STABILITY OF SYMBOLIC FEATURE GROUP IMPORTANCE IN THE CONTEXT OF MULTI-MODAL MUSIC CLASSIFICATION

Igor Vatolkin
TU Dortmund University
Department of Computer Science
igor.vatolkin@udo.edu

Cory McKay
Marianopolis College
Department of Liberal and Creative Arts
cory.mckay@mail.mcgill.ca

ABSTRACT

Multi-modal music classification creates supervised models trained on features from different sources (modalities): the audio signal, the score, lyrics, album covers, expert tags, etc. A concept of "multi-group feature importance" not only helps to measure the individual relevance of features of a feature type under investigation (such as the instruments present in a piece), but also serves to quantify the potential for further improving classification by adding features from other feature types or extracted from different kinds of sources, based on a multi-objective analysis of feature sets after evolutionary feature selection. In this study, we investigate the stability of feature group importance when different classification methods and different measures of classification quality are applied. Since musical scores are particularly helpful in deriving semantically meaningful, robust genre characteristics, we focus on the feature groups analyzed by the jSymbolic feature extraction software, which describe properties associated with instrumentation, basic pitch statistics, melody, chords, tempo, and other rhythmic aspects. These symbolic features are analyzed in the context of musical information drawn from five other modalities, and experiments are conducted involving two datasets, one small and one large. The results show that, although some feature groups can remain similarly important compared to others, differences can also be evident in various applications, and can depend on the particular classifier and evaluation measure being used. Insights drawn from this type of analysis can potentially be helpful in effectively matching specific features or feature groups to particular classifiers and evaluation measures in future feature-based MIR research.

1. INTRODUCTION AND RELATED WORK

There are music research scenarios where manually designed features can have value, and where gaining understanding of which features are particularly effective for various classification tasks can be of central importance. Researchers seeking to gain domain knowledge can be interested in more than just optimizing performance. Those working on composer attribution or genre classification, for example, might be interested in learning about the specific qualities that delineate musical style, so comparing the performance of different musically meaningful features can provide important insight. There can also be situations where only very limited training data are available, such as when there is only so much extant music by a given composer, and where data augmentation techniques can only improve matters so much. In such situations, deep learning that self-learns features can be less useful, and handmade features become valuable, as do approaches for selecting more promising features when many are available.

Multi-modal features drawn from a variety of data sources can be of particular interest. Although there can be redundancy across types of musical data (e.g., both audio and symbolic data specify pitch), such information is not always equally accessible (e.g., pitch is harder to extract from dense polyphonic audio), and there are other times when different source types contain complementary information (e.g., album art can provide cultural information that is inaccessible from audio or scores). Multi-modal research often plays an essential role in musicology, and involving sources as diverse as concert programs, manuscript illuminations, critical accounts, contemporary visual portrayals, and scores themselves. MIR research can and has similarly taken advantage of multi-modal information, both for improving performance in an engineering sense and in learning more about the connections between different types of musical data in a scientific or musicological sense. There is a rich MIR literature involving two modalities, but relatively few studies have considered more [1–6], and relatively few multi-modal public datasets are available (standouts are described in [5, 7–10]). There are also several papers that provide useful summaries of multi-modal MIR research [11–14].

Of course, with more types of data come more features, and the curse of dimensionality becomes a concern. Although handmade features can help, since they tend to represent information more concisely than raw data, too many can still present problems when training data are limited. Although there is substantial literature on dimensionality reduction (e.g., [15]), and important related MIR studies [4, 16–23], some approaches sacrifice feature independence (e.g., PCA), which compromises interpretability,

and feature selection approaches can be sensitive to particular classifier and evaluation methodologies.

So, it can be of value to get a sense of how consistently features perform across different classification and evaluation methodologies, in order to gain a deeper musical understanding of which features and groups of features are most important in delineating classes of interest. This paper is concerned with exactly this: attempting to measure the stability in importance of various features for a given music classification task, which can ultimately help to avoid overfitting the choice of features (or modalities), and can provide insights into underlying musical meaning.

We focus here on symbolic features, as they tend to be particularly interpretable, especially when collaborating with musicologists or theorists. This research maintains a fundamentally multi-modal character, as it also uses information from five other source types to ground feature stability measurements in a broader context: we used the same audio, album cover image, playlist co-occurrence, semantic tag, and lyric text feature data as [10]. These data are particularly useful because they include both a small but clean dataset and a large but noisy dataset, permitting stability to be measured in both scenarios. All symbolic features were extracted with jSymbolic [24], which provides access to many features usefully grouped into feature types whose relative stability can be compared.

We selected genre classification as the test domain for this research, but there is nothing about the techniques we propose that is specific to genre. Furthermore, the same techniques could just as easily be refocused on modalities other than symbolic music. It is also important to note that there are fundamental concerns related to the evaluation of musical classification [25–28]; although there is insufficient space to detail these here, this is essential reading for MIR researchers working in classification.

2. MULTI-GROUP FEATURE IMPORTANCE

We have proposed the concept of *multi-group feature importance* in [10], as an extension of earlier work [29–31]. A feature selection scenario based on two objectives was constructed, with the goal of simultaneously minimizing balanced relative error m_{BRE} (defined as a mean of relative errors for positive and negative instances) and maximizing the share of features g_i belonging to a given group i under investigation, such as rhythmic descriptors. This approach makes it possible to answer not only the questions "what is the lowest error rate achieved by rhythmic features in predicting a musical category?" and "which are the best rhythmic features for the current prediction task?", but also "how can performance be improved when features of other groups, domains, or modalities are also introduced?"

While the formal details of "importance" and the mathematical backing of multi-objective optimization are left to [10], we will briefly illustrate the core ideas here with the help of Figure 1. The connected circles in Subfigure (a) show a *non-dominated front* after feature selection, which simultaneously maximizes the share of rhythmic features g_{RHYTHM} and minimizes m_{BRE}, when features from all six modalities are considered (i.e. not just other symbolic features). Each circle corresponds to a feature set that is not *dominated* by any other feature set. This means that no other feature sets have been identified after feature selection that are "better" than the one under consideration, in combined terms of the two measures being optimized (i.e., using *a greater share* of rhythmic descriptors and achieving *smaller* classification error at the same time). For instance, the feature set in the upper right corner contains only rhythmic descriptors ($g_{RHYTHM} = 1.0$), but $m_{BRE} = 0.4236$ is rather high. No other feature sets that contain only rhythmic descriptors have lower errors, meaning that they are dominated by the upper right corner set, and are not shown in the plot. The feature set in the bottom left corner has the smallest $m_{BRE} = 0.0139$, but has only 37.86% rhythmic descriptors. Other circled feature sets between these corner solutions consist of various trade-offs between the two measures, and are also not dominated by any other found feature sets. The non-dominated fronts in the figures are constructed after ten feature selection repetitions and an independent evaluation on the reserved test set created with music tracks used neither for training the models nor for evaluating the selected feature sets.

Subfigure (a) indicates that rhythmic descriptors do not perform well for Traditional Blues music. This is not only because the error is high in general (this can occur, for example, if a category is too hard to predict or is badly defined), but also because the error is substantially reduced if other features (e.g., audio features) are allowed to contribute to the feature sets used to train the classification models. So, the overall *multi-group feature importance* of rhythmic descriptors is low, which is indicated graphically by the large area shaded grey on the graph. As the number of all possible non-empty feature sets is typically very high ($2^N - 1$ for N features), an evolutionary algorithm is proposed in [10] to explore many different combinations of features in a non-deterministic way, guided by a process inspired by natural evolution, where feature sets can be randomly changed by *mutation* that switches feature dimensions on or off. Only fitter feature sets survive after an exit condition is fulfilled, such as a cap on the overall number of mutations.

In [10], we have also investigated the importances of different modalities and proposed the complementary concept of *multi-group feature redundancy*. However, despite a large number of experiments, the results should be treated with caution because of two limitations of the study: only random forest classifiers were used and balanced relative error was the only measure of classification quality. It was therefore unclear whether estimated importances would remain similar if other classification methods were applied or other evaluation measures used. The present study addresses both of these gaps, by investigating how stable multi-group feature performance is when classification method or evaluation measure are varied.

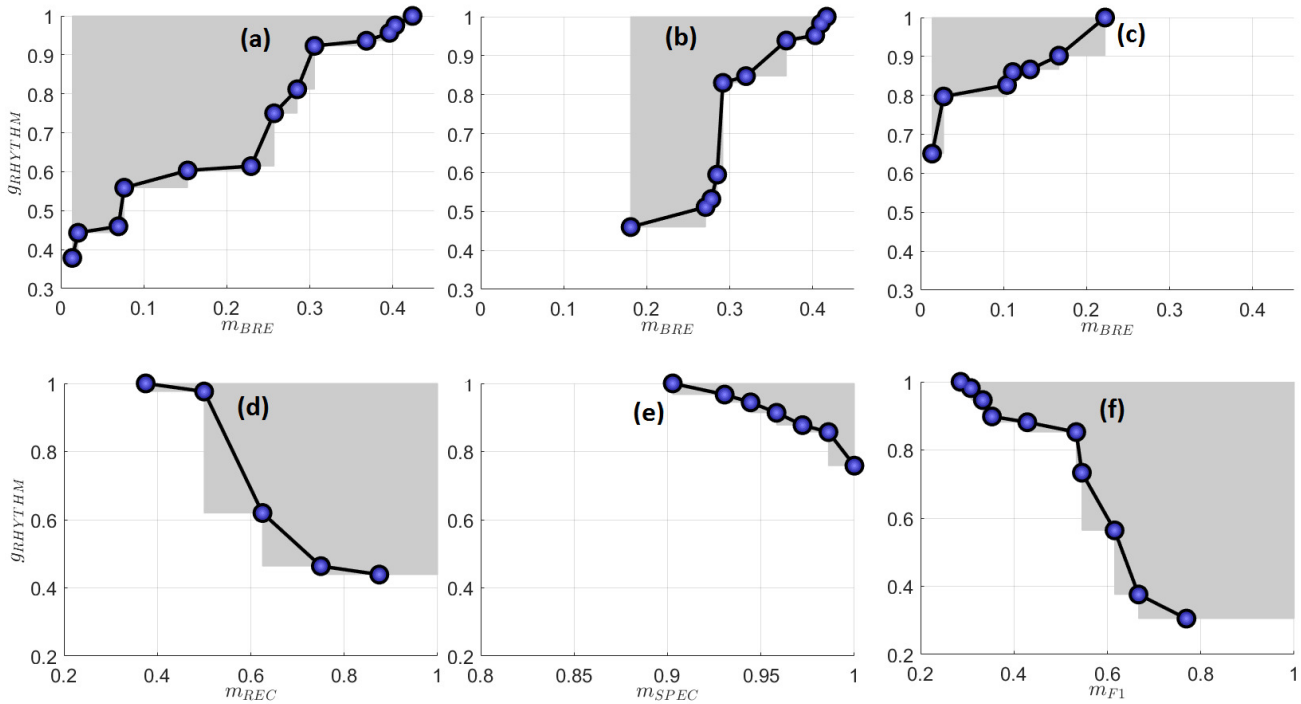

Figure 1. Non-dominated fronts after multi-objective feature selection for the identification of Traditional Blues in the SLAC dataset [2], with rhythmic descriptors the feature group being focused on. Top row: the share of rhythmic descriptors g_{RHYTHM} is maximized and the balanced relative error m_{BRE} is minimized using random forest (a), k-nearest neighbors (b), or support vector machine (c) classifiers. Bottom row: random forest classifiers are used to maximize both g_{RHYTHM} and recall m_{REC} (d), specificity m_{SPEC} (e), or F1-measure m_{F1} (f).

3. DESIGN OF EXPERIMENTS

We conducted our experiments on two pre-existing datasets, namely LMD-aligned [32] and SLAC [2], involving a total of 20 genres and sub-genres. While LMD-aligned is a larger dataset, it has problems with noisiness and structure (e.g., it is unbalanced, with an overly large representation of Pop/Rock songs). SLAC is smaller but carefully designed with, for example, an equal number of pieces belonging to each genre, each of which can be broken into two equally represented sub-genres. We made use of multi-modal features extracted from 1,575 Lakh and 250 SLAC tracks, which have been made publicly available. [1]

Table 1 shows sample jSymbolic features, divided into feature groups; the full list of symbolic (and other) features is available online. [2] We excluded the dynamics features used in [10], as dynamics are often inconsistently encoded.

In the first part of this study we measure the importance of eight groups of symbolic features (relative to features from all six modalities considered) not only with the random forests (RF) classifiers used in [10], but also with k-nearest neighbor (kNN) and support vector machine (SVM) classifiers. As can be observed in Figure 1, subfigures (a)-(c), notable differences in importance can be evident when the classification method is changed. Of course, a limitation of this study is that all classifiers are applied with default hyperparameters in the AMUSE framework [33]: 100 trees for RF, $k = 1$ for kNN and a linear kernel for SVM. In practice, varying hyperparameters may well also introduce meaningful variance in measured feature importances.

We chose to omit deep neural networks from our experiments, despite their popularity in MIR, for two reasons. First, they typically learn their own features, which can make it difficult to analyze the relative relevance of interpretable semantic descriptors for a given musical category. Second, the typically very large number of parameters they involve can lead to overfitting in situations with limited available data, such as when a musical category is defined by a small number of "positive" and "negative" tracks, as in real-world situations where a listener may wish to provide only a few labeled examples to train a supervised classification model.

In the second part of this study we vary the evaluation measures used in the two-objective feature selection. Although m_{BRE} is generally a good choice for binary classification tasks, as it can help to measure performance with unbalanced test sets [34, p.344], it is not often used in MIR classification studies. We have therefore extended the setup from [10] with three additional evaluation criteria. Recall and specificity measure classification errors associated with, respectively, instances annotated in the ground truth as belonging or not belonging to a category [34, p.342]. F1-measure is a weighted combination of precision and recall that, like m_{BRE}, can be useful for less

[1] https://zenodo.org/record/5651429
[2] https://doi.org/10.5334/tismir.67.s1

Group	Feature Examples
Pitch	First pitch, last pitch, major or minor, pitch class histogram, pitch variability, range
Melodic	Amount of arpeggiation, direction of melodic motion, melodic intervals, repeated notes
Chords	Chord type histogram, dominant seventh chords, variability of number of simultaneous pitches
Rhythm	Initial time signature, metrical diversity, note density per quarter note, prevalence of dotted notes
Tempo	Initial tempo, mean tempo, minimum and maximum note duration, note density and its variation
Instrument presence	Note prevalences of pitched and unpitched instruments, pitched instruments present
Instrument prevalence	Prevalences of individual instruments/instrument groups: acoustic guitar, string ensemble, etc.
Texture	Average number of independent voices, parallel fifths and octaves, voice overlap

Table 1. Sample jSymbolic [24] features grouped into eight semantically meaningful groups.

balanced datasets [34, p.344]. The potential for differences in measured importances based on recall, specificity, and F1 is demonstrated in subfigures (d), (e), and (f).

We conducted 24,000 feature selection experiments involving 20 genres and sub-genres $\times$ 3 cross-validation folds $\times$ 8 feature groups $\times$ 5 new combinations of classifiers and measures $\times$ 10 statistical repetitions, with runtimes between about 2 and 50 hours per experiment.

4. DISCUSSION OF RESULTS

Tables 2 and 3 show multi-group feature importances for all 20 classes, with each value indicating results aggregated over three folds. In each row, the importance values across the eight feature groups are shown with color: the group with the highest importance (compared to the other feature groups in the same row) is marked in deep red, and the group with the lowest importance in deep blue. The four "more important" feature groups are shown in shades of red, and four "less important" groups in shades of blue.

The instrument presence feature group seems to be the most important group in most of the experiments, with the instrument prevalence group being the most important in a few others. Features measuring pitch statistics are the most important in five experiments (e.g., RnB prediction using kNN and m_{BRE}). Interestingly, instrument prevalence is the least important group for almost all LMD-aligned genres (even when instrument presence is the most important group). Another intriguing result is that melodic features seem to be particularly unimportant for the Classical SLAC genre and its ClassBaroq and ClassRomant sub-genres.

As anticipated in Figure 1, these results reveal differences between importances when the classifier is varied. Such difference may be particularly meaningful when a cell's color changes from red to blue or vice versa: for example, chords are among the four more important groups for predicting Country using RF and SVM, but are among the less important groups when kNN is used. On the other hand, for this same genre tempo-based features become more important using kNN, not only relatively (the cell goes from blue to red), but also with respect to mean importance values (0.861 instead of 0.656 and 0.651, respectively). This supports our preliminary suspicion that the choice of classifier can have a strong impact, even for "robust" features groups (e.g., symbolic characteristics arguably describe musical properties in concise and clearly understood ways compared to audio descriptors). However, this is certainly not always the case: for 480 combinations of 20 classes, 8 feature groups, and 3 classifiers, in 62 cases (12.92%) the cell shade remains the same when the classifier is changed for a fixed genre and feature group.

When the evaluation measures are varied for a fixed classifier (note that here results were restricted to RF), the changes seem to be slightly less impactful: for 640 related combinations (20 classes, 8 feature groups, 4 measures), the shade changed in only 53 cases (8.28%).

A randomly chosen decision will assign a feature group to be either "more" or "less" relatively important with an expected probability of 25% for three classifiers (1 case with all more important values, 1 case with all less important values, and 6 remaining cases), and with 12.5% for the four measures. Although this interpretation is not perfect, as the change of a light red shade to a light blue shade will indicate a switch between "more" and "less" important, our main inference from the complete study is that it is not enough to claim that some feature group is "generally" more or less important for a particular musical category. Our results suggest that it is indeed necessary to accompany feature group evaluations with specification of the classifier and evaluation measure used. More general claims about the suitability of particular feature groups should be substantiated with broader experiments involving multiple classifiers and evaluation measures (and, perhaps, classifier hyper-parameters).

5. CONCLUSIONS

This paper studied the stability of multi-group feature importance when classifiers and evaluation measures are varied. Eight symbolic feature groups were focused on within a multi-modal classification context involving features extracted from six modalities (symbolic and five others). Various combinations of features and classification approaches were used to predict genres and sub-genres for two datasets with publicly available pre-extracted feature values. The results show that, although in most cases the relative importance of individual feature groups is not affected by the choice of classification method or evaluation measure, either or both of these do nonetheless have an influence in a non-negligible number of cases. Multiple parameters of the feature importance estimation chain can impact determination of which musical properties are more or less relevant for a particular musical category.

In the future, we plan to continue our experiments by

		Pitch	Melodic	Chords	Rhythm	Tempo	Instr. Pres.	Instr. Prev.	Texture
Country	RF	0.796±0.05	0.675±0.07	0.798±0.02	0.820±0.04	0.656±0.03	0.945±0.01	0.614±0.02	0.700±0.03
	kNN	0.924±0.08	0.704±0.07	0.832±0.03	0.865±0.04	0.861±0.06	0.952±0.03	0.645±0.02	0.779±0.03
	SVM	0.756±0.03	0.688±0.04	0.785±0.04	0.730±0.03	0.651±0.05	0.861±0.03	0.633±0.01	0.665±0.02
	RF-F1	0.610±0.09	0.439±0.03	0.576±0.10	0.654±0.05	0.413±0.11	0.889±0.01	0.285±0.12	0.381±0.08
	RF-Rec	0.849±0.01	0.722±0.06	0.823±0.03	0.882±0.02	0.755±0.08	0.965±0.01	0.668±0.06	0.733±0.08
	RF-Spec	0.760±0.03	0.678±0.06	0.768±0.04	0.785±0.04	0.618±0.06	0.927±0.02	0.604±0.09	0.674±0.03
Electronic	RF	0.825±0.04	0.774±0.02	0.815±0.01	0.871±0.02	0.754±0.07	0.951±0.02	0.697±0.03	0.746±0.02
	kNN	0.867±0.07	0.748±0.11	0.861±0.06	0.914±0.06	0.954±0.01	0.964±0.02	0.700±0.04	0.708±0.05
	SVM	0.824±0.04	0.773±0.01	0.770±0.05	0.836±0.05	0.768±0.00	0.916±0.01	0.683±0.03	0.758±0.03
	RF-F1	0.695±0.05	0.583±0.05	0.681±0.02	0.771±0.02	0.542±0.08	0.934±0.01	0.450±0.01	0.553±0.02
	RF-Rec	0.877±0.03	0.751±0.08	0.830±0.05	0.911±0.03	0.738±0.12	0.944±0.01	0.718±0.04	0.759±0.04
	RF-Spec	0.757±0.09	0.738±0.05	0.748±0.02	0.835±0.02	0.742±0.04	0.956±0.01	0.660±0.03	0.733±0.01
Pop	RF	0.752±0.05	0.694±0.07	0.735±0.04	0.806±0.04	0.701±0.04	0.927±0.02	0.626±0.06	0.713±0.04
	kNN	0.946±0.04	0.898±0.01	0.908±0.01	0.982±0.01	0.961±0.03	0.993±0.00	0.837±0.06	0.906±0.08
	SVM	0.799±0.08	0.718±0.08	0.777±0.02	0.780±0.04	0.781±0.01	0.875±0.02	0.738±0.02	0.795±0.04
	RF-F1	0.788±0.02	0.668±0.06	0.782±0.02	0.786±0.03	0.732±0.05	0.916±0.04	0.649±0.05	0.748±0.01
	RF-Rec	0.803±0.10	0.729±0.04	0.769±0.11	0.870±0.07	0.722±0.13	0.959±0.02	0.643±0.03	0.745±0.09
	RF-Spec	0.575±0.08	0.510±0.10	0.593±0.11	0.612±0.06	0.497±0.04	0.899±0.03	0.475±0.09	0.569±0.08
RnB	RF	0.808±0.04	0.689±0.10	0.807±0.05	0.844±0.03	0.716±0.02	0.958±0.02	0.621±0.09	0.697±0.07
	kNN	0.921±0.06	0.601±0.08	0.778±0.00	0.892±0.06	0.818±0.20	0.920±0.06	0.402±0.02	0.743±0.07
	SVM	0.764±0.07	0.699±0.03	0.780±0.04	0.788±0.04	0.735±0.01	0.889±0.03	0.663±0.10	0.691±0.07
	RF-F1	0.711±0.06	0.418±0.16	0.603±0.03	0.661±0.03	0.448±0.17	0.892±0.00	0.306±0.07	0.417±0.03
	RF-Rec	0.818±0.01	0.778±0.03	0.834±0.03	0.861±0.04	0.738±0.06	0.972±0.00	0.665±0.04	0.679±0.08
	RF-Spec	0.829±0.06	0.683±0.02	0.738±0.02	0.832±0.02	0.704±0.04	0.930±0.01	0.601±0.07	0.637±0.02
Rock	RF	0.726±0.06	0.602±0.03	0.713±0.07	0.748±0.03	0.655±0.03	0.917±0.02	0.598±0.02	0.657±0.05
	kNN	0.949±0.05	0.856±0.03	0.919±0.02	0.961±0.04	0.983±0.01	0.984±0.01	0.813±0.05	0.948±0.02
	SVM	0.768±0.03	0.756±0.01	0.766±0.02	0.780±0.03	0.710±0.04	0.886±0.04	0.625±0.03	0.686±0.04
	RF-F1	0.767±0.03	0.641±0.08	0.742±0.01	0.773±0.05	0.679±0.08	0.928±0.02	0.608±0.03	0.663±0.04
	RF-Rec	0.808±0.04	0.693±0.10	0.766±0.05	0.887±0.01	0.784±0.06	0.968±0.01	0.641±0.07	0.731±0.01
	RF-Spec	0.612±0.08	0.446±0.04	0.586±0.08	0.603±0.05	0.482±0.07	0.859±0.06	0.509±0.12	0.508±0.05
Blues	RF	0.923±0.03	0.742±0.07	0.858±0.06	0.902±0.03	0.759±0.05	0.975±0.00	0.769±0.10	0.723±0.13
	kNN	0.743±0.06	0.650±0.20	0.659±0.11	0.841±0.04	0.725±0.14	0.945±0.00	0.576±0.04	0.624±0.07
	SVM	0.792±0.02	0.557±0.14	0.665±0.20	0.795±0.08	0.584±0.04	0.916±0.01	0.621±0.13	0.627±0.03
	RF-F1	0.806±0.09	0.551±0.20	0.656±0.11	0.801±0.07	0.534±0.04	0.953±0.02	0.645±0.03	0.525±0.12
	RF-Rec	0.984±0.01	0.865±0.07	0.910±0.10	0.982±0.02	0.891±0.05	0.994±0.00	0.904±0.09	0.865±0.12
	RF-Spec	0.943±0.03	0.826±0.10	0.954±0.02	0.899±0.01	0.894±0.01	0.984±0.00	0.898±0.05	0.857±0.06
Classical	RF	0.934±0.02	0.782±0.04	0.962±0.01	0.949±0.01	0.872±0.02	0.996±0.00	0.945±0.05	0.885±0.05
	kNN	0.858±0.04	0.690±0.08	0.844±0.04	0.829±0.04	0.811±0.01	0.990±0.01	0.945±0.05	0.791±0.08
	SVM	0.872±0.01	0.706±0.06	0.915±0.01	0.836±0.04	0.837±0.04	0.979±0.02	0.963±0.03	0.812±0.03
	RF-F1	0.848±0.04	0.679±0.04	0.910±0.04	0.878±0.04	0.722±0.04	0.989±0.01	0.889±0.10	0.738±0.08
	RF-Rec	0.987±0.01	0.925±0.07	0.989±0.02	0.982±0.02	0.952±0.03	1.000±0.00	1.000±0.00	0.983±0.03
	RF-Spec	0.953±0.02	0.836±0.07	0.981±0.02	0.960±0.00	0.892±0.04	0.992±0.01	0.925±0.07	0.905±0.06
Rock	RF	0.933±0.00	0.887±0.03	0.911±0.07	0.917±0.03	0.833±0.05	0.995±0.00	0.915±0.04	0.905±0.02
	kNN	0.791±0.15	0.716±0.09	0.753±0.08	0.802±0.13	0.596±0.32	0.995±0.00	0.804±0.01	0.702±0.11
	SVM	0.808±0.02	0.827±0.09	0.816±0.09	0.797±0.04	0.758±0.03	0.982±0.00	0.866±0.06	0.749±0.06
	RF-F1	0.799±0.03	0.705±0.06	0.843±0.05	0.804±0.02	0.628±0.05	0.984±0.01	0.787±0.04	0.743±0.03
	RF-Rec	0.960±0.03	0.971±0.03	0.982±0.03	0.969±0.03	0.910±0.07	1.000±0.00	1.000±0.00	0.974±0.02
	RF-Spec	0.956±0.03	0.878±0.01	0.955±0.01	0.947±0.01	0.870±0.02	0.994±0.00	0.928±0.02	0.941±0.06
Jazz	RF	0.967±0.01	0.924±0.02	0.965±0.03	0.966±0.00	0.900±0.02	0.996±0.00	0.908±0.02	0.888±0.06
	kNN	0.886±0.03	0.566±0.05	0.764±0.10	0.772±0.11	0.754±0.11	0.985±0.01	0.606±0.11	0.618±0.23
	SVM	0.898±0.03	0.832±0.04	0.888±0.05	0.777±0.05	0.821±0.04	0.985±0.00	0.820±0.01	0.813±0.04
	RF-F1	0.913±0.01	0.815±0.05	0.929±0.05	0.898±0.04	0.837±0.10	0.990±0.00	0.805±0.02	0.748±0.10
	RF-Rec	0.972±0.02	0.966±0.04	0.991±0.01	0.988±0.02	0.933±0.04	0.998±0.00	0.981±0.03	0.927±0.01
	RF-Spec	0.970±0.02	0.952±0.01	0.983±0.02	0.967±0.01	0.971±0.01	0.999±0.00	0.940±0.01	0.941±0.01
Rap	RF	0.967±0.01	0.920±0.02	0.930±0.05	0.928±0.02	0.861±0.02	0.996±0.00	0.849±0.03	0.912±0.04
	kNN	0.955±0.01	0.731±0.10	0.839±0.04	0.828±0.04	0.797±0.07	0.979±0.00	0.711±0.09	0.555±0.06
	SVM	0.880±0.04	0.705±0.07	0.887±0.06	0.793±0.01	0.788±0.04	0.955±0.01	0.798±0.07	0.823±0.07
	RF-F1	0.918±0.03	0.796±0.01	0.818±0.10	0.847±0.07	0.748±0.06	0.992±0.01	0.762±0.03	0.737±0.21
	RF-Rec	0.990±0.02	0.936±0.01	0.988±0.02	0.959±0.06	0.936±0.03	0.999±0.00	0.960±0.04	0.969±0.04
	RF-Spec	0.975±0.02	0.945±0.01	0.939±0.06	0.950±0.03	0.898±0.03	0.999±0.00	0.938±0.02	0.910±0.05

Table 2. Multi-group symbolic feature importances for five LMD-aligned genres (top half of the table) and five SLAC parent genres (bottom half of the table), aggregated over 3 folds. F1: F1-measure; Rec: recall; Spec: specificity. The first three rows of each genre block were evaluated with balanced relative error m_{BRE}.

measuring the impact of other feature types and modalities. We will also further examine the effects of varying other parameters in the experimental setup, such as classifier hyper-parameters, and also systematically consider as-

		Pitch	Melodic	Chords	Rhythm	Tempo	Instr. Pres.	Instr. Prev.	Texture
BluesModern	RF	0.922±0.05	0.711±0.07	0.775±0.04	0.701±0.18	0.732±0.19	0.965±0.04	0.826±0.05	0.837±0.14
	kNN	0.762±0.07	0.772±0.13	0.631±0.08	0.698±0.03	0.773±0.08	0.966±0.04	0.536±0.07	0.774±0.16
	SVM	0.865±0.03	0.602±0.02	0.660±0.16	0.758±0.12	0.641±0.13	0.946±0.03	0.684±0.17	0.614±0.07
	RF-F1	0.916±0.07	0.488±0.39	0.603±0.40	0.704±0.33	0.511±0.29	0.929±0.09	0.510±0.27	0.711±0.22
	RF-Rec	0.769±0.20	0.629±0.17	0.640±0.07	0.498±0.30	0.895±0.10	0.991±0.02	0.885±0.20	0.794±0.07
	RF-Spec	0.984±0.01	0.966±0.02	0.970±0.03	0.982±0.02	0.963±0.01	0.997±0.00	0.928±0.02	0.962±0.04
BluesTradit	RF	0.836±0.10	0.664±0.10	0.909±0.02	0.949±0.03	0.778±0.18	0.971±0.01	0.844±0.09	0.673±0.14
	kNN	0.798±0.12	0.619±0.14	0.755±0.09	0.851±0.07	0.768±0.10	0.935±0.04	0.805±0.18	0.537±0.10
	SVM	0.844±0.05	0.672±0.14	0.618±0.15	0.773±0.14	0.748±0.10	0.964±0.05	0.598±0.13	0.608±0.11
	RF-F1	0.852±0.04	0.595±0.23	0.645±0.13	0.852±0.04	0.629±0.24	0.960±0.04	0.694±0.19	0.505±0.16
	RF-Rec	0.872±0.10	0.747±0.17	0.943±0.10	0.905±0.14	0.944±0.10	0.999±0.00	0.624±0.45	0.782±0.38
	RF-Spec	0.993±0.01	0.911±0.05	0.916±0.05	0.971±0.03	0.918±0.10	0.991±0.01	0.922±0.09	0.855±0.14
ClassBaroq	RF	0.914±0.07	0.747±0.26	0.921±0.06	0.804±0.11	0.815±0.14	0.997±0.00	0.816±0.03	0.860±0.14
	kNN	0.821±0.06	0.486±0.40	0.754±0.14	0.654±0.03	0.572±0.18	0.985±0.01	0.964±0.04	0.421±0.12
	SVM	0.901±0.02	0.595±0.04	0.856±0.05	0.774±0.06	0.625±0.04	0.985±0.01	0.909±0.08	0.740±0.10
	RF-F1	0.851±0.21	0.520±0.14	0.701±0.30	0.688±0.21	0.593±0.08	0.987±0.02	0.728±0.18	0.539±0.17
	RF-Rec	0.995±0.01	0.861±0.13	0.971±0.05	0.875±0.22	0.778±0.29	1.000±0.00	0.883±0.11	0.976±0.02
	RF-Spec	0.970±0.02	0.922±0.08	0.986±0.01	0.988±0.01	0.954±0.04	1.000±0.00	1.000±0.00	0.934±0.07
ClassRomant	RF	0.831±0.12	0.710±0.14	0.940±0.04	0.893±0.07	0.781±0.12	1.000±0.00	0.895±0.14	0.670±0.09
	kNN	0.888±0.09	0.670±0.25	0.789±0.12	0.947±0.05	0.817±0.14	0.983±0.01	0.804±0.13	0.722±0.23
	SVM	0.817±0.07	0.723±0.18	0.952±0.04	0.860±0.05	0.738±0.04	0.983±0.02	0.971±0.02	0.746±0.11
	RF-F1	0.744±0.04	0.510±0.23	0.864±0.09	0.855±0.03	0.579±0.13	1.000±0.00	0.731±0.10	0.621±0.09
	RF-Rec	0.792±0.14	0.724±0.24	0.974±0.02	0.848±0.16	0.480±0.25	0.999±0.00	1.000±0.00	0.806±0.09
	RF-Spec	0.979±0.02	0.942±0.05	0.999±0.00	0.993±0.01	0.980±0.01	0.999±0.00	0.986±0.01	0.971±0.02
JazzBop	RF	0.957±0.01	0.767±0.15	0.855±0.04	0.824±0.08	0.863±0.03	0.966±0.03	0.818±0.10	0.865±0.03
	kNN	0.761±0.11	0.604±0.07	0.838±0.04	0.751±0.02	0.711±0.08	0.994±0.00	0.713±0.16	0.572±0.06
	SVM	0.844±0.01	0.720±0.06	0.840±0.02	0.784±0.06	0.816±0.06	0.973±0.03	0.713±0.12	0.759±0.05
	RF-F1	0.788±0.03	0.609±0.35	0.722±0.16	0.811±0.03	0.681±0.03	0.962±0.05	0.571±0.15	0.568±0.09
	RF-Rec	0.949±0.04	0.653±0.33	0.774±0.06	0.904±0.11	0.965±0.03	0.954±0.08	1.000±0.00	0.839±0.18
	RF-Spec	0.994±0.01	0.985±0.02	0.995±0.01	0.994±0.01	0.991±0.01	1.000±0.00	0.960±0.05	0.930±0.05
JazzSwing	RF	0.957±0.01	0.863±0.07	0.911±0.06	0.946±0.05	0.932±0.04	0.947±0.04	0.876±0.05	0.861±0.08
	kNN	0.939±0.07	0.840±0.10	0.887±0.08	0.825±0.12	0.707±0.14	0.983±0.01	0.738±0.05	0.749±0.22
	SVM	0.870±0.06	0.823±0.05	0.903±0.07	0.840±0.07	0.807±0.06	0.946±0.03	0.822±0.07	0.800±0.10
	RF-F1	0.868±0.07	0.658±0.05	0.788±0.10	0.888±0.03	0.604±0.16	0.943±0.03	0.759±0.16	0.650±0.08
	RF-Rec	0.988±0.02	0.942±0.06	0.840±0.16	0.943±0.08	0.901±0.12	0.962±0.03	0.867±0.13	0.828±0.26
	RF-Spec	0.994±0.01	0.966±0.01	0.978±0.02	0.996±0.01	0.964±0.02	0.999±0.00	0.978±0.00	0.979±0.02
RapHardcore	RF	0.845±0.00	0.790±0.07	0.898±0.06	0.880±0.03	0.688±0.18	0.965±0.01	0.693±0.11	0.883±0.11
	kNN	0.816±0.09	0.685±0.14	0.801±0.13	0.725±0.07	0.661±0.06	0.969±0.03	0.677±0.24	0.686±0.02
	SVM	0.768±0.10	0.779±0.05	0.820±0.13	0.718±0.18	0.712±0.12	0.913±0.07	0.719±0.10	0.813±0.06
	RF-F1	0.842±0.14	0.727±0.05	0.649±0.23	0.949±0.05	0.838±0.09	0.978±0.01	0.576±0.39	0.842±0.16
	RF-Rec	0.660±0.27	0.700±0.24	0.981±0.02	0.985±0.03	0.888±0.07	0.970±0.01	0.829±0.15	0.954±0.05
	RF-Spec	0.993±0.01	0.959±0.01	0.985±0.01	0.987±0.02	0.964±0.03	1.000±0.00	0.972±0.02	0.989±0.01
RapPop	RF	0.851±0.08	0.637±0.03	0.926±0.04	0.883±0.06	0.787±0.19	0.953±0.01	0.644±0.16	0.658±0.17
	kNN	0.907±0.09	0.629±0.11	0.790±0.12	0.736±0.04	0.732±0.12	0.987±0.01	0.643±0.17	0.668±0.08
	SVM	0.854±0.07	0.776±0.08	0.863±0.09	0.682±0.10	0.625±0.32	0.895±0.06	0.511±0.17	0.756±0.09
	RF-F1	0.835±0.12	0.626±0.38	0.862±0.11	0.755±0.22	0.637±0.32	0.971±0.03	0.710±0.25	0.582±0.21
	RF-Rec	0.972±0.05	0.761±0.18	0.910±0.16	0.529±0.09	0.603±0.41	0.957±0.07	0.734±0.06	0.681±0.48
	RF-Spec	0.982±0.01	0.965±0.04	0.961±0.02	0.961±0.04	0.960±0.02	1.000±0.00	0.935±0.04	0.944±0.03
RockAltern	RF	0.787±0.06	0.768±0.10	0.815±0.07	0.891±0.02	0.824±0.08	0.971±0.01	0.728±0.09	0.681±0.05
	kNN	0.658±0.14	0.678±0.07	0.639±0.10	0.830±0.14	0.751±0.08	0.952±0.05	0.423±0.24	0.593±0.19
	SVM	0.670±0.12	0.679±0.13	0.797±0.08	0.804±0.14	0.751±0.11	0.903±0.07	0.695±0.10	0.592±0.04
	RF-F1	0.535±0.20	0.458±0.05	0.764±0.14	0.740±0.05	0.597±0.13	0.911±0.07	0.442±0.15	0.498±0.24
	RF-Rec	0.578±0.19	0.794±0.19	0.855±0.04	0.884±0.03	0.847±0.22	0.942±0.04	0.698±0.08	0.765±0.10
	RF-Spec	0.977±0.03	0.918±0.06	0.987±0.02	0.980±0.03	0.907±0.11	1.000±0.00	0.963±0.04	0.934±0.01
RockMetal	RF	0.914±0.03	0.840±0.06	0.873±0.09	0.847±0.11	0.840±0.04	0.988±0.00	0.950±0.03	0.840±0.09
	kNN	0.986±0.02	0.797±0.12	0.797±0.18	0.835±0.04	0.720±0.30	0.985±0.01	0.795±0.17	0.601±0.17
	SVM	0.758±0.25	0.788±0.05	0.915±0.04	0.826±0.03	0.750±0.08	0.981±0.02	0.946±0.04	0.810±0.03
	RF-F1	0.943±0.04	0.605±0.14	0.742±0.15	0.739±0.18	0.630±0.15	0.985±0.01	0.802±0.04	0.775±0.04
	RF-Rec	0.965±0.03	0.703±0.23	0.795±0.26	0.757±0.16	0.823±0.05	0.991±0.01	1.000±0.00	0.807±0.16
	RF-Spec	0.995±0.01	0.972±0.02	0.985±0.00	0.992±0.01	0.965±0.03	1.000±0.00	0.992±0.01	0.990±0.01

Table 3. Multi-group symbolic feature importances for ten SLAC sub-genres, aggregated over 3 folds. F1: F1-measure; Rec: recall; Spec: specificity. The first three rows of each sub-genre block were evaluated with balanced relative error m_{BRE}.

pects like extraction times, statistical properties, and suitability for data augmentation. We will also run further trials in order to be able to apply more developed statistical significance testing.

6. ACKNOWLEDGMENTS

The experiments were carried out on the Linux HPC cluster at TU Dortmund (LiDO3), which was partially funded by the Large-Scale Equipment Initiative by the German Research Foundation (DFG) grant 271512359. The second author's work was funded by the Fonds de recherche du Québec – Société et culture, under grants 2021-CHZ-282456 and 2022-CHZ-309882.

7. REFERENCES

[1] C. McKay and I. Fujinaga, "Combining features extracted from audio, symbolic and cultural sources," in *Proc. of the 9th International Conference on Music Information Retrieval, ISMIR*, 2008, pp. 597–602.

[2] C. McKay, J. A. Burgoyne, J. Hockman, J. B. L. Smith, G. Vigliensoni, and I. Fujinaga, "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*, 2010, pp. 213–218.

[3] B. McFee and G. R. G. Lanckriet, "Hypergraph models of playlist dialects," in *Proc. of the 13th International Society for Music Information Retrieval Conference, ISMIR*, 2012, pp. 343–348.

[4] R. Panda, R. Malheiro, B. Rocha, A. Oliveira, and R. P. Paiva, "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, 2013.

[5] S. Oramas, O. Nieto, F. Barbieri, and X. Serra, "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*, 2017, pp. 23–30.

[6] S. Oramas, F. Barbieri, O. Nieto, and X. Serra, "Multimodal deep learning for music genre classification," *Transactions of the International Society for Music Information Retrieval*, vol. 1, no. 1, pp. 4–21, 2018.

[7] N. Orio, D. Rizo, R. Miotto, M. Schedl, N. Montecchio, and O. Lartillot, "Musiclef: a benchmark activity in multimodal music information retrieval," in *Proc. of the 12th International Society for Music Information Retrieval Conference, ISMIR*, 2011, pp. 603–608.

[8] G. Meseguer-Brocal, A. Cohen-Hadria, and G. Peeters, "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*, 2018, pp. 431–437.

[9] D. Bogdanov, A. Porter, H. Schreiber, J. Urbano, and S. Oramas, "The acousticbrainz genre dataset: Multi-source, multi-level, multi-label, and large-scale," in *Proc. of the 20th International Society for Music Information Retrieval Conference, ISMIR*, 2019, pp. 360–367.

[10] I. Vatolkin and C. McKay, "Multi-objective investigation of six feature source types for multi-modal music classification," *Transactions of the International Society for Music Information Retrieval*, vol. 5, no. 1, pp. 1–19, 2022.

[11] R. Mayer and A. Rauber, "Multimodal aspects of music retrieval: Audio, song lyrics - and beyond?" in *Advances in Music Information Retrieval*, Z. W. Ras and A. Wieczorkowska, Eds. Springer, 2010, pp. 333–363.

[12] P. Knees and M. Schedl, "A survey of music similarity and recommendation from music context data," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 10, no. 1, pp. 2:1–2:21, 2013.

[13] D. Jannach, I. Vatolkin, and G. Bonnin, "Music data: beyond the signal level," in *Music Data Analysis: Foundations and Applications*, C. Weihs, D. Jannach, I. Vatolkin, and G. Rudolph, Eds. CRC Press, 2017, pp. 197–215.

[14] F. Simonetta, S. Ntalampiras, and F. Avanzini, "Multimodal music information processing and retrieval: Survey and future challenges," in *Proc. of the International Workshop on Multilayer Music Representation and Processing, MMRP*, 2019, pp. 10–18.

[15] I. Guyon, M. Nikravesh, S. Gunn, and L. A. Zadeh, Eds., *Feature Extraction. Foundations and Applications*, ser. Studies in Fuzziness and Soft Computing. Berlin Heidelberg: Springer, 2006, vol. 207.

[16] I. Fujinaga, "Machine recognition of timbre using steady-state tone of acoustic musical instruments," in *Proc. of the International Computer Music Conference, ICMC*, 1998, pp. 207–210.

[17] R. Fiebrink and I. Fujinaga, "Feature selection pitfalls and music classification," in *Proc. of the 7th International Conference on Music Information Retrieval, ISMIR*, 2006, pp. 340–341.

[18] S. Doraisamy, S. Golzari, N. M. Norowi, M. N. Sulaiman, and N. I. Udzir, "A study on feature selection and classification techniques for automatic genre classification of traditional malay music," in *Proc. of the 9th International Conference on Music Information Retrieval, ISMIR*, J. P. Bello, E. Chew, and D. Turnbull, Eds., 2008, pp. 331–336.

[19] C. N. Silla Jr., A. L. Koerich, and C. A. A. Kaestner, "A feature selection approach for automatic music genre classification," *International Journal of Semantic Computing*, vol. 3, no. 2, pp. 183–208, 2009.

[20] R. Mayer, A. Rauber, P. J. P. de León, C. Pérez-Sancho, and J. M. Iñesta, "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*. ACM, 2010, pp. 53–56.

[21] P. Saari, T. Eerola, and O. Lartillot, "Generalizability and simplicity as criteria in feature selection: Application to mood classification in music," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1802–1812, 2011.

[22] S.-C. Lim, J.-S. Lee, S.-J. Jang, S.-P. Lee, and M. Y. Kim, "Music-genre classification system based on spectro-temporal features and feature selection," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 4, pp. 1262–1268, 2012.

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

[24] C. McKay, J. Cumming, and I. Fujinaga, "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*, 2018, pp. 348–354.

[25] B. L. Sturm, "A survey of evaluation in music genre recognition," in *Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation, 10th International Workshop, AMR*, 2012, pp. 29–66.

[26] ——, "Two systems for automatic music genre recognition: What are they really recognizing?" in *Proc. of the 2nd International ACM Workshop on Music Information Retrieval with User-Centered and Mulitmodal Strategies, MIRUM*, 2012, pp. 69–74.

[27] ——, "Classification accuracy is not enough," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 371–406, 2013.

[28] ——, "Evaluating music emotion recognition: Lessons from music genre recognition?" in *IEEE International Conf. on Multimedia and Expo Workshops, ICMEW*, 2013, pp. 1–6.

[29] I. Vatolkin, M. Preuß, and G. Rudolph, "Multi-objective feature selection in music genre and style recognition tasks," in *Proc. of the 13th Annual Genetic and Evolutionary Computation Conference, GECCO*, N. Krasnogor and P. L. Lanzi, Eds. ACM Press, 2011, pp. 411–418.

[30] I. Vatolkin, "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, EMO*. Springer, 2015, pp. 529–543.

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

[32] C. Raffel, "Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching," Ph.D. dissertation, Graduate School of Arts and Sciences, Columbia University, 2016.

[33] I. Vatolkin, P. Ginsel, and G. Rudolph, "Advancements in the music information retrieval framework AMUSE over the last decade," in *Proc. of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2021, pp. 2383–2389.

[34] C. Weihs, D. Jannach, I. Vatolkin, and G. Rudolph, Eds., *Music Data Analysis: Foundations and Applications*. CRC Press, 2016.