high activation level" (see Section A.2 in the supplementary material). Groups of symbolic features describe different music properties based on the categorization of features in jSymbolic (McKay et

	LMD-aligned genres					SLAC genres				
Group	**Country**	**Electronic**	**Pop**	**RnB**	**Rock**	**Blues**	**Classical**	**Rock**	**Jazz**	**Rap**
Audio Signal	0.9982±0.00	**0.9715±0.01**	0.9995±0.00	0.9954±0.01	0.9753±0.04	0.9492±0.09	1.0000±0.00	1.0000±0.00	1.0000±0.00	1.0000±0.00
Model-Based	0.9998±0.00	0.9998±0.00	**0.9769±0.02**	0.9989±0.00	**0.9569±0.01**	0.9453±0.09	1.0000±0.00	1.0000±0.00	1.0000±0.00	1.0000±0.00
Symbolic	0.9949±0.00	0.9987±0.00	0.9997±0.00	**0.9820±0.02**	0.9998±0.00	**0.9229±0.13**	**0.9875±0.01**	**0.9919±0.01**	**0.9908±0.00**	1.0000±0.00
Album Cover	0.9994±0.00	0.9996±0.00	1.0000±0.00	1.0000±0.00	0.9986±0.00	1.0000±0.00	1.0000±0.00	1.0000±0.00	1.0000±0.00	1.0000±0.00
Lyrics	**0.9934±0.01**	0.9866±0.00	0.9881±0.02	0.9974±0.00	0.9999±0.00	0.9992±0.00	1.0000±0.00	1.0000±0.00	1.0000±0.00	**0.9988±0.00**
	SLAC sub-genres									
Group	**BluesMod**	**BluesTra**	**ClassBar**	**ClassRom**	**JazzBop**	**JazzSwi**	**RapHar**	**RapPop**	**RockAlt**	**RockMet**
Audio Signal	**0.8889±0.16**	1.0000±0.00	1.0000±0.00	**0.8814±0.12**	**0.9506±0.09**	1.0000±0.00	**0.8716±0.13**	1.0000±0.00	**0.8125±0.22**	0.9845±0.02
Model-Based	0.9675±0.05	0.9996±0.00	1.0000±0.00	0.9435±0.10	0.9593±0.07	1.0000±0.00	0.9044±0.09	**0.9414±0.05**	0.9057±0.08	**0.9372±0.05**
Symbolic	0.9168±0.14	**0.8320±0.28**	**0.8824±0.13**	0.9799±0.03	0.9873±0.02	**0.8036±0.31**	0.9564±0.07	0.9871±0.01	0.8957±0.16	0.9902±0.02
Album Cover	0.9999±0.00	0.9918±0.01	0.9848±0.03	1.0000±0.00	1.0000±0.00	1.0000±0.00	1.0000±0.00	1.0000±0.00	1.0000±0.00	1.0000±0.00
Lyrics	0.9881±0.01	1.0000±0.00	0.9880±0.02	0.9802±0.03	0.9818±0.01	1.0000±0.00	0.9627±0.02	0.9745±0.02	0.9755±0.03	0.9900±0.02

Table 5: Normalized multi-group feature redundancy (R_h*) comparison of the five feature types left after excluding playlist features (lower values are better). The mean and standard deviation are shown across three folds. The best value for each class is in bold. Deep red indicates the best mean R_h* and deep blue the worst (equal values are possible).

	LMD-aligned genres					SLAC genres				
Group	**Country**	**Electronic**	**Pop**	**RnB**	**Rock**	**Blues**	**Classical**	**Rock**	**Jazz**	**Rap**
Audio Signal: Low-Level	**0.956±0.01**	**0.969±0.01**	0.875±0.09	**0.965±0.01**	0.896±0.04	**0.988±0.00**	0.997±0.00	**0.999±0.00**	**0.997±0.00**	**0.999±0.00**
Audio Signal: Semantic	0.923±0.00	0.957±0.02	0.881±0.07	0.940±0.02	0.875±0.01	0.970±0.01	**0.998±0.00**	0.984±0.00	0.988±0.00	0.992±0.00
Audio Signal: Str. Compl.	0.813±0.04	0.848±0.03	0.805±0.04	0.815±0.01	0.732±0.07	0.924±0.03	0.978±0.02	0.975±0.01	0.973±0.01	0.974±0.02
Model-Based: Instruments	0.859±0.03	0.888±0.04	0.812±0.06	0.880±0.02	0.831±0.02	0.934±0.01	0.975±0.02	0.977±0.00	0.975±0.00	0.989±0.01
Model-Based: Moods	0.857±0.02	0.877±0.01	0.868±0.05	0.883±0.02	0.848±0.04	0.950±0.01	0.990±0.01	0.991±0.00	0.980±0.01	0.997±0.00
Model-Based: Various	0.899±0.01	0.937±0.02	0.856±0.07	0.928±0.00	0.874±0.02	0.959±0.02	0.995±0.00	0.992±0.01	0.984±0.01	0.999±0.00
Symbolic: Pitch	0.796±0.05	0.825±0.04	0.752±0.05	0.808±0.04	0.726±0.06	0.923±0.03	0.934±0.02	0.933±0.00	0.967±0.01	0.967±0.01
Symbolic: Melodic	0.675±0.07	0.774±0.02	0.694±0.07	0.689±0.10	0.602±0.03	0.742±0.07	0.782±0.04	0.887±0.03	0.924±0.02	0.920±0.02
Symbolic: Chords	0.798±0.02	0.815±0.01	0.735±0.04	0.807±0.05	0.713±0.07	0.858±0.06	0.962±0.01	0.911±0.07	0.965±0.03	0.930±0.05
Symbolic: Rhythm	0.820±0.04	0.871±0.02	0.806±0.04	0.844±0.03	0.748±0.03	0.902±0.03	0.949±0.01	0.917±0.03	0.966±0.00	0.928±0.02
Symbolic: Tempo	0.656±0.03	0.754±0.07	0.701±0.04	0.716±0.02	0.655±0.03	0.759±0.05	0.872±0.02	0.833±0.05	0.900±0.02	0.861±0.02
Symbolic: Instr. Presence	0.945±0.01	0.951±0.02	**0.927±0.02**	0.958±0.02	**0.917±0.02**	0.975±0.00	0.996±0.00	0.995±0.00	0.996±0.00	0.996±0.00
Symbolic: Instruments	0.614±0.02	0.697±0.03	0.626±0.06	0.621±0.09	0.598±0.02	0.769±0.10	0.945±0.05	0.915±0.04	0.908±0.02	0.849±0.03
Symbolic: Texture	0.700±0.03	0.746±0.02	0.713±0.04	0.697±0.07	0.657±0.05	0.723±0.13	0.885±0.05	0.905±0.02	0.888±0.06	0.912±0.04
Symbolic: Dynamics	0.216±0.04	0.304±0.03	0.416±0.01	0.318±0.05	0.321±0.07	0.473±0.18	0.663±0.10	0.477±0.03	0.614±0.15	0.538±0.05
	SLAC sub-genres									
Group	**BluesMod**	**BluesTra**	**ClassBar**	**ClassRom**	**JazzBop**	**JazzSwi**	**RapHar**	**RapPop**	**RockAlt**	**RockMet**
Audio Signal: Low-Level	**0.982±0.01**	**0.984±0.01**	0.934±0.01	0.982±0.01	0.984±0.01	**0.977±0.00**	**0.994±0.01**	0.940±0.03	0.968±0.04	**0.999±0.00**
Audio Signal: Semantic	0.952±0.03	0.945±0.03	0.947±0.03	0.975±0.02	**0.987±0.01**	0.971±0.01	0.967±0.04	0.930±0.02	0.948±0.04	0.976±0.00
Audio Signal: Str. Compl.	0.875±0.02	0.819±0.12	0.832±0.10	0.981±0.01	0.942±0.02	0.918±0.01	0.888±0.04	0.896±0.03	0.914±0.00	0.976±0.01
Model-Based: Instruments	0.930±0.05	0.904±0.13	0.875±0.10	0.888±0.17	0.952±0.03	0.919±0.06	0.885±0.07	0.901±0.04	0.899±0.02	0.971±0.02
Model-Based: Moods	0.939±0.04	0.859±0.06	0.939±0.05	0.917±0.09	0.978±0.02	0.928±0.03	0.974±0.00	**0.976±0.01**	0.945±0.03	0.998±0.00
Model-Based: Various	0.978±0.01	0.908±0.09	0.894±0.03	0.983±0.02	0.973±0.01	0.934±0.04	0.979±0.02	0.916±0.05	0.958±0.02	0.993±0.01
Symbolic: Pitch	0.922±0.05	0.836±0.10	0.914±0.07	0.831±0.12	0.957±0.01	0.957±0.01	0.845±0.00	0.851±0.08	0.787±0.06	0.914±0.03
Symbolic: Melodic	0.711±0.07	0.664±0.10	0.747±0.26	0.710±0.14	0.767±0.15	0.863±0.07	0.790±0.07	0.637±0.03	0.768±0.10	0.840±0.06
Symbolic: Chords	0.775±0.04	0.909±0.02	0.921±0.06	0.940±0.04	0.855±0.04	0.911±0.06	0.898±0.06	0.926±0.04	0.815±0.07	0.873±0.09
Symbolic: Rhythm	0.701±0.18	0.949±0.03	0.804±0.11	0.893±0.07	0.824±0.08	0.946±0.05	0.880±0.03	0.883±0.06	0.891±0.02	0.847±0.11
Symbolic: Tempo	0.732±0.19	0.778±0.18	0.815±0.14	0.781±0.12	0.863±0.03	0.932±0.04	0.688±0.18	0.787±0.19	0.824±0.08	0.840±0.04
Symbolic: Instr. Presence	0.965±0.04	0.971±0.01	**0.997±0.00**	**1.000±0.00**	0.966±0.03	0.947±0.04	0.965±0.01	0.953±0.01	**0.971±0.01**	0.988±0.00
Symbolic: Instruments	0.826±0.05	0.844±0.09	0.816±0.03	0.895±0.14	0.818±0.10	0.876±0.05	0.693±0.11	0.644±0.16	0.728±0.09	0.950±0.03

Symbolic: Texture	0.837±0.14	0.673±0.14	0.860±0.14	0.670±0.09	0.865±0.03	0.861±0.08	0.883±0.11	0.658±0.17	0.681±0.05	0.840±0.09
Symbolic: Dynamics	0.329±0.31	0.626±0.05	0.513±0.38	0.603±0.18	0.411±0.19	0.588±0.21	0.271±0.10	0.457±0.14	0.423±0.05	0.565±0.11

Table 6: Comparison of 15 feature sub-groups from the modalities Audio Signal, Model-Based, and Symbolic with respect to normalized multi-group importance I_h^*, after $h(m_e\downarrow, g_k\uparrow)$ FS optimization. Higher values are better. Mean values and standard deviations across the three folds are reported. The highest mean I_h^* value for each genre is in bold; cells with higher values are marked in red, and cells with lower values in blue.

al., 2018), with two exceptions: features that simply measure instrument presence are separated from other instrumental statistics, and "rhythm-related features that are influenced by tempo" are also separated out into their own group. The groups selected here are only one possible way to distinguish features by their properties; finer or different groupings may be appropriate for future experiments (e.g., properties of string instruments).

Low-level audio features have the highest mean I_h^* values in 12 of the 20 classes, symbolic instrument presence features in 5, audio semantic in 2, and model-based in 1. These groups seem to be particularly important, as they also contribute to more than half of the second-highest mean I_h^* values: instrument presence (7), semantic (4), and low-level (4), and moods (1). The least important sub-group is symbolic dynamics, which may be explained by its small size (only 4 dimensions) and inconsistent MIDI encoding practice.[17] Interestingly, the symbolic group "instruments", which measures higher-level features like the number of pitched instruments or electric instrument prevalence, performs worse than simpler statistics of instrument presence for all classes, and is even the second worst for six genres (but has a better performance on the SLAC sub-genres).

There is some variance across folds (see Section C.3 in the supplementary material for individual fold I_h^* values), especially in the smaller SLAC classes, but some interesting outcomes can still be identified. The Classical sub-genres Baroque and Romantic, as well as Pop, can be identified very well with only symbolic features noting the instruments present. Model-based mood features are very important for RapPop and RockMet. RnB, SLAC Rock and RapHar can be best recognized with low-level audio signal descriptors, which also have the highest I_h^* values for two of the three Blues, Jazz, Rap, RockAlt, and RockMet folds. Dynamics features perform the worst for most classes, but not for two folds of the BluesTra, ClassBar, and ClassRom sub-genres.

These results open up interesting avenues for future research. Understanding which types of features grouped by semantic, statistical, or extraction properties work best for different genres may not only provide musically interesting insights, but may also give useful hints as to what types of new features might be most promising to investigate and develop further.

7 CONCLUSIONS

This article introduces two novel statistics based on multi-objective evolutionary feature selection and non-dominated hypervolumes. These allow one to measure how "important" or "redundant" various feature groups are with respect to the identification of given classes. These statistics respectively permit the investigation of the extent to which it is possible to achieve as high a classification performance as possible while 1) maintaining as pure a feature group as possible, or 2) excluding to the greatest extent possible the feature group under investigation. This not only allows one to gain insights into which types of features or data might be most interesting to investigate from a strictly performance-oriented perspective, but can also result in ontological and other high-level insights with domain-specific value.

Our experimental results[18] suggest genre classification benefits from combining diverse features drawn from multi-

An essential area of future work is the construction of expanded and unified publicly accessible multi-modal MIR datasets, as accessing matched data is currently difficult, as seen in Section 5.1: individual independent linked data resources can become inaccessible over time, or identifiers used to match items between them can change or disappear, for example. We argue that the substantial potential of expanded multi-modal research, both pure and applied, makes the construction of high-quality multi-modal MIR datasets well worth the effort.

modal data. For example, we found that for all predicted music genres or styles the best-performing feature subset was comprised of multiple feature types, even when our methodology explicitly attempted to favor pure feature groups. The most effective feature types varied with the particular genre under consideration, and the performance of different feature sub-groups for a given type of data could also vary widely in a way dependent on the genre. However, some feature groups did perform better than others overall: playlist features tended to be the most discriminative, and album cover features the least.

This highlights the potential of more MIR research involving comparative multi-modal feature analysis. The work presented here both introduces evolutionary techniques that can be usefully applied to this purpose and integrates a broad variety of features drawn from six different types of musical data, more than any previous research. The proposed approaches also facilitate the comparison of different types of features drawn from the same kind of data. It is notable that the techniques proposed here are quite general; there is nothing about them that is specific to musical genre, or even to music. They can be applied to any classification domain involving features.

It is hoped that this work will stimulate further general research on multi-modal classification. Human listeners consume and create music in ways that are cognizant of the audio signal, symbolic musical abstractions, lyrics, cultural context, etc., and MIR can benefit from similarly considering a fuller scope of relevant information and its interrelations. The MIR community is uniquely positioned to address this kind of work, with its broad range of disciplines and techniques touching on many kinds of musical information, and there are many promising areas for future research using the multi-objective evolutionary approaches we propose here. The feature selection experiments can be repeated or re-evaluated with respect to other evaluation criteria, and compared to other feature selection algorithms. Types of music classification other than genre can be investigated, such as cover song detection, or mood or artist identification. Additional types of data can also be added, such as music videos, and a broader range of evolutionary approaches can be experimented with.

NOTES

1 In this paper, the terms "modality" and "source type" are used to refer to a specific kind of musical data (e.g., lyric texts or album cover images). A "source" indicates information belonging to a given modality for a particular instance (e.g., a text file specifying lyrics for a piece). A "feature" denotes a quantitative measure that can be calculated from sources of a given modality (e.g., the number of different instruments specified in scores). "Multi-modal" refers to work involving more than one source type.

2 Earlier work on music genre recognition that has received less attention is highlighted by Sturm (2012a).

3 Function maximization can alternatively be employed (e.g., by multiplying by -1). Different strategies can be applied for selecting and evaluating concrete "trade-off" solutions while minimizing $m_1,\ldots,m_o$; the one employed in our study (based on hypervolumes and non-dominated sorting) is introduced in Section 4.1. A list of alternatives is provided, e.g., by Audet et al. (2021).

4 We also follow another optimization strategy with the minimization of both m_e and g_k, as discussed in Section 4.2.

5 Although genre recognition, this article's sample domain, can be addressed as a multi-class problem (or even multi-label, which better reflects the reality of genre), we restrict this study to binary classification to obtain individual strengths and weaknesses of modalities and feature sub-groups for distinct classes.

6 Once again, maximization follows a similar procedure.

7 Very small errors are achieved for some classes using playlist features, as discussed in Section 6.1; the evaluation without playlist features is explicitly addressed in Section 6.2.

8 All feature values extracted from both of these datasets, along with relevant metadata, are available at: *https://zenodo.org/record/5651429*.

9 *https://colinraffel.com/projects/lmd*, accessed on: 13.11.2021.

10 *http://millionsongdataset.com/musixmatch*, accessed on: 13.11.2021.

11 The SLAC cultural data consist of statistics derived from search engine hit counts and Last.fm tag counts for predetermined search strings based on genre, archived from the time of SLAC's original publication. These data were excluded from the present study, as it was not possible to update it.

12 *https://zenodo.org/record/4571050#.YD1SlmhKj-g*, accessed on: 13.11.2021.

13 *https://www.discogs.com*, accessed on: 13.11.2021.

14 *https://musicbrainz.org*, accessed on: 13.11.2021.

15 *https://www.allmusic.com*, accessed on: 13.11.2021.

16 These features represent co-occurrences of tracks with music categories from the source publication (Vatolkin et al., 2014), not with the SLAC/LMD-aligned genre categories used in this research; the integration of these playlist statistics can also be understood as transfer learning, as with the Section 5.3 features.

17 Some MIDI files are encoded using live encoding capture, typically with a keyboard, while others are encoded with score editing or sequencing software, which might encode dynamics differently.

18 It is important to acknowledge that genre recognition suffers from fundamental labeling and evaluation concerns, as discussed above, and that large feature sets are associated with complexity issues. This provides important context when considering this and other research in genre classification and large-scale feature selection.

ADDITIONAL FILE

source, multi-level, multi-label, and large-scale. In *Proc. of*

The additional file for this article can be found as follows:

- **Supplementary Material.** A: Feature Lists; B: Results of Statistical Tests; C: Experiment Results for Three Folds. DOI: *https://doi.org/10.5334/tismir.67.s1*

ACKNOWLEDGEMENTS

Portions of this work were generously supported by (1) the German Research Foundation (DFG), under grant 336599081, and (2) the Fonds de recherche du Québec – Société et culture, under grants 2021-CHZ-282456 and 2022-CHZ-309882. The experiments were carried out on the Linux HPC cluster at TU Dortmund (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as grant 271512359.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Igor Vatolkin *orcid.org/0000-0002-9454-9402*
TU Dortmund University, Department of Computer Science, Germany

Cory McKay *orcid.org/0000-0003-3214-8862*
Marianopolis College, Department of Liberal and Creative Arts, Canada

REFERENCES

Amaldi, E., and **Kann, V.** (1998). On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 209(1–2):237–260. DOI: *https://doi.org/10.1016/S0304-3975(97)00115-1*

Audet, C., Bigeon, J., Cartier, D., Digabel, S. L., and **Salomon, L.** (2021). Performance indicators in multiobjective optimization. *European Journal on Operational Research*, 292(2):397–422. DOI: *https://doi.org/10.1016/j.ejor.2020.11.016*

Bertin-Mahieux, T., Ellis, D. P. W., Whitman, B., and **Lamere, P.** (2011). The Million Song Dataset. In *Proc. of the 12th International Society for Music Information Retrieval Conference, ISMIR*, pages 591–596.

Bischoff, K., Firan, C. S., Paiu, R., Nejdl, W., Laurier, C., and **Sordo, M.** (2009). Music mood and theme classification – a hybrid approach. In *Proc. of the 10th International Society for Music Information Retrieval Conference, ISMIR*, pages 657–662.

Bogdanov, D., Porter, A., Schreiber, H., Urbano, J., and **Oramas, S.** (2019). The AcousticBrainz Genre Dataset: Multi-*the 20th International Society for Music Information Retrieval Conference, ISMIR*, pages 360–367.

Bonnin, G., and **Jannach, D.** (2014). Automated generation of music playlists: Survey and experiments. *ACM Computing Surveys*, 47(2):26:1–26:35. DOI: *https://doi.org/10.1145/2652481*

Breiman, L. (2001). Random forests. *Machine Learning Journal*, 45(1):5–32. DOI: *https://doi.org/10.1023/A:1010933404324*

Cataltepe, Z., Yaslan, Y., and **Sonmez, A.** (2007). Music genre classification using MIDI and audio features. *EURASIP Journal of Applied Signal Processing*, 2007(1):150–150. DOI: *https://doi.org/10.1155/2007/36409*

Celma, Ò. (2010). *Music Recommendation and Discovery – The Long Tail, Long Fail, and Long Play in the Digital Music Space*. Springer. DOI: *https://doi.org/10.1007/978-3-642-13287-2*

Choi, K., Fazekas, G., Sandler, M. B., and **Cho, K.** (2017). Transfer learning for music classification and regression tasks. In *Proc. of the 18th International Society for Music Information Retrieval Conference, ISMIR*, pages 141–149.

Costa, Y. M. G., Oliveira, L. S., and **Silla Jr., C. N.** (2017). An evaluation of convolutional neural networks for music classification using spectrograms. *Applied Soft Computing*, 52:28–38. DOI: *https://doi.org/10.1016/j.asoc.2016.12.024*

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and **Bray, C.** (2004). Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22.

Dannenberg, R. B., Thom, B., and **Watson, D.** (1997). A machine learning approach to musical style recognition. In *Proc. of the International Computer Music Conference, ICMC*, pages 344–347.

Dhanaraj, R., and **Logan, B.** (2005). Automatic prediction of hit songs. In *Proc. of the 6th International Conference on Music Information Retrieval, ISMIR*, pages 488–491.

Doraisamy, S., Golzari, S., Norowi, N. M., Sulaiman, M. N., and **Udzir, N. I.** (2008). A study on feature selection and classification techniques for automatic genre classification of traditional Malay music. In Bello, J. P., Chew, E., and Turnbull, D., editors, *Proc. of the 9th International Conference on Music Information Retrieval, ISMIR*, pages 331–336.

Dunker, P., Nowak, S., Begau, A., and **Lanz, C.** (2008). Content-based mood classification for photos and music: A generic multi-modal classification framework and evaluation approach. In *Proc. of the 1st ACM SIGMM International Conference on Multimedia Information Retrieval, MIR*, pages 97–104. DOI: *https://doi.org/10.1145/1460096.1460114*

Fiebrink, R., and **Fujinaga, I.** (2006). Feature selection pitfalls and music classification. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR)*, pages 340–341.

Fujinaga, I. (1998). Machine recognition of timbre using steady-state tone of acoustic musical instruments. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 207–210.

Guyon, I., Nikravesh, M., Gunn, S., and **Zadeh, L. A.,** editors (2006). *Feature Extraction: Foundations and Applications*, volume 207 of *Studies in Fuzziness and Soft Computing*. Springer, Berlin Heidelberg. DOI: *https://doi.org/10.1007/978-3-540-35488-8*

Harris, Z. S. (1954). Distributional structure. *WORD*, 10(2–3):146–162. DOI: *https://doi.org/10.1080/00437956.1954.11659520*

Hastie, T., Tibshirani, R., and **Friedman, J.** (2009). *The Elements of Statistical Learning*. Springer, New York. DOI: *https://doi.org/10.1007/978-0-387-84858-7*

Hu, X., Choi, K., and **Downie, J. S.** (2017). A framework for evaluating multimodal music mood classification. *Journal of the Association for Information Science and Technology*, 68(2):273–285. DOI: *https://doi.org/10.1002/asi.23649*

Huang, Y., Lin, S., Wu, H., and **Li, Y.** (2014). Music genre classification based on local feature selection using a self-adaptive harmony search algorithm. *Data Knowledge Engineering*, 92:60–76. DOI: *https://doi.org/10.1016/j.datak.2014.07.005*

Jannach, D., Vatolkin, I., and **Bonnin, G.** (2017). Music data: Beyond the signal level. In Weihs, C., Jannach, D., Vatolkin, I., and Rudolph, G., editors, *Music Data Analysis: Foundations and Applications*, pages 197–215. CRC Press.

Knees, P., and **Schedl, M.** (2013). A survey of music similarity and recommendation from music context data. *ACM Transactions on Multimedia Computing, Communications and Applications*, 10(1):2:1–2:21. DOI: *https://doi.org/10.1145/2542205.2542206*

Kohavi, R., and **John, G. H.** (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1–2):273–324. DOI: *https://doi.org/10.1016/S0004-3702(97)00043-X*

Kudo, M., and **Sklansky, J.** (2000). Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition*, 33(1):25–41. DOI: *https://doi.org/10.1016/S0031-3203(99)00041-2*

Lamere, P. (2008). Social tagging and music information retrieval. *Journal of New Music Research*, 37(2):101–114. DOI: *https://doi.org/10.1080/09298210802479284*

Lartillot, O., and **Toiviainen, P.** (2007). MIR in Matlab (II): A toolbox for musical feature extraction from audio. In *Proc. of the 8th International Conference on Music Information Retrieval, ISMIR*, pages 127–130.

Laurier, C., Grivolla, J., and **Herrera, P.** (2008). Multimodal music mood classification using audio and lyrics. In *Seventh International Conference on Machine Learning and Applications*, pages 688–693. DOI: *https://doi.org/10.1109/ICMLA.2008.96*

Le, Q., and **Mikolov, T.** (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st*

Logan, B., Kositsky, A., and **Moreno, P.** (2004). Semantic analysis of song lyrics. In *IEEE International Conference on Multimedia and Expo, ICME*, volume 2, pages 827–830. DOI: *https://doi.org/10.1109/ICME.2004.1394328*

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110. DOI: *https://doi.org/10.1023/B:VISI.0000029664.99615.94*

Martí, R., Lozano, J. A., Mendiburu, A., and **Hernando, L.** (2018). Multi-start methods. In Martí, R., Pardalos, P. M., and Resende, M. G. C., editors, *Handbook of Heuristics*, pages 155–175. Springer International Publishing, Cham. DOI: *https://doi.org/10.1007/978-3-319-07124-4_1*

Martin, R., and **Nagathil, A. M.** (2009). Cepstral modulation ratio regression (CMRARE) parameters for audio signal analysis and classification. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, pages 321–324. DOI: *https://doi.org/10.1109/ICASSP.2009.4959585*

Mauch, M., and **Levy, M.** (2011). Structural change on multiple time scales as a correlate of musical complexity. In Klapuri, A. and Leider, C., editors, *Proc. of the 12th International Society for Music Information Retrieval Conference, ISMIR*, pages 489–494.

Mayer, R., and **Rauber, A.** (2010). Multimodal aspects of music retrieval: Audio, song lyrics – and beyond? In Ras, Z. W. and Wieczorkowska, A., editors, *Advances in Music Information Retrieval*, pages 333–363. Springer. DOI: *https://doi.org/10.1007/978-3-642-11674-2_15*

Mayer, R., Rauber, A., de León, P. J. P., Pérez-Sancho, C., and **Iñesta, J. M.** (2010). Feature selection in a Cartesian ensemble of feature subspace classifiers for music categorisation. In *Proc. of the 3rd International Workshop on Machine Learning and Music (MML)*, pages 53–56. ACM. DOI: *https://doi.org/10.1145/1878003.1878021*

McFee, B., and **Lanckriet, G. R. G.** (2012). Hypergraph models of playlist dialects. In *Proc. of the 13th International Society for Music Information Retrieval Conference, ISMIR*, pages 343–348.

McKay, C., Burgoyne, J. A., Hockman, J., Smith, J. B. L., Vigliensoni, G., and **Fujinaga, I.** (2010). Evaluating the genre classification performance of lyrical features relative to audio, symbolic and cultural features. In *Proc. of the 11th International Society for Music Information Retrieval Conference, ISMIR*, pages 213–218.

McKay, C., Cumming, J., and **Fujinaga, I.** (2018). jSymbolic 2.2: Extracting features from symbolic music for use in musicological and MIR research. In *Proc. of the 19th International Society for Music Information Retrieval Conference, ISMIR*, pages 348–354.

International Conference on International Conference on Machine Learning (ICML), volume 32, pages 1188–1196. *JMLR.org*.

Lim, S.-C., Lee, J.-S., Jang, S.-J., Lee, S.-P., and **Kim, M. Y.** (2012). Music-genre classification system based on spectro-temporal features and feature selection. *IEEE Transactions on Consumer Electronics*, 58(4):1262–1268. DOI: *https://doi.org/10.1109/TCE.2012.6414994*

McKay, C., and **Fujinaga, I.** (2006). Musical genre classification: Is it worth pursuing and how can it be improved? In *Proc. of the 7th International Conference on Music Information Retrieval, ISMIR*, pages 101–106.

McKay, C., and **Fujinaga, I.** (2008). Combining features extracted from audio, symbolic and cultural sources. In *Proc. of the 9th International Conference on Music Information Retrieval, ISMIR*, pages 597–602.

Meseguer-Brocal, G., Cohen-Hadria, A., and **Peeters, G.** (2018). DALI: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm. In *Proc. of the 19th International Society for Music Information Retrieval Conference, ISMIR*, pages 431–437.

Müller, M., and **Ewert, S.** (2011). Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features. In *Proc. of the 12th International Society for Music Information Retrieval Conference, ISMIR*, pages 215–220.

Neumayer, R., and **Rauber, A.** (2007). Integration of text and audio features for genre classification in music information retrieval. In *Proc. of the 29th European Conference on IR Research, ECIR*, pages 724–727. DOI: *https://doi.org/10.1007/978-3-540-71496-5_78*

Oramas, S., Barbieri, F., Nieto, O., and **Serra, X.** (2018). Multimodal deep learning for music genre classification. *Transactions of the International Society for Music Information Retrieval*, 1(1):4–21. DOI: *https://doi.org/10.5334/tismir.10*

Oramas, S., Nieto, O., Barbieri, F., and **Serra, X.** (2017). Multi-label music genre classification from audio, text and images using deep features. In *Proc. of the 18th International Society for Music Information Retrieval Conference, ISMIR*, pages 23–30.

Orio, N., Rizo, D., Miotto, R., Schedl, M., Montecchio, N., and **Lartillot, O.** (2011). MusiCLEF: A benchmark activity in multimodal music information retrieval. In *Proc. of the 12th International Society for Music Information Retrieval Conference, ISMIR*, pages 603–608.

Panda, R., Malheiro, R., Rocha, B., Oliveira, A., and **Paiva, R. P.** (2013). Multi-modal music emotion recognition: A new dataset, methodology and comparative analysis. In *Proc. of the 10th International Symposium on Computer Music Multidisciplinary Research, CMMR*. Springer.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., [illegible]

A., editors, *Algorithms from and for Nature and Life – Classification and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 529–537. Springer. DOI: *https://doi.org/10.1007/978-3-319-00035-0_54*

Saari, P., Eerola, T., and **Lartillot, O.** (2011). Generalizability and simplicity as criteria in feature selection: Application to mood classification in music. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1802–1812. DOI: *https://doi.org/10.1109/TASL.2010.2101596*

Schindler, A. (2019). *Multi-Modal Music Information Retrieval: Augmenting Audio-Analysis with Visual Computing for Improved Music Video Analysis*. PhD thesis, Faculty of Informatics, TU Wien.

Schreiber, H. (2015). Improving genre annotations for the Million Song Dataset. In *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR*, pages 241–247.

Sigtia, S., and **Dixon, S.** (2014). Improved music feature learning with deep neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pages 6959–6963. DOI: *https://doi.org/10.1109/ICASSP.2014.6854949*

Silla Jr., C. N., Koerich, A. L., and **Kaestner, C. A. A.** (2009). A feature selection approach for automatic music genre classification. *International Journal of Semantic Computing*, 3(2):183–208. DOI: *https://doi.org/10.1142/S1793351X09000719*

Simonetta, F., Ntalampiras, S., and **Avanzini, F.** (2019). Multimodal music information processing and retrieval: Survey and future challenges. In *Proc. of the International Workshop on Multilayer Music Representation and Processing, MMRP*, pages 10–18. DOI: *https://doi.org/10.1109/MMRP.2019.00012*

Sturm, B. L. (2012a). A survey of evaluation in music genre recognition. In *10th International Workshop on Adaptive Multimedia Retrieval:Semantics, Context, and Adaptation, AMR*, pages 29–66. DOI: *https://doi.org/10.1007/978-3-319-12093-5_2*

Sturm, B. L. (2012b). [illegible]

Brucher, M., Perrot, M., and **Duchesnay, E.** (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Raffel, C. (2016). *Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching*. PhD thesis, Graduate School of Arts and Sciences, Columbia University. DOI: *https://doi.org/10.1109/ICASSP.2016.7471641*

Řehůřek, R., and **Sojka, P.** (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50.

Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, 3:1371–1382.

Rötter, G., Vatolkin, I., and **Weihs, C.** (2013). Computational prediction of high-level descriptors of music personal categories. In Lausen, B., den Poel, D. V., and Ultsch,

Sturm, B. L. (2012b). Two systems for automatic music genre recognition: What are they really recognizing? In *Proc. of the 2nd International ACMWorkshop on Music Information Retrieval with User-Centered and Mulitmodal Strategies, MIRUM*, pages 69–74. DOI: *https://doi.org/10.1145/2390848.2390866*

Sturm, B. L. (2013a). Classification accuracy is not enough. *Journal of Intelligent Information Systems*, 41(3):371–406. DOI: *https://doi.org/10.1007/s10844-013-0250-y*

Sturm, B. L. (2013b). Evaluating music emotion recognition: Lessons from music genre recognition? In *IEEE International Conference on Multimedia and Expo Workshops, ICMEW*, pages 1–6. DOI: *https://doi.org/10.1109/ICMEW.2013.6618342*

Tzanetakis, G., and **Cook, P. R.** (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302. DOI: *https://doi.org/10.1109/TSA.2002.800560*

Vatolkin, I. (2015). Exploration of two-objective scenarios on supervised evolutionary feature selection: A survey and a case study (application to music categorisation). In *Proc. of the 8th International Conference on Evolutionary Multi-Criterion Optimization*, pages 529–543. Springer. DOI: *https://doi.org/10.1007/978-3-319-15892-1_36*

Vatolkin, I., Bonnin, G., and **Jannach, D.** (2014). Comparing audio features and playlist statistics for music classification. In *Analysis of Large and Complex Data – Second European Conference on Data Analysis, ECDA*, pages 437–447. DOI: *https://doi.org/10.1007/978-3-319-25226-1_37*

Vatolkin, I., Preuß, M., and **Rudolph, G.** (2011). Multiobjective feature selection in music genre and style recognition tasks. In Krasnogor, N. and Lanzi, P. L., editors, *Proc. of the 13th Annual Genetic and Evolutionary Computation Conference (GECCO)*, pages 411–418. ACM Press. DOI: *https://doi.org/10.1145/2001576.2001633*

Vatolkin, I., Rudolph, G., and **Weihs, C.** (2015). Evaluation of album effect for feature selection in music genre recognition. In *Proc. of the 16th International Society for Music Information Retrieval Conference, ISMIR*, pages 169–175.

Vatolkin, I., Theimer, W. M., and **Botteck, M.** (2010). AMUSE (Advanced MUSic Explorer): A multitool framework for music data analysis. In *Proc. of the 11th International Society for Music Information Retrieval Conference, ISMIR*, pages 33–38.

Weihs, C., Jannach, D., Vatolkin, I., and **Rudolph, G.,** editors (2017). *Music Data Analysis: Foundations and Applications*. CRC Press. DOI: *https://doi.org/10.1201/9781315370996*

Wilkes, B. (2019). Analyse von bild-, text- und audiobasierten Merkmalen für die Klassifikation von Musikgenres. Master's thesis, Department of Computer Science, TU Dortmund.

Witten, I. H., Frank, E., Hall, M. A., and **Pal, C. J.** (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, Burlington, Massachusetts.

Zangerle, E., Tschuggnall, M., Wurzinger, S., and **Specht, G.** (2018). Alf-200k: Towards extensive multimodal analyses of music tracks and playlists. In Pasi, G., Piwowarski, B., Azzopardi, L., and Hanbury, A., editors, *Advances in Information Retrieval*, pages 584–590. Springer. DOI: *https://doi.org/10.1007/978-3-319-76941-7_48*

Zitzler, E. (2012). Evolutionary multiobjective optimization. In Rozenberg, G., Bäck, T., and Kok, J. N., editors, *Handbook of Natural Computing, Volume 2*, pages 871–904. Springer, Berlin Heidelberg. DOI: *https://doi.org/10.1007/978-3-540-92910-9_28*

TO CITE THIS ARTICLE:
Vatolkin, I., & McKay, C. (2022). Multi-Objective Investigation of Six Feature Source Types for Multi-Modal Music Classification. *Transactions of the International Society for Music Information Retrieval*, 5(1), pp. 1–19. DOI: *https://doi.org/10.5334/tismir.67*

Submitted: 17 June 2020 Accepted: 02 December 2021 Published: 24 January 2022