Can't trust the feeling?
How open data reveals unexpected behavior of high-level music descriptors

Liem, C.C.S.; Mostert, C.

Publication date
2020

Document Version
Final published version

Published in
Proceedings of the 21st International Society for Music Information Retrieval Conference

Citation (APA)
Liem, C. C. S., & Mostert, C. (2020). Can't trust the feeling? How open data reveals unexpected behavior of high-level music descriptors. In *Proceedings of the 21st International Society for Music Information Retrieval Conference* (pp. 240-247)

CAN'T TRUST THE FEELING? HOW OPEN DATA REVEALS UNEXPECTED BEHAVIOR OF HIGH-LEVEL MUSIC DESCRIPTORS

Cynthia C. S. Liem
Delft University of Technology
Delft, The Netherlands
c.c.s.liem@tudelft.nl

Chris Mostert
Delft University of Technology
Delft, The Netherlands
chrismostert@outlook.com

ABSTRACT

Copyright restrictions prevent the widespread sharing of commercial music audio. Therefore, the availability of resharable pre-computed music audio features has become critical. In line with this, the AcousticBrainz platform offers a dynamically growing, open and community-contributed large-scale resource of locally computed low-level and high-level music descriptors. Beyond enabling research reuse, the availability of such an open resource allows for renewed reflection on the music descriptors we have at hand: while they were validated to perform successfully under lab conditions, they now are being run 'in the wild'. Their response to these more ecological conditions can shed light on the degree to which they truly had construct validity. In this work, we seek to gain further understanding into this, by analyzing high-level classifier-based music descriptor output in AcousticBrainz. While no hard ground truth is available on what the true value of these descriptors should be, some oracle information can still be derived, relying on semantic redundancies between several descriptors, and multiple feature submissions being available for the same recording. We report on multiple unexpected patterns found in the data, indicating that the descriptor values should not be taken as absolute truth, and hinting at directions for more comprehensive descriptor testing that are overlooked in common machine learning evaluation and quality assurance setups.

1. INTRODUCTION

In many music information retrieval (MIR) applications, it is useful to include information related to music content. However, many large-scale music audio collections of interest cannot legally be shared as-is. As a compromise, efforts have been undertaken to locally pre-compute music audio descriptors and make these available through APIs or as part of research datasets. Parties without in-house access to large audio corpora need to rely on such data for subsequent use. Indeed, large-scale pre-computed descriptor corpora have been feeding into further machine learning pipelines, empowering music applications, facilitating benchmarking initiatives [1, 2], and leading to inferences and statements about the nature of music preferences and listening behavior at an unprecedented scale [3–6].

Audio-based music descriptors are commonly divided into low- and high-level descriptors. Low-level descriptors can closely be related to the audio signal, while high-level descriptors are more semantically understandable to humans. This does not make high-level descriptors easier to extract; many of them cannot objectively and directly be measured in the physical world, and thus consider *constructs* rather than physically measurable phenomena.

The performance of automated music descriptor extraction procedures is reported according to the common evaluation methodologies in the field. For descriptors based on supervised machine learning, this normally includes a performance report on a test set that was partitioned out of the original dataset and not seen during training, or on cross-validation outcomes. However, descriptors that are reported and assumed to be successful may still be prone to sensitivities not explicitly accounted for in their design and evaluation. In lower-level music descriptors, implementations of MFCC and chroma descriptors showed sensitivities to different audio encoding formats [7], while common textual descriptions of audio extractor pipelines turned out insufficiently specific to yield reproducible results [8]. For higher-level descriptors, seemingly well-performing trained music genre classifiers turned out to be unexpectedly sensitive to subtle, humanly interpretable audio transformations [9]. Such sensitivities are not restricted to music genre classification; for example, trade-offs between accuracy and semantic robustness have also been observed in deep music representations [10]. Generally, in many MIR tasks, ground truth relies on human judgement and labeling. This may be imprecise and subjective, leading to low inter-rater agreement. In its turn, this leads to questions on whether a clear-cut ground truth exists at all, while this often is fundamental to machine learning techniques and their evaluation [11–14].

Can we tell whether automated descriptors are as trustable as initially assumed? Do they truly measure what they are intended to measure? Do they match broader, less explicitly encoded assumptions we have on them? These are important questions to ask: in case of negative an-

swers, the descriptors may not provide a valid basis for subsequent work to build upon. However, finding sensitivities that were unnoticed in original evaluation contexts is non-trivial, requiring a broader, more meta-analytic perspective. In this work, we focus on this, by providing an analysis of music descriptor values obtained through the AcousticBrainz [15] platform. By soliciting community-contributed submissions of locally run, but largely standardized music feature extractors, the platform offers a large-scale perspective on music that 'people felt worth the upload'. As such, it offers a more ecological 'in-the-wild' data perspective than what was studied in the lab, when the descriptors were originally designed. Indeed, through cross-collection evaluation procedures employing independent ground truth validation sets, several well-known genre classification models were shown not to generalize well beyond their original evaluation datasets [16].

The AcousticBrainz data is unusually transparent and rich: more so than e.g. the popular Million Song Dataset [17]. Many descriptor fields are available for each submission, multiple submissions can be added for the same MusicBrainz recording, each submission is encoded with additional metadata on characteristics of the input audio and the extractor software, and the extractor software is open source [18]. We use this richness to comprehensively analyze existing computed descriptor values in AcousticBrainz. Rather than relying on explicit and clear-cut ground truth, we look at the data through a meta-scientific lens, and impose more general assumptions on descriptor behavior, inspired by psychological and software testing techniques. This way, we will reveal several unexpected patterns in the descriptor values. As original music audio is not attached to the descriptor entries, we will not (yet) be able to fully replicate how descriptors were computed, nor will we be able to recreate experimental conditions on this data, in which possible reasons for unexpected behavior can cleanly be statistically controlled. Still, our analysis will help in pinpointing concrete directions towards future controlled studies.

In the remainder of this paper, we will discuss related work in Section 2. Then, we will introduce the data used for our analyses in Section 3, after which we will present analyses into intra-dataset correlations (Section 4), descriptor stability (Section 5), and descriptor value distributions (Section 6), followed by the conclusion and an outlook towards future work.

2. RELATED WORK

In conducting science, it is non-trivial to assess whether the outcomes we are observing, the inferences we are making and the conclusions we are drawing are truly correct. These questions of *validity* were first acknowledged in the domain of psychological testing, where the focus was on measuring psychological constructs: abstracted human characteristics (e.g. 'conscientiousness') that are not directly and physically observable, but that can still be measured (e.g. through well-designed surveys). Various subcategories of validity exist [19]. Among these, one of the most intuitive to understand, yet hardest to pinpoint, is the notion of *construct validity*: the question whether a measurement procedure can indeed be considered to yield a *"measure of some attribute or quality which is not "operationally defined""* [20].

The traditional viewpoint on ways to assess construct validity, is to consider a measure procedure as part of a *nomological network*, and relate its outcomes to those of other procedures, that have previously been shown to be valid [20]; in practice, in much of psychological research, this is done by assessing correlations between construct measurements that are theorized to have an interpretable relation to one another. This does create dependencies under uncertainty, still boiling down to a philosophical question of 'what the first truth is to start with'—something that may be disproven during the research process, as more evidence will come in and further comparisons are being made. It has therefore been argued that comprehensive inquiry into construct validity will not only lead to better assessments, but also leads to fundamental questionings and improvements of the complete scientific process [21].

Within MIR, while comprehensive meta-scientific questions on this have not been asked, criticisms of current evaluation practices, referring to the notions of both validity and reliability and the way in which they have been used in the Information Retrieval field, have been presented by Urbano et al. [22]. In addition, Sturm's criticisms of 'horse systems' in MIR [9] (machine learning-based systems that performance-wise appear to make humanly intelligent decisions, but that turn out to pick up on irrelevant confounds in data) can again be related to construct validity.

As a method to assess whether a system is a 'horse system', Sturm proposes to investigate how systems react to input data transformations that are considered 'irrelevant' (i.e. imperceptible) to humans. Interestingly, this technique has been used in another research field focused on 'testing': the field of software testing, in which it would be called *metamorphic testing* [23]. While software testing appears to be a much more objective and precise procedure than psychological testing, from a formal, logical perspective, many real-life programs may actually be considered non-testable, and the problem of determining whether a software artefact is bug-free is undecidable [24]. While one cannot pinpoint one exact oracle truth, it still may be possible to *derive* partial oracle truth through transformations based on known data relationships [25], e.g. by applying input transformations that should not change a system's output, which is done in metamorphic testing.

3. ACOUSTICBRAINZ

In our studies, we study descriptor values as found through the AcousticBrainz platform. More specifically, we will depart from the most recent high-level descriptor data dump obtained through the AcousticBrainz website [1]. We are interested in the high-level descriptors, as they should

[1] `https://AcousticBrainz.org/download`.
The data dump used in our analyses is `AcousticBrainz-highlevel-json-20150130.tar.bz2`

mimic humanly understandable semantic concepts, which should be relatable in humanly interpretable ways.

The data dump considers 1,805,912 entries of community-contributed high-level descriptor values, that can be broken down into genres, moods, and other categories (e.g. danceability); a full overview can be found in ([15], Table 4). Unless indicated otherwise, our analyses will consider this full data dump. In all cases, descriptor values consider classification outputs, obtained through machine learning; for each possible class label within a descriptor (e.g., `jazz` in the genre_dortmund classifier), the classifier confidence for that class label is given as a float value. The performance of each of the classifiers is documented on the AcousticBrainz website; where possible, performance is reported on publicly available datasets [2] .

4. INTRA-DATASET CORRELATIONS

Following the psychological concept of the nomological network, one way to assess validity is to assess how the outcomes of related measurement procedures correlate with each other. For this, we take advantage of *semantic redundancy* within the AcousticBrainz high-level descriptors. For example, several musical genres literally re-occur as class labels within the various genre classifiers. Then, it is not unrealistic to assume that, given the same audio input, the output of alternative jazz classifiers should positively correlate. Furthermore, some 'softer' assumptions on meaningful relationships can be made: e.g., aggressive music is likely not relaxed, and happy music is likely not sad. We defined multiple of these relationships for which we would expect to observe (strong) positive correlations between classifier label predictions, and computed their Pearson correlations. The results are displayed in Table 1.

The found correlations were unexpected; we were especially surprised by the very low correlations found for the genre classifiers, while they should target the same concepts. A scatter plot of `rock` classifier confidences in genre_rosamerica and genre_tzanetakis (which yielded a negative correlation) is given in Figure 1. It appears that confidences outcomes do not uniformly distribute over the full [0.0, 1.0] confidence range; we will investigate this further in the following sections.

Out of all 'softer' assumptions that were compared, the lowest correlation (.13) is between `happy` and `not sad`, implying that music classified as happy could be sad at the same time. The classifiers used in AcousticBrainz indeed allow for this, as separate binary classifiers exist for happy and sad moods; however, this contradicts Russell's 2D circumplex model of affect [26], in which happiness and sadness would have opposite scores on the valence dimension.

5. STABILITY

Our correlation analyses showed unexpected results. However, as different classifiers were trained on different datasets, they may have considered different characteristics of the input data. Inspired by the idea of derived oracles,

[2] `https://AcousticBrainz.org/datasets/accuracy`

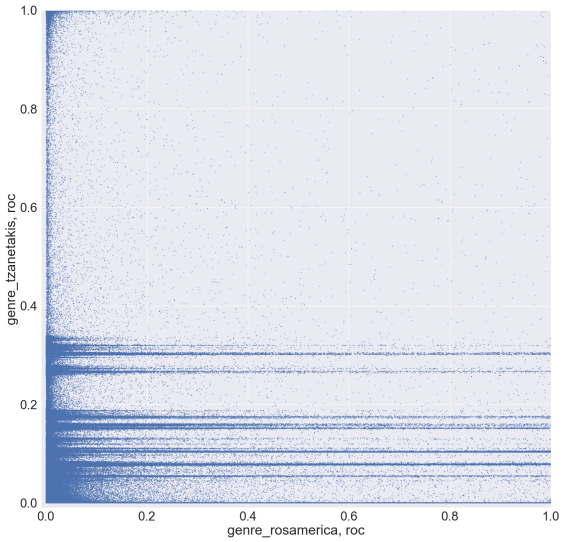


Figure 1: Scatter plot of classifier confidences. Each point indicates an AcousticBrainz submission, with confidences for genre_rosamerica, `roc` and genre_tzanetakis, `roc`.

we can however also consider relationships that should be closer to the identity, and thus should lead to (nearly) identical outcomes.

In AcousticBrainz, multiple submissions can be made for the same MusicBrainz recording ID (MBID). Semantically, a MusicBrainz recording really references one and the same recording. So while users may have encoded the recording audio in different ways, and may be using different versions of the feature extractor, we should intuitively be able to assume that re-submissions of one and the same recording should yield descriptor values that are very close to one another. In other words, we wish for re-submissions for the same MBID to display *stability*.

For this, we need to consider the MBIDs in our data dump that have more than one associated submission. Filtering for this led to a corpus of 941,018 submissions for 299,097 different MBIDs. If n submissions are available for a given MBID, a given classifier c and a given classifier label l, the corresponding classifier confidences for these submissions can now be grouped into a population (MBID, c, l) of size n. Considering we have k unique MBIDs in our dataset (in our case, k = 299,097), we can then enumerate the populations as $[(\text{MBID}_1, c, l), (\text{MBID}_2, c, l), ..., (\text{MBID}_k, c, l)]$, and operate within and/or across them when calculating instability metrics.

We consider two alternative ways to quantify instability. First, for each of the submission populations, we can compute the variance observed for classifier confidences, for each label l in classifier c. As there may be a varying amount of submissions within a population, we normalize for this by computing the *pooled variance* $\overline{\text{var}}(c, l)$ over our filtered corpus as follows:

$$\overline{\text{var}}(c, l) = \frac{\Sigma_{i=1}^{k}(n_i \times \text{var}((\text{MBID}_i, c, l)))}{\Sigma_{i=1}^{k} n_i} \quad (1)$$

Classifier, label A	Classifier, label B	Pearson's r	p
genre_rosamerica, `cla`	genre_tzanetakis, `cla`	.29	<.001
genre_dortmund, `rock`	genre_rosamerica, `roc`	.24	<.001
genre_dortmund, `jazz`	genre_rosamerica, `jaz`	.22	<.001
genre_dortmund, `pop`	genre_rosamerica, `pop`	.11	<.001
genre_dortmund, `jazz`	genre_tzanetakis, `jaz`	.08	<.001
genre_rosamerica, `pop`	genre_tzanetakis, `pop`	.06	<.001
genre_rosamerica, `hip`	genre_tzanetakis, `hip`	.05	<.001
genre_rosamerica, `jaz`	genre_tzanetakis, `jaz`	.02	<.001
genre_dortmund, `blues`	genre_tzanetakis, `blu`	.01	<.001
genre_dortmund, `pop`	genre_tzanetakis, `pop`	-.05	<.001
genre_dortmund, `rock`	genre_tzanetakis, `roc`	-.06	<.001
genre_rosamerica, `roc`	genre_tzanetakis, `roc`	-.07	<.001
mood_aggressive, `aggressive`	mood_relaxed, `not_relaxed`	.59	<.001
mood_acoustic, `acoustic`	mood_electronic, `not_electronic`	.58	<.001
danceability, `danceable`	mood_party, `party`	.53	<.001
mood_electronic, `electronic`	genre_dortmund, `electronic`	.48	<.001
danceability, `danceable`	genre_rosamerica, `dan`	.33	<.001
mood_happy, `happy`	mood_party, `party`	.20	<.001
mood_happy, `happy`	mood_sad, `not_sad`	.13	<.001

Table 1: Pearson correlations between high-level classifier outcomes, theorized to positively correlate with another.

where n_i is the sample size of the ith population in our enumeration.

As there are multiple possible labels within the same classifier, but we want to discuss outcomes at the classifier level, we then take the mean pooled variance, $\overline{\overline{\text{var}(c)}}$, over all possible labels $l \in L_c$ for classifier c.

When using variances, classifier confidences are considered to be informative. Alternatively, one could choose to rather consider each classifier label as a binary label. To reflect this perspective, for each population and for each classifier, we can compute the normalized information entropy $\hat{H}(\text{MBID}_i, c)$, which uses the Shannon entropy [27], but normalizes by the amount of possible labels $|L_c|$ for c:

$$\begin{aligned}&\hat{H}(\text{MBID}_i, c)\\ &= -\Sigma_{l \in L_c} \frac{P((\text{MBID}_i, c, l)) \log_2 P((\text{MBID}_i, c, l))}{\log_2 |L_c|}\\ &= -\Sigma_{l \in L_c} P((\text{MBID}_i, c, l)) log_{|L_c|} P((\text{MBID}_i, c, l))\end{aligned} \tag{2}$$

where $P((\text{MBID}_i, c, l))$ is the probability of label l in classifier c, following the observed empirical distribution within the population corresponding to MBID_i. Then, to have a weighted measure per classifier over the whole filtered corpus, we calculate the pooled normalized entropy $\overline{\hat{H}}(c)$, similarly to how we computed the pooled variance.

While we want for descriptor values to be stable within a submission, it is usually not the intention that for a given descriptor, the classifier would be so stable that it always predicts a single l throughout the whole corpus. This e.g. happens for the genre_dortmund classifier, which unrightfully classifies many AcousticBrainz submissions as electronic music, as also noticed in [16]. To quantify the unbiasedness of a classifier, we compute the normalized entropy for each classifier over our complete (unfiltered) corpus, denoted as $\hat{H}(c)_{all}$. A higher $\hat{H}(c)_{all}$ denotes a more uniform distribution over the different possible class labels for c across the corpus, and thus lower classifier bias.

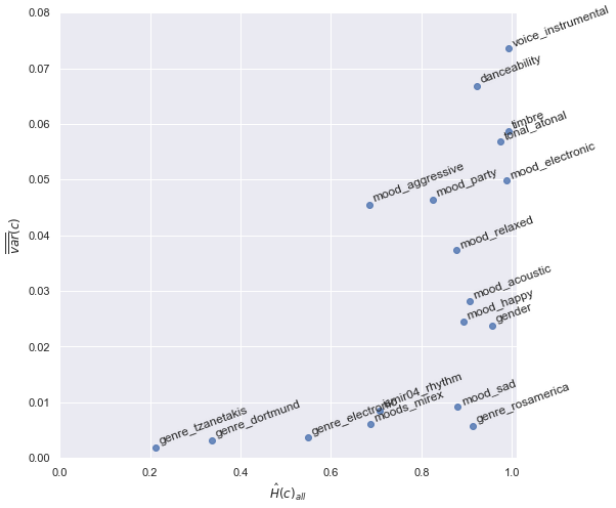
Plots in which we illustrate $\overline{\overline{\text{var}(c)}}$ and $\overline{\hat{H}}(c)$ (pooled with regard to recordings with multiple submissions) vs. $\hat{H}(c)_{all}$ (taken across the whole, unfiltered corpus) are shown in Figure 2. As we can see, indeed, the genre classifiers turn out stable but highly biased. While in most cases, observed trends are comparable for the two possible instability measures, some exceptions are found, most notably on the gender classifier, which is considered stable when using $\overline{\overline{\text{var}(c)}}$, but unstable when using $\overline{\hat{H}}(c)$. Seemingly, confidences for this classifier are close to 0.5, meaning that male/female classifications easily flip within a submission.

6. VALUE DISTRIBUTIONS

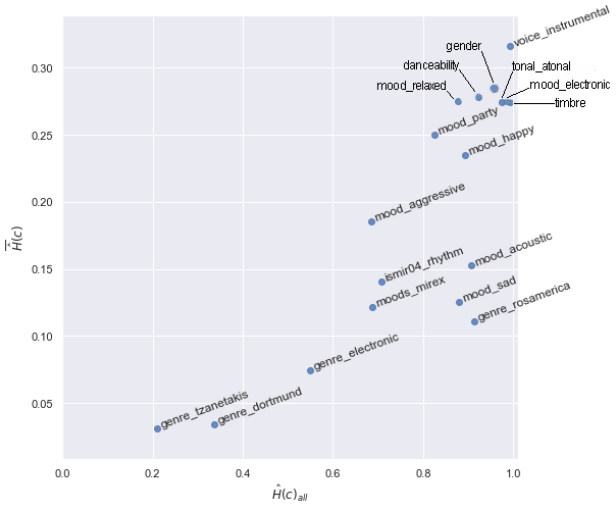
From Figure 1, it was observed that descriptor values clustered together in small bands. This behavior occurs for several genre and mood classifiers. To illustrate this, Figure 3 displays a histogram of descriptor values for the mood_acoustic, mood_relaxed, mood_electronic and mood_sad classifiers, as observed across the complete AcousticBrainz corpus. Some confidence values seem disproportionally represented: in the histogram, sharp spikes occur for mood_acoustic, mood_relaxed, mood_electronic, and a minor spike for mood_sad.

There are various reasons why this may be the case. Possibly, the community may have fed skewed data to the classifier. Alternatively, the feature extractor may have shown anomalous responses to specific inputs. For each submission, we have rich metadata, which e.g. includes information about audio codecs, bit rates, song lengths, and software library versions that were used when the submission was created. While, in the absence of a conscious experimental design underlying the data, we cannot cleanly test for contributions of individual facets, we still can examine *whether major distributional differences occur for submissions with scores within the anomalous-looking spikes, when comparing these to submissions with scores outside of these*.

For this, for each of the classifiers, we manually define range intervals for the classifier confidences, within which



(a) Instability based on mean pooled variance $\overline{\overline{\text{var}}(c)}$.



(b) Instability based on pooled normalized entropy $\overline{\hat{H}}(c)$.

Figure 2: Submission instability vs. corpus-wide unbiasedness ($\hat{H}(c)_{all}$).

we consider a submission to belong to an anomalous classifier confidence value spike. We then compare the metadata value distributions of submissions within each classifier spike to those of submissions that do not occur in any of the four anomalous spikes (1,239,882 submissions for 855,266 unique MBID recordings).

To investigate whether the observed anomalies may have been skewed towards any particular genre, we also study a subset of our corpus, which was cross-matched against the AcousticBrainz genre dataset [28]. More specifically, we only kept MBIDs which also occurred in all three publicly available ground truth sets (Discogs, last.fm and tagtraum) of the AcousticBrainz genre dataset, reducing the corpus to 402,279 submissions for 164,826 unique MBID recordings. Examining confidence value distributions for this filtered dataset, we still observed the same anomalous spikes for the same range intervals. Therefore, we will apply the same range intervals as before to select values associated to anomaly spikes, and will

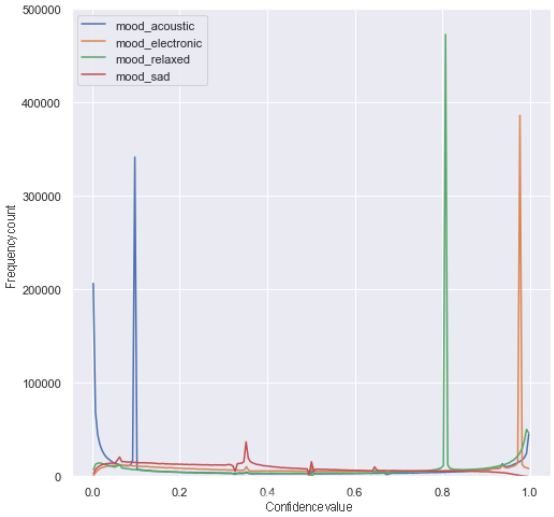


Figure 3: Histogram of descriptor values for several classifiers, considered across the whole corpus.

again compare distributional differences between these and non-anomalous submissions (now amounting to 267,394 submissions for 128,687 unique MBID recordings), in this case to see whether certain genres are overrepresented in the anomalous spikes. For each classifier of interest, an overview of anomalous spike interval ranges and counts of corresponding unique recording MBIDs and submissions is given in Table 2.

To quantify distributional differences, we use the Jensen-Shannon (JS) distance metric:

$$JS\_distance(p,q) = \sqrt{\frac{D(p\|m) + D(q\|m)}{2}} \quad (3)$$

where m is the pointwise mean of p and q and D is the Kullback-Leibler (KL) divergence [29]. The JS distance is based on the JS divergence [30]; as advantages over the KL divergence, the JS divergence is symmetric and always has a finite value within the [0, 1] range [31].

For each metadata category in our overall corpus, and for each genre category in our genre-filtered corpus, we calculate the JS distance between the frequency occurrence profiles of category values, counted over all submissions within an anomalous spike, vs. all submissions without any anomalous spike. As some categories can assume many different values (e.g. replay_gain), we only do comparisons for values that occur at least 10 times in both frequency profiles. JS distance values for the metadata comparisons are listed in Table 3, while JS distance for the genre comparisons are listed in Table 4.

As can be observed in Table 3, comparing submissions within and outside of the anomalous spikes, major distributional differences are found for used extractor software versions. These go up to the level of Essentia Git commit and build versions that were used for low-level feature extraction. In addition, we also observe distributional differences for bit_rate and codec, likely confirming earlier observations [7] that low-level feature extractors may display sensitivities with regard to different audio codecs and

Classifier	Anomalous range	Full		Genre	
		#MBIDs	#submissions	#MBIDs	#submissions
mood_acoustic, `acoustic`	[0.09, 0.10]	282,605	358,747	60,261	94,268
mood_relaxed, `relaxed`	[0.805, 0.815]	373,555	485,184	72,739	119,050
mood_electronic, `electronic`	[0.972, 0.982]	315,626	401,151	64,944	101,915
mood_sad, `sad`	[0.346, 0.362]	57,697	75,688	8,854	14,242

Table 2: Details of anomalous spike data slices used for distributional comparisons. For each classifier of interest, we indicate the classifier confidence range for which a submission was considered to be anomalous. We also list the counts of unique MBID recordings and overall submissions, both for the full corpus and our genre-filtered corpus.

	`acoustic`	`relaxed`	`electronic`	`sad`
bit_rate	.42	.32	.39	.17
codec	.34	.26	.32	.06
length	.15	.15	.15	.32
lossless	.28	.21	.27	.02
essentia_low	.61	.52	.59	.15
essentia_git_sha_low	.67	.58	.66	.23
essentia_build_sha_low	.70	.62	.69	.24

Table 3: JS distances between frequency profiles over metadata categories, for anomalous vs. non-anomalous submissions considering the four classifiers of interest. For metadata categories that are not listed, found JS distances were always 0.

	`acoustic`	`relaxed`	`electronic`	`sad`
Discogs	.12	.09	.11	.11
last.fm	.14	.12	.13	.14
tagtraum	.14	.11	.13	.14

Table 4: JS distances between frequency profiles over genre categories, for anomalous vs. non-anomalous submissions considering the four classifiers of interest.

compression rates. In contrast, Table 4 shows that JS distances are equivalent and low across genre taxonomies and types of anomalies: from this, it seems more likely that the anomalies were caused by submission extraction contexts, rather than the inclusion of anomalous data.

7. CONCLUSIONS AND FUTURE WORK

In this work, we analyzed patterns in high-level descriptor values in AcousticBrainz. As we showed, while the descriptors were successfully validated under lab conditions, they show unexpected behavior in the wild, raising questions on the extent to which they have construct validity.

The unexpected behavior could have two potential causes. First of all, **the construct underlying several high-level descriptors may be conceptually problematic by itself**. For example, the concept of genre [32], as well as its use in machine learning classification tasks [33] has been criticized by musicologists and musicians. Furthermore, within music psychology, there have been findings that sad music does not necessarily elicit sad emotions [34, 35]. Further interdisciplinary research will be needed to better understand these phenomena.

Our current analyses also accumulated evidence that **the AcousticBrainz community confronted the descriptors with audio and extraction contexts that were too different from the contexts on which classifiers originally were trained**. It should be noted that original training datasets for the classifiers were far smaller in size (several hundreds to thousands of data points) than the current scale of AcousticBrainz, and that this logically may not have managed capturing all intricacies of larger-scale, ecologically valid data. However, our analyses suggest that anomalous behavior may also be due to audio codecs, compression rates and different versions of software implementations and builds that were used during extraction, which are rarely explicitly considered and reported in evaluation setups. As for the software versions, it should further be noted that, while we focused on high-level descriptors, all found differences occurred in the extraction procedures of low-level descriptors (feature representations), while the high-level machine learning models stayed constant. Thus, low-level descriptor performance should explicitly stay in scope when studying high-level descriptors.

With this work, we wished to shed light on current challenges regarding the reproducibility and generalizability of research outcomes, and on elements of processing pipelines that are under-represented in applied machine learning and signal processing literature, yet play a critical role for the pipeline's performance [8, 36]. Inspired by literature in both psychological and software testing, we also offered several possible strategies to assess descriptor validity, even in the absence of a clear ground truth.

While we exposed several potentially problematic patterns, we explicitly do not wish for this work to be seen as a criticism of AcousticBrainz and/or Essentia. No other MIR resource or API currently offers similar levels of transparency that allow for analyses like we performed here, and we would like to explicitly thank the teams behind these initiatives for their openness. It also is this openness that will allow for us to perform further research in the near future—with more systematic testing strategies and experimental designs—towards more holistic quality assurance procedures for applied machine learning procedures in the context of humanly-interpretable signal data.

8. REFERENCES

[1] A. Schindler and R. Mayer and A. Rauber, "Facilitating Comprehensive Benchmarking Experiments on the Million Song Dataset," in *Proceedings of the 13th Conference of the International Society for Music Information Retrieval (ISMIR 2012)*, 2012.

[2] D. Bogdanov, A. Porter, J. Urbano, and H. Schreiber, "The MediaEval 2018 AcousticBrainz Genre Task: Content-based Music Genre Recognition from Multiple Sources," in *MediaEval Benchmark Workshop*, 2018.

[3] J. Serrà, A. Corral, M. Boguñá, M. Haro, and J. L. Arcos, "Measuring the Evolution of Contemporary Western Popular Music," *Scientific Reports*, vol. 2, 2012.

[4] M. Interiano, K. Kazemi, L. Wang, J. Yang, Z. Yu, and N. L. Komarova, "Musical trends and predictability of success in contemporary songs in and out of the top charts," *Royal Society Open Science*, vol. 5, no. 171274, 2018.

[5] M. Park, J. Thom, S. Mennicken, H. Cramer, and M. Macy, "Global music streaming data reveal diurnal and seasonal patterns of affective preference," *Nature Human Behaviour*, vol. 3, no. 3, pp. 230–236, 2019.

[6] E. Zangerle, R. Huber, M. Vötter, and Y.-H. Yang, "Hit Song Prediction: Leveraging Low-and High-Level Audio Features," in *Proceedings of the 20th Conference of the International Society for Music Information Retrieval (ISMIR 2019)*, 2019.

[7] J. Urbano, D. Bogdanov, P. Herrera, E. Gómez, and X. Serra, "What is the Effect of Audio Quality on the Robustness of MFCCs and Chroma Features?" in *Proceedings of the 15th Conference of the International Society for Music Information Retrieval (ISMIR 2014)*, 2014.

[8] B. McFee, J. W. Kim, M. Cartwright, J. Salamon, R. M. Bittner, and J. P. Bello, "Open-Source Practices for Music Signal Processing Research: Recommendations for Transparent, Sustainable, and Reproducible Audio Research," *IEEE Signal Processing Magazine*, vol. 36, 2019.

[9] B. L. Sturm, "A Simple Method to Determine if a Music Information Retrieval System is a "Horse"," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1636–1644, 2014.

[10] J. Kim, J. Urbano, C. C. S. Liem, and A. Hanjalic, "Are Nearby Neighbors Relatives? Testing Deep Music Embeddings," *Frontiers in Applied Mathematics and Statistics*, vol. 5, p. 53, 2019.

[11] A. Flexer and T. Grill, "The Problem of Limited Inter-rater Agreement in Modelling Music Similarity," *Journal of New Music Research*, vol. 45, no. 3, pp. 239–251, 2016.

[12] A. Flexer and T. Lallai, "Can we increase inter- and intra-rater agreement in modeling general music similarity?" in *Proceedings of the 20th Conference of the International Society for Music Information Retrieval (ISMIR 2019)*, 2019.

[13] H. V. Koops, W. B. de Haas, J. A. Burgoyne, J. Bransen, A. Kent-Muller, and A. Volk, "Annotator subjectivity in harmony annotations of popular music," *Journal of New Music Research*, vol. 48, no. 3, pp. 232–252, 2019.

[14] S. Balke, J. Abeßer, J. Driedger, C. Dittmar, and M. Müller., "Towards evaluating multiple predominant melody annotations in jazz recordings," in *Proceedings of the 17th Conference of the International Society for Music Information Retrieval (ISMIR 2016)*, 2016.

[15] A. Porter, D. Bogdanov, R. Kaye, R. Tsukanov, and X. Serra, "AcousticBrainz: A Community Platform for Gathering Music Information Obtained from Audio," in *Proceedings of the 16th Conference of the International Society for Music Information Retrieval (ISMIR 2015)*, 2015.

[16] D. Bogdanov, A. Porter, P. Herrera, and X. Serra, "Cross-collection evaluation for music classification tasks," in *Proceedings of the 17th Conference of the International Society for Music Information Retrieval (ISMIR 2016)*, 2016.

[17] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere, "The Million Song Dataset," in *Proceedings of the 12th Conference of the International Society for Music Information Retrieval (ISMIR 2011)*, 2011.

[18] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. R. Zapata, and X. Serra, "Essentia: An Audio Analysis Library for Music Information Retrieval," in *Proceedings of the 14th Conference of the International Society for Music Information Retrieval (ISMIR 2013)*, 2013.

[19] W. R. Shadish, T. D. Cook, and D. T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin, 2002.

[20] L. J. Cronbach and P. E. Meehl, "Construct Validity in Psychological Tests," *Psychological Bulletin*, vol. 52, p. 281–302, 1955.

[21] G. T. Smith, "On Construct Validity: Issues of Method and Measurement," *Psychological Assessment*, vol. 17, no. 4, pp. 396–408, 2005.

[22] J. Urbano, M. Schedl, and X. Serra, "Evaluation in Music Information Retrieval," *Journal of Intelligent Information Systems*, vol. 31, pp. 345–369, 2013.

[23] T. Y. Chen, S. C. Cheung, and S. M. Yiu, "Metamorphic Testing: A New Approach for Generating Next Test Cases," Hong Kong University of Science and Technology, Tech. Rep., 1998.

[24] E. J. Weyuker, "On Testing Non-testable Programs," *The Computer Journal*, vol. 25, no. 4, pp. 465–470, 1982.

[25] E. T. Barr, M. Harman, P. McMinn, M. Shahbaz, and S. Yoo, "The Oracle Problem in Software Testing: A Survey," *IEEE Transactions on Software Engineering*, vol. 41, no. 5, pp. 507–525, 2015.

[26] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 1980.

[27] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[28] 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 *Proceedings of the 20th Conference of the International Society for Music Information Retrieval (ISMIR 2019)*, 2019.

[29] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[30] D. Endres and J. Schindelin, "A new metric for probability distributions," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1858–1860, 2003.

[31] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[32] C. C. S. Liem, A. Rauber, T. Lidy, R. Lewis, C. Raphael, J. D. Reiss, T. Crawford, and A. Hanjalic, "Music Information Technology and Professional Stakeholder Audiences: Mind the Adoption Gap," in *Dagstuhl Follow-Ups*. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2012, vol. 3.

[33] B. L. Sturm, "Classification accuracy is not enough: On the evaluation of music genre recognition systems," *Journal of Intelligent Information Systems*, vol. 41, pp. 371–406, 2013.

[34] J. K. Vuoskoski and T. Eerola, "Can sad music really make you sad? Indirect measures of affective states induced by music and autobiographical memories." *Psychology of Aesthetics, Creativity, and the Arts*, vol. 6, no. 3, pp. 204–213, 2012.

[35] A. Kawakami, K. Furukawa, K. Katahira, and K. Okanoya, "Sad music induces pleasant emotion," *Frontiers in Psychology*, vol. 4, 2013.

[36] M. F. Dacrema, S. Boglio, P. Cremonesi, and D. Jannach, "A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research," *arXiv preprint arXiv:1911.07698*, 2019.