PREPRINT

Automatic subgenre classification in an electronic dance music taxonomy

Antonio Caparrini[a], Javier Arroyo[b,c], Laura Pérez-Molina[d] and Jaime Sánchez-Hernández[b]

[a]Management Solutions. Madrid, Spain.
[b]Facultad de Informática, Universidad Complutense de Madrid, Spain
[c]Instituto de Tecnología del Conocimiento. Universidad Complutense de Madrid, Spain
[d]Everis. Madrid, Spain

ARTICLE HISTORY
Compiled April 22, 2020

ABSTRACT
Electronic dance music (EDM) is a genre where thousands of new songs are released every week. The list of EDM subgenres considered is long, but it also evolves according to trends and musical tastes.

With this in view, we have retrieved two sets of over 2,000 songs separated by more than a year. Songs belong to the top 100 list of an EDM website taxonomy of more than 20 subgenres that changed in the period considered. We test the effectiveness of automatic classification on these sets and delve into the results to determine, for example, which subgenres perform better and worse, how the performance of some subgenres change in the two sets, or how some subgenres are often confused with one another. We illustrate confusion among subgenres by a graph and interpret it as a taxonomic map of EDM. We also assess the deterioration of the performance of the classifier of the first set when used to classify the second one. Finally, we study how the new subgenres that appear in the second set relate to the old ones with the help of the classifier of the first set.

As a result, this work illustrates the main challenges that EDM poses to automatic classification and provides insights into where are the limits of this approach.

KEYWORDS
music subgenre classification; machine learning; audio feature extraction; MIR; electronic dance music

1. Introduction

Both the music industry and listeners use genres as a practical way to classify songs according to their characteristics. Ascribing a song to a genre reflects that such a song shares a number of those characteristics with the songs of that genre. The characteristics that define a specific genre can be influenced by culture, and geographical or time factors, and this fact can lead to ill-defined genre labels (Scaringella, Zoia, & Mlynek, 2006).

However, genres are still instrumental, and they are widely employed for description and classification (Aucouturier & Pachet, 2003). This applies all the more to the case of

This article has been accepted for publication in Journal of New Music Research, published by Taylor Francis.
CONTACT Email: javier.arroyo@fdi.ucm.es

music in digital format, where digital repositories contain thousands of songs that need to be classified or labelled for identification and retrieval. In those cases, the process of manually assigning the genre label is time-consuming and complex (Aucouturier & Pachet, 2003).

As a result of this need, the field of automatic genre classification emerged almost twenty years ago (Li, Ogihara, & Li, 2003; Scaringella et al., 2006; Tzanetakis & Cook, 2002). Moreover, it has gained relevance over the years, as shown by the classification challenge (Defferrard, Mohanty, Carroll, & Salathé, 2018) that took place in the WWW 2018 Conference with the FMA data set (Defferrard, Benzi, Vandergheynst, & Bresson, 2017).

Despite this interest, automatic subgenre classification has been significantly less studied, probably because it focuses on a more restricted context. However, we consider that the case of electronic dance music (EDM) merit particular attention in automatic subgenre classification due to its distinguished features, explained below.

Firstly, due to EDM burgeoning production where hundreds or even thousands of new songs are added every week in online music shops. Given that some those shops have taxonomies containing more than 20 subgenres, automatic classification could alleviate the work needed to assign a subgenre label to each song.

Secondly, because subgenres evolve very rapidly in EDM. Producers, journalists, and consumers are always eager to promote new micro-genres (Collins, 2012). The subgenre proliferation is probably unequalled within other types of music. Vitos (2014) shows a well-informed analysis on this phenomenon related to some distinctive characteristics of EDM, such as the narrow interconnection between genre cultures and industry. New subgenres appear, and sometimes old subgenres are redefined as the music taste changes or change the terms used by the aficionados to refer to music styles (or even the intended meanings of such terms). While the appearance of new genres is assumed in genre taxonomies (Scaringella et al., 2006), in EDM these changes happen more intensely.

Finally, Vitos (2014) mentions that there is also an excessive taxonomy that obeys to DJs requirements of exact categorisations in order to use similar tracks in their performances (that will also be informative for the audiences). Over-categorisation may result in somewhat arbitrary and not unified taxonomies with ill-defined subgenres or subgenres defined according to extra-musical elements. In this respect, automatic classification can help to identify potentially ill-defined or redundant subgenres.

All these aspects make EDM a stimulating challenge for automatic classification. In our work, we will consider *Beatport*, one of the most famous EDM web sites for DJs and fans. Beatport classifies each song into a single subgenre from a broad taxonomy of over 20 subgenres. According to the interview with the Beatport VP of Marketing (Cole, 2016), Beatport reportedly has '25000 new releases most weeks'[2] and a committee review the taxonomy every six months 'to retire genres and bring in new ones' with the aim of 'staying ahead of the trends'.[3] As a result, Beatport provides an adequate testing ground for studying automatic subgenre classification in EDM and the impact of changes in the taxonomy and subgenre evolution over time.

More precisely, in this work, we have gathered two data sets of over 2000 EDM songs (100 song per subgenre in the Beatport taxonomy), each data set separated by more than a year and with some differences in the list of subgenres. We will study

[2]According to the filtering tool implemented in the web, 3,209,396 tracks were released from 1 January 2016 to 31 December 2018, which means more than 20,0000 tracks per week. See `https://www.beatport.com/tracks/all?start-date=2016-1-1&end-date=2018-12-31`

[3]Some changes in Beatport genres are reported by Williams (2017a,b).

the performance of machine learning automatic classification in both data sets, with and without retraining of the classifier. Our experiment will look into results to try to answer the following questions:

- How accurate are machine learning classifiers considering taxonomies of over 20 subgenres?
- Does the accuracy change in both data sets? Is retraining necessary after one year?
- Are some subgenres more challenging to discriminate or the classification error distributes evenly among subgenres?
- Is there any intense confusion among a pair (or a group) of subgenres? Is this a potential cause of redundancy or overlapping in the taxonomy?
- Are the answers to the two previous questions consistent in both data sets?
- Can we use the classifier to relate the new subgenres that appear in the second data set with the older ones?

The analysis of the results and the classification error will provide some answers to these questions and will raise some new questions about the very own nature of some subgenres and the evolution of the taxonomy.

The rest of the article proceeds as follows. Section 2 reviews the related work. Section 3 describes the process of collecting and transforming the data and the set up of the machine learning classifiers. Section 4 discusses the results of the experiments. Finally, Section 5 concludes with some remarks and future work.

2. Related work

As we have pointed in the previous section, automatic subgenre classification has been substantially less studied than genre classification. However, folk music classification has received significant attention, mostly due to its interest from the ethnomusicological point of view. As a result, there are works about folk genres from different parts of the world including Latin America (Silla, Koerich, & Kaestner, 2008), Europe (Boot, Volk, & de Haas, 2016; Chai & Vercoe, 2001; Hillewaere, Manderick, & Conklin, 2012), India (Sridharan, Moh, & Moh, 2018) or China (Liu, Yang, & Chen, 2008). The taxonomies in these works represent narrower settings of more closely related labels. However, they are not so broad and changing over time as it is the case of EDM taxonomies.

EDM has also received attention for automatic classification. This genre has some prominent features such as rhythmic patterns, tempo and repetition, as in the case of ballroom dance music (Dixon, Pampalk, & Widmer, 2004). Diakopoulos, Vallis, Hochenbaum, Murphy, & Kapur (2009) apply this kind of features to EDM for automatic subgenre classification. They consider six EDM subgenres that are selected by their diversity and popularity at that time and obtain a classification rate of 75.2%.

Leimeister, Gaertner, & Dittmar (2014) refine the extraction of rhythmic pattern features by using source separation techniques for discriminating bass drum and snare drum events. Their classifier works on eight subgenres, and the use of both tempo and rhythmic pattern features allows to increase the accuracy from 66% to 71% with respect to the results only considering tempo features. Other works have further refined the use of segmentation techniques in music similarity to analyze both rhythm and timbre similarity in EDM, where these features are prominent (Panteli, Bogaards, & Honingh, 2014; Panteli, Rocha, Bogaards, & Honingh, 2017). However, the techniques have not been evaluated in a subgenre classification setting.

Gomez Camara (2017) also proposes automatic subgenre classification of EDM. It uses Juno,[4] the UK-based online record shop, for collecting a data set of 1200 songs. This work considers only six subgenres from the Juno taxonomy, which has more than twenty subgenres and that for some subgenres considers a second level. Some problematic subgenres are excluded such as *hard-house* that has a tempo of 118-135 or 140 BPM, depending on the source. Similarly, the work reportedly do not include some some songs with intricate rhythm patterns. The classification accuracy in the experiment varies when considering tempo alone (60.5%), or combined with beat histograms (68.4%), and also with timbral features, namely Mel Frequency Cepstral Coefficients and pitch histograms (74.6%).

Along these lines, our work focuses on automatic classification in a real-life scenario. More precisely, we study an EDM taxonomy with a broad list of subgenres, probably some of them overlapping or ill-defined, and highly changing over time. This kind of taxonomies offer a challenge for automatic genre classification and also provides an stimulating setting to empirically assess subgenre definition.

3. Material and methods

In our experiment, for each EDM subgenre of Beatport, we retrieved the top 100 tracks of its ranking on November 29, 2016. Beatport included 23 subgenres at that time, so we collected 2300 songs as *Set 1*. We repeated the retrieval 14 months later, on February 4, 2018. Then Beatport considered 29 subgenres, so we retrieved 2900 songs as *Set 2*.

The taxonomy changed notably in that period since new subgenres were added, and some others removed. The changes seem to follow the Beatport policy of regularly updating the subgenre list to follow the trends in EDM. In particular, Funk R&B is only present in Set 1, while the following subgenres appear only in Set 2: Afro House, Funk Disco, Funk House, Garage, Left Bass, Left House Techno and Trap.

Interestingly, 100 songs appear both in Set 1 and 2, which means that some songs are still popular 14 months later. In 67 of them, there was no subgenre change. However, there was a subgenre change in 33 songs. Table 1 shows the subgenre changes. Most of them involve subgenres that either appeared in or were dropped removed from Set 2, that is, songs that were reclassified after the taxonomy changes. Most notably, 15 songs from the deprecated Funk R&B subgenre were reclassified as the new subgenre Funk Disco, which probably means that Funk Disco and Funk R&B are related. However, in two songs the subgenre changed without involving deprecated or new subgenres (the change was from Electro House to Big Room in both songs). Thus, in Beatport songs may change their subgenre. Since the number is small and the reasons for the changes unknown to us, we did not remove the songs from the sets.

For each song, Beatport makes available a song preview of 120 seconds (obtained from any part of the song), sampled at 22050 Hz in mono format. Each fragment is characterized by audio features extracted using pyAudioAnalysis (Giannakopoulos, 2015), but also by rhythmic features extracted with Essentia (Bogdanov et al., 2013).

For the automatic classification task, we considered four machine learning algorithms (namely, decision trees, random forests, extremely randomized trees and gradient tree boosting). Each algorithm was fine-tuned by hyperparameter optimization

[4] `www.juno.co.uk`

using genetic algorithms in a stratified k-fold cross-validation setting.

We analyze in detail the classification results, paying attention to the performance across subgenres, confusion between subgenres and the evolution of the classification performance in the two sets of songs.

More details on the experiment are given below.

3.1. Audio feature extraction

In order to predict the subgenre of the songs, we first have to extract features from the audio files. For that purpose, we follow a well-known approach in audio analysis, similar to that described in Peeters (2004). More precisely, we divide the audio signal of each song into non-overlapping time frames of 50 msec, compute some (instantaneous) audio features for each time frame, and, then, summarize the values of each feature along time using descriptive statistics, more precisely, the mean and the standard deviation.

For this task, we use pyAudioAnalysis, the open-Source Python library for audio signal analysis (Giannakopoulos, 2015), and, consequently, we extract the 34 audio features implemented in that library, and calculate the mean and standard deviation of each feature for a total of 68 variables describing each song. As a result, we include time-domain features extracted from the raw signal (e.g. Zero Crossing Rate, Energy and Entropy of Energy), frequency-domain features (such as those extracted from the spectrum or the chroma vector) and cepstral-domain features (more precisely, 13 Mel Frequency Cepstral Coefficients). According to Tzanetakis, Essl, & Cook (2001), these are timbral features that fall into the category of musical 'surface features' in contrast to rhythm features that focus on the most salient periodicities of the signal, in particular, tempo.

Regarding the estimation of tempo, pyAudioAnalysis generates two values: the BPM estimation and a confidence value that represents the overall dominance of the detected beat rate. The confidence is relevant as some songs combine different tempos. It is important to remark that these features are not estimated for each frame, but for the whole audio track. However, given the prominent role of tempo and rhythmic patterns in EDM and that pyAudioAnalysis do not provide more features to represent them, we decided to use Essentia (Bogdanov et al., 2013) to include more tempo and rhythmic features. In particular, we use its BPM estimation some additional rhythmic features available in the library, namely: first and second peaks properties, danceability, onset ratio and several measurements about beats loudness.

We decided to consider two BPM measurements (Essentia's and pyAudionalysis') due to the inherent complexity of tempo induction and the many different ways to estimate it (Gouyon et al., 2006). Tempo estimation in pyAudioAnalysis is not very sophisticated (Giannakopoulos, 2015), and the procedure that follows the Essentia Rhythm Extractor[5] seems to be more refined, combining several periodicity functions and beats estimators. The results will serve us to analyze whether the two BPM measurements are truly redundant or whether they provide different information that can be exploited by the classifiers.

As a result, the classifiers will use 92 input variables. Table 2 shows a summary of all the features considered and their corresponding input variables. The resulting data sets are available for download.[6]

[5] `https://essentia.upf.edu/documentation/reference/std_RhythmExtractor.html`

[6] Set 1 can be downloaded from `https://www.kaggle.com/caparrini/electronic-music-features-201611`

3.2. *Classification Methods*

In Beatport, each song belongs only to a subgenre. Hence, we will carry out a single-output multiclass classification task, where only one label, i.e. one subgenre, is correct.

For each data set we have performed classification with four different machine learning classifier algorithms.[7] Namely, decision trees (Breiman, Friedman, Olshen, & Stone, 1984), random forests (Breiman, 2001), extremely randomized trees (Geurts, Ernst, & Wehenkel, 2006) and gradient tree boosting (Friedman, 2001). These methods include the classic decision trees, but also three different ensemble classifiers, which are known to be more accurate than any of its members if the classifiers are accurate and diverse (Hansen & Solomon, 1990).

The approach followed is similar to the one that resulted best in the work by Murauer & Specht (2018) for the WWW 2018 genre classification challenge (Defferrard et al., 2018), where ensemble methods such as extreme gradient tree boosting and extremely randomized trees on numerical audio features extracted with Essentia provided better results than those from deep neural networks on the same features or than those from convolutional neural networks on spectrogram data.

All the considered algorithms are tree-based classifiers, which we consider relevant, because they incorporate 'feature selection' and in this way, they can cope with a high number of variables of presumably very different importance for the classification. In addition, they are white-box classifiers and provide ways to measure the relevance of the features used for the classification, which makes possible to study the classification process, in addition to the results themselves.

For each machine learning algorithm, we look for the optimal values of its parameters using genetic algorithms.[8] For each individual in the population, i.e. for each parameter combination, we use as fitness value the inverse of the mean classification error in the validation of a stratified k-fold cross-validation setting with $k = 10$. The parameters used on the genetic algorithm search are:

- Epochs: 30
- Initial population: 100
- Mate method: two-point crossover
- Selection: Select the best individual among two randomly chosen individuals, 100 times
- Polynomial mutation as implemented in original NSGA-II algorithm

The best individual is selected, i.e. the best parameter combination obtained in the genetic optimization.

Finally, we compare the best individuals of each of the four machine learning algorithms considered. For the comparison, we use stratified k-fold cross-validation with $k = 10$. The validation is stratified to preserve the same amount of songs of each subgenre in each fold. We also use the same data set partition (i.e. the same folds) for all the experiments to eliminate the impact of the random partition when comparing the performance of the classifiers.

-beatporttop100 and and Set 2 from `https://www.kaggle.com/caparrini/electronic-music-features-201802-beatporttop100`

[7] Algorithms implemented in Scikit-learn. `http://scikit-learn.org/`

[8] The Python library DEAP, `http://deap.readthedocs.io/en/master/`

3.3. Description of the analysis of the results

We analyze the results in two different ways:

- Analysis of automatic classification, which includes the following aspects:
  - Classification performance in general and for each subgenre.
  - Confusion matrices and graphs between subgenres.
  - Importance of the audio features for classification.
  - Repeatability of the classification results for the two data sets.
- Analysis of the performance of the Set 1 classifier in Set 2:
  - Evolution of the performance, i.e., is the classifier trained with Set 1 still useful for classifying songs retrieved after more than a year? Are its results notably worst than those obtained from the classification algorithm trained with the songs in Set 2?
  - Studying the relation of the new subgenres, i.e., those present in Set 2 but not in Set 1, with the older ones. We use the classifier trained with the Set 1 to classify the songs of the new subgenres. If the songs from a new subgenre are mainly classified in a few subgenres from the old taxonomy, then those subgenres are somehow related to the new ones.

The objective interpretation of the results guides our analysis. However, on some occasions that are marked appropriately, we use our knowledge in EDM to interpret them in the light of human perception. This interpretation is inevitably subjective and hence debatable, but we believe it enhances the analysis.

The code of the experiment is available for download.[9]

4. Results

4.1. Analysis of automatic classification

4.1.1. Classification performance

Table 3 shows the accuracy of the four machine learning algorithms considered in the k-fold cross-validation setting. Not surprisingly, decision trees are the worst classifier in both sets, due to their simplicity. However, the rest of the methods obtain similar results. Their mean values are not significantly different from a statistical point of view at $P = 0.05$ level according to an ANOVA test adjusted with Bonferroni correction.[10]

According to the mean accuracy values, the gradient tree boosting classifier was the best in Set 1 (accuracy of 59%), while the extremely randomized trees algorithm performed best in Set 2 (accuracy of 48.2%). Accuracy notably decreases in Set 2 (almost 10 points for most classifiers). The main reason behind this result could be the fact that the taxonomy in Set 2 has more subgenres than in Set 1. However, other factors, such as subgenre redefinition could take place as well, as we will see later.

Our best accuracy values are still far from those that are obtained with state-of-art methods in typical genre classification data sets, see for example the work by Nanni, Costa, Aguiar, Silla Jr., & Brahnam (2018) who report over 90% in three popular data sets (GTZAN, ISMIR 2004, and Latin Music Database). However, our results are closer to those reported by EDM subgenre classification studies with a limited set

[9]The code can be downloaded from `https://github.com/Caparrini/pyGenreClf`

[10]The same conclusions are drawn using a McNemar test (at $P = 0.05$ level and adjusted with Bonferroni correction) to compare the performance of the classifiers pairwise.

of subgenres, which report accuracy values around 75% in reduced taxonomies of six and eight subgenres. With these precedents, we consider our accuracy values are fair, taking into account the standard features extracted, the high number of subgenres (over 20 in both sets) and the subgenre proximity (in principle, subgenres should be stylistically closer among themselves than genres and, consequently, more prone to confusion).

We will only report the results of the best classifier for each set. However, we compared the agreement of the best classifiers (random forest, extremely randomized trees and gradient tree boosting) in each set using the Kappa coefficient.[11] Results in Table 4 show high agreement between all methods, considering the high number of subgenres. This means that the classification induced by these methods tend to agree. Thus, the in-depth analysis of the results of the best classifier that will be reported below is not expected to be greatly different to that from other classifier.

Furthermore, we will show the aggregated results of the $k = 10$ validation subsamples. Hence they do not belong to a specific classifier but to the aggregation of $k = 10$ classifiers. The aim is to generalize the conclusions that can be obtained.

In Table 5, we show classification performance measures for each subgenre in Set 1 and Set 2. More precisely, for each subgenre, we show its precision (true positives divided by the number of true positives and false positives), recall (the number of true positives divided by the number of true positives and false negatives) and F1 score (the harmonic mean of precision and recall). These measures are related to the classifier exactness, completeness and a balanced combination of both features, respectively.

The precision values widely vary. In Set 1 precision ranges from 0.89 of Hardcore Hard Techno to 0.39 of Techno and Deep House, while in Set 2 it ranges from 0.78 of Hard Dance and Psy Trance to 0.21 of House. The lowest precision value in Set 2 is very low (21%), but is still a significant improvement with respect to that from a random choice classifier in Set 2, i.e, one that randomly assigns the subgenre to a new song, which would have a expected precision of 3.4% (1 out of 29) for all the genres.

Similar variation can be observed in the recall values as well. In general, low performance values could indicate ill-defined subgenres, or subgenres defined by extra-musical features, or by musical features that cannot be identified by our approach. From a pragmatic perspective, automatic classification in this experiment still seems of great help at least for suggesting classification in most, if not all, subgenres.

In both sets, the five subgenres with better performance according to the F1 score are the same: Drum & Bass, Hard Dance, Hardcore Hard Techno, Psy Trance and Trance. Thus, they seem very apt subgenres for automatic classification using the approach followed in this work.

On the other hand, the worst F1 values are for Techno (0.38) in Set 1, while in Set 2 is House (0.18). The explanation could be that these subgenres are related to others with similar features. Interestingly, in the case of House, its performance is notably better in Set 1 (0.40), but still is one of the worst performances in that set. This means that the House subgenre is one of the worst characterized in both cases. However, Techno, being the worst subgenre in Set 1 according to the F1 score, improves its performance in Set 2 (0.54) and is the sixth-best subgenre in terms of F1 score. It could mean that the Techno label is more consistent in Set 2 than it was in Set 1. We can find the opposite case in Glitch Hop where the F1 score decreases from 0.72 in Set 1 to 0.51 in Set 2.

These cases could hint at stylistic changes in the subgenres in the dates considered,

[11] The Kappa coefficient measures inter-rate agreement and ranges from 0 to 1

and even different interactions with other subgenres in the taxonomy. In any case, the classification performance of many other subgenres changes in both sets. This means that the taxonomy changes and the stylistic trends affect in different ways to the subgenres considered.

Regarding the new subgenres, the performance of the classifier is poor if we consider that new subgenres supposedly are supposedly created to better refer to new music styles. Our classifier is not able to effectively discriminate them (it obtains F1 scores of less than 0.5). Again, the reasons may be varied, including limitations of our approach, imprecise subgenre definition or extra-musical aspects.

4.1.2. Subgenre confusion matrix

Figures 1 and 2 show the confusion matrices for Set 1 and Set 2, respectively. In Figure 1 we can see strong confusions among some subgenres, but in an assymetrical way. More precisely:

- More Electro House songs are classified as Big Room than vice versa,
- More Hip Hop songs are classified as Dubstep than vice versa, and
- More Progressive House songs are classified as Deep House than vice versa, but the confusion is more balanced in this case.

We can also see some subgenres with very low values in the main diagonal, which means that many of their songs are misclassified. This is the case of Dance, Techno, House and Deep House, which are confused with many genres. These subgenres probably are too-broad from a stylistic point of view and, hence, the confusion.

In Figure 2, we can see lighter shades of blue in the main diagonal, and many non-white cells in the rest of the matrix. This means that the confusion among subgenres increases in Set 2 with respect to that in Set 1. However, a careful look at both figures reveal interesting phenomena, for example:

- House in Set 2 seems to be hurt by the appearance of the subgenre Funk House not present in Set 1. In both sets Tech House is often confused with House, which makes sense. However, the impact of Minimal and Future House changes across both sets. So confusion among subgenres is not consistent in both taxonomies.
- The case of Techno is the opposite, as more Techno songs are correctly classified in Set 2. A careful look at the confusion matrices reveals that in Set 1 Techno songs were misclassified as Progressive House, Tech House and Minimal in 31 occasions and that this error now takes place only nine times. Thus, Techno seems to be more consistently defined in Set 2 .
- Left Bass is a new subgenre that is confused with many other subgenres. This means that our classifier cannot distinguish it clearly, perhaps because the subgenre label is loosely used or because our approach can not grasp its distinctive features.
- Funk Disco and Funk House are new subgenres that share the ‘funk’ vein but do not confuse much. In the case of Funk Disco the sources of confusion are Dance and Indie Dance/ Nu disco, while for Funk House are Future House and House.

4.1.3. Subgenre confusion graph

Since the resulting matrix is very difficult to analyze for the human eye given the high number of classes (subgenres) and the subtleties of the confusion, we propose to use a directed graph to present the confusion among subgenres. We will build such a graph following these ideas:

- A node i is added for each subgenre. Its size will be proportional to the true positives.
- An edge between node i and node j is added if there are songs in the subgenre i misclassified in the subgenre j (this allows edges in both directions between nodes). We represent additional information within edges:
  - Line thickness is proportional to the total number of songs misclassified, that is, songs from subgenre i classified as subgenre j and vice versa.
  - Arrow head size. The arrow head from node i to node j is only present if there are songs of subgenre i misclassified as of subgenre j, being the size of the arrow head size proportional to the number of songs misclassified.
  - To clarify the visualisation of the graph, edges with weight less than or equal to 5 have been filtered out.

We use the tool *Force Atlas 2*, which is a force directed layout: it simulates a physical system in order to arrange a network (Jacomy, Venturini, Heymann, & Bastian, 2014). Nodes repulse each other like charged particles, while edges attract the nodes they connect like springs. It aims to represent structural proximity, facilitating the analysis, and in our particular case, the analysis of classification confusion between subgenres.

We will only analyze the confusion graph of Set 2 shown in Figure 3, for the sake of brevity. This set has more subgenres and is more recent. The most remarkable aspect in this graph is that subgenre location and interconnection can also be explained with the BPM. To support the explanation below, the distribution of the BPMs is shown in Figure 4.

- House is a small node, which may indicate that it is an ill-defined subgenre (of course, this is only a hint that should be verified by a musicologist). Around House, there are subgenres with a similar BPM around 125 (Minimal, Progressive House, Funk House, Afro House,...).
- Funk Disco and Indie Dance / Nu disco have slower BPMs (115-122).
- In the bottom and right sides of the graph are located subgenres with greater BPMs. Techno songs have BPMs around 125 with tracks up to 130, Hardcore Hard Techno around 130, Trance between 130-135 and Psy Trance above 135. This progression can be appreciated visually on the graph.
- Hard Dance has the greatest BPM with 150.
- Dubstep and Trap have BPMs between 120-160, but the beat is halftime.
- Reggae Dub, Glitch Hop and Hip Hop shares BPMs between 100-140, sometimes halftime.
- Drum And Bass have a characteristic BPM of 87, but it can be higher, and some tracks have BPMs that are shared with Left Bass.
- Next to Future House, we have Electro House and then Big Room. Big Room with a BPM around 130 and Electro House between 125-130, makes a progression from Future House with BPM more like the 'House' group of 125. Here we notice that BPM is not the only important feature, because these subgenres are clearly separated from others with similar BPMs.

It is important to remark that from our point of view, many of the errors are somehow understandable since they involve closely related subgenres, for example, subgenres that share a common origin, or subgenres that are a specialization of others or that combine stylistic features of others.

The subgenres Electro House, Big Room and Future House have an interesting relation. In September 2016, Beatport added three new labels to their system: Dance, Big Room and Future House. This was an attempt to classify the EDM that was proliferating at that time. Citing an electronic music aficionado *'Big Room is essentially everything carried over from Electro House plus all the big anthem house. [...] Future House is essentially big room handbag, with tracks moved over from House, Deep House, Dance, and Progressive House'*.[12] The confusion between Electro House and Big Room makes sense given that Big Room was originally Electro House. The same way, the relation between these two genres with Future House is explained.

More opinions from the same aficionado: *'Dance is a rather plain category. It is the genre used by generic music services like iTunes and Google Play to cover everything electronic'*. This vagueness is probably what we are viewing on the confusion graph. The Dance node is small because of its low true positives rate. It is confused with several genres, and the most relevant confusion is with Future House which is, as mentioned before, a very popular electronic music (it is played in many music radio stations). So Dance seems to be a catch-all subgenre that is strongly related to the most popular EDM songs and subgenres.

Deep House, Minimal and Progressive House are also confused. Listening to these genres, all of them have the same BPM and blurred limits. From our point of view, Minimal is less 'dark' and have fewer sound richness, while Deep is more percussive and instrumental, and Progressive is more melodic than the others. However, these are subtle differences even for electronic music aficionados.

As a result, we can read the confusion graph of the automatic classifier as a map of subgenres of electronic music from the perspective of a human expert. For example, we can see the House family on the left (marked BPMs), below Techno and faster genres (Hard Techno, Trance, Psy Trance), the other way through Future House the modern faster genres Electro House and Big Room. On the top-right of the graph are the genres with more descriptive rhythm features and several BPMs, slower like the 'Hip Hop' group, faster such as Dubstep and Trap, or pretty fast, such as Drum And Bass and Left Bass.

The resulting groups or families in the map make sense to human experts. Thus, errors from automatic classification seem to be related mainly to the stylistic similarity of subgenres.

4.1.4. Feature importance

In tree-based classifiers, it is possible to measure the importance of each feature in classification. Figure 5 shows the top 10 features for the best classifiers in Set 1 and Set 2. In both cases, we can see that the *Essentia* BPM is clearly the most relevant feature, and that the second one is the first peak of the BPM histogram estimated by *Essentia*.

The next two features in both sets are the BPM estimated by *pyAudioAnalysis* and a confidence measure of the BPM estimated by *pyAudioAnalysis*, but not in the same order. Moreover, their difference with respect to the rest of the features is not so pronounced.

The prominent role of the BPM features proves the importance of rhythm to classify subgenres in EDM. It is important to remark that the BPM measures from *Essentia* and from *pyAudioAnalysis* are not redundant, but complementary. In some experiments not reported here, we have detected that classification performance decreases if

[12] `https://radleymarx.com/dj-biz/beatport-updates-their-genre-categories-big-room-future-house-and-more/`

we remove one of them, and that the ranking of the importance of the BPM features is always the same, being the *Essentia* BPM more important than the *pyAudioAnalysis* one.

The non-redundancy of BPM variables somehow reinforces the idea that the combination of the approaches is better than any of them, as appreciated by Gouyon et al. (2006) when evaluating the performance of BPM estimation algorithms. In our case, the classifier exploits the two BPM estimates, the information provided of the BPM confidence variable, which is related to the tempo dominance, and that from the first peak of the histogram. This fact opens up new research pathways, such as including more BPM estimates and other tempo features.

Regarding the rest of the top features, in both rankings, we can see time-domain, spectral-domain and cepstral-domain features. In this sense, we can say that these families of features also play an important role in the classification, even if less prominent than the tempo-related features.

4.1.5. Conclusions about the repeatability of the results in the two sets

Since in both sets of songs we have used the same methods, we can analyze the repeatability of the results over time,

The main conclusion is that in both sets of songs, it was possible to build a competent automatic classifier, and that the classifiers relied heavily on the tempo features.

Another important conclusion is that the analysis of the results confirmed the dynamic nature of EDM subgenres. Firstly, the classifier performance notably changed across sets. Secondly, while the classifier performance in some subgenres was similarly good or bad in both sets, in some others experienced considerable changes. What factors do explain these results? On the one hand, the human factor behind subgenre definition and labelling. On the other hand, the experiment design because the top 100 songs of each subgenre probably biased the data set towards popular or fashionable songs at each moment. Another potential cause is the inability of the approach followed to apprehend some musical nuances that differentiate subgenres.

In any case, the results make clear that the dynamic nature of EDM poses a challenge for automatic classifiers, which may quickly became obsolete. We will delve into this aspect below.

4.2. Analysis of the performance of the Set 1 classifier in Set 2

In this section, we want to analyze how the performance of an electronic music classifier evolves along time, that is, we want to determine whether the classifier trained some time ago can still produce a reliable classification.

In order to asses that, we will retrain the best classifier of Set 1, i.e. the Gradient Tree Boosting classifier, using the whole data of Set 1. Retraining is needed because the results reported are the aggregation of the $k = 10$ validation sub-samples, i.e. using ten classifiers. As already mentioned, the subgenre Funk R&B is not present in Set 2, but we did not remove these songs when retraining the Set 1 classifier since we wanted to test precisely the performance of the 'old' classifier.

The classifier trained with Set 1 will be used to classify the songs of Set 2 with some exceptions:

- We removed the songs from the new seven subgenres of Set 2.
- We removed the 70 songs that appear in both data sets and do not belong to

the new subgenres. More precisely, we removed the 67 songs appear in both data sets with the same subgenre and the three songs that changed to a subgenre already existing in Set 1 (see Section 3 for the details of the repeated songs).

As a result, our trimmed Set 2 has 2130 songs belonging to 22 subgenres. The subgenre that loses most songs is Dance with ten songs discarded, while six subgenres do not lose any song.

The accuracy of the Set 1 classifier in Set 2 is 0.403, thus, the performance has remarkably decreased from 0.548 that is the aggregated accuracy of Set 1 classifier.[13] If we compare it with the aggregated accuracy of the Set 2 classifier, which is 0.501 (considering only the common subgenres in Set 1 and Set 2), we still appreciate a notable deterioration. Hence, we can conclude that retrieving a new set of songs after some months and training a new classifier is highly recommended because changes in the taxonomies and music trends hurt the classifier performance.

In Table 6, we show the performance for each subgenre. More precisely, we show the precision, the recall, the F1 score and the difference between the F1 score obtained by the Set 1 and the Set 2 classifiers in Set 2. The F1 scores are not truly comparable, since the one from Set 1 (see Table 5) is the aggregation of the classifiers of the $k = 10$ validation folds. However, we added the difference to facilitate an approximate comparison.

The negative differences in F1 for most subgenres suggest that Set 1 classifier is less effective than the one freshly trained. However, five values are positive meaning that for five subgenres the old classifier was able to grasp the subgenre features better; the most remarkable subgenre in this aspect is Dubstep with an increase of 0.1 in the F1 when using the old classifier, that is, the one trained with Set 1. This may be explained by the songs present in Set 2 that are more similar to those in Set 1, probably because in the Set 2 classifier the inclusion of Trap songs misled the classifier.

On the other hand, the subgenres where the Set 1 classifier deteriorates more are Trance (-0.59) and Techno (-0.33). The case of Trance is very curious because the performance of both classifiers in Table 5 is good, and the confusion matrices do not change substantially. Thus, we would expect the Set 1 classifier to perform well in Trance songs of Set 2, but it does not happen. Hence, most likely, the subgenre has gone through stylistic changes during that period, but is very consistent in both sets. The case of Techno is different because the performance of the Set 1 classifier was really poor, as noted in Table 5, and it improves when is used to classify the Set 2 songs. This evolution points to a different and more consistent use of the Techno label in Set 2 than in Set 1.

Some subgenres such as Drum & Bass and Psy Trance show good performance (F1 score over 0.7) and little change (0.04 and -0.08 respectively) meaning that these subgenres did not change much in the period considered. However, the fluctuations that the rest of the table shows reveal that taxonomy changes affect in very different ways to the subgenre performance. Hence, this result reinforces the idea of the need to retrain a classifier after some time. More precisely, when some subgenres have probably gone through stylistic changes.

[13]Accuracy values excluding the Funk R&B subgenre not present in Set 2 and the songs that appear in both Set 1 and 2

4.2.1. Relating the new subgenres to the older ones with Set 1 classifier

In this section, we use the Set 1 classifier to try to establish relations between new subgenres and the older ones. These relations could provide hints about the genealogy of the new subgenres. However, this is a speculative and in some way, naive study. The musical influence of a genre on a later one is a non-trivial issue (Collins, 2012) and requires a more sophisticated study informed, among other things, by the musicological features of the genres.

When we classify the songs from the new subgenres in Set 2 with the classifier trained with Set 1, we aim to find whether there are 'old' subgenres that are strongly confused with the new ones. If we find that most songs from a new subgenre are labelled as from an old one, then these subgenres probably share stylistic aspects. Figure 6 shows the results.

For the case of Afro House we find that 32 songs are labelled as Minimal, 31 as Indie Dance / Nu disco and 14 as Deep House. Interestingly, only the confusion with the Deep House appears in the confusion matrix in Figure 2. So perhaps the relation of Afro House with Minimal and Indie Dance / Nu disco is due to some stylistic similarities with these subgenres at the time Set 1 was retrieved.

In the case of Funk Disco, 33 songs are classified as Funk R&B, which is a subgenre that disappeared in Set 2. As noted in Section 3 some of the Funk R&B songs were re-labelled as Funk Disco, and as a result, this new subgenre shares some of its features. Again there is a strong confusion with Indie Dance / Nu disco, which means that the latter probably is a catch-all subgenre or one that groups different styles.

Interestingly, it does not happen the same in the case of Funk House, because only two songs are labelled as the deprecated Funk R&B. In this case, they are mostly classified as Tech House (41) and House (32), meaning that Funk House most likely can be considered as a subcategory of these subgenres. Interestingly, Funk House is barely confused with these two subgenres in Set 2, which means that the use of the label Funk House in Set 2 is coherent.

The Garage subgenre shows a strong confusion with Breaks (26 hits), but also important ones (of 10 or more hits) with other four subgenres. The confusion with the Breaks subgenre is the only one that remains in the confusion matrix of Set 2, so perhaps the new subgenre cleared the rest of them.

In the case of Left Bass, it is confused with 20 subgenres, and the one with more hits is Electronica Downtempo with 21, while there are other five subgenres with around 10 songs classified. This means that the stylistic relations of Left Bass are not clear. Moreover, the Set 2 classifier has a low recall in this subgenre (0.29) and the confusion matrix in Set 2 shows that there are no strong confusion with any other subgenre. This either means that the songs from this subgenre do not share strong stylistic features or that our classifier is not able to grasp them.

Left House Techno has a strong relationship with Minimal with 34 hits and to a lesser extent with Electronica Downtempo with 18 hits, but no strong confusion appear in the confusion matrix in Set 2. However, the F1 score is 0.44, which means that the classifier has not grasped the subgenre features very effectively.

Interestingly, most of the songs of the Trap subgenre are classified as Hip Hop (31) and Dubstep (24). From our point, this makes perfect sense, since Trap music is characterized by the slow tempos of Dubstep and the drum sounds of Hip Hop. While the confusion with Hip Hop disappears in the confusion matrix in Figure 2, the confusion with Dubstep still exists, which means that both subgenres share essential features and that in some cases, the classifier is not able to discriminate them.

5. Discussion

We have addressed the problem of automatic subgenre classification in EDM in a realistic setting, i.e., considering a widely used taxonomy and popular songs at different times. With this aim, we take as reference Beatport, a famous website of EDM songs. This source provides a big collection of samples of songs with their corresponding classification according to a given taxonomy, which is very suitable for our goals. However, our methodology could be ported to any other EDM database with minor changes.

We follow a standard approach and obtain satisfactory results given the complexity of the task, that is, classifying a song in a broad taxonomy of subgenres. As pointed out in Section 2, some works obtain better results but considering reduced taxonomies: Diakopoulos et al. (2009) consider six subgenres and obtain a classification rate of 75.2%; Leimeister et al. (2014) obtain 71% with eight subgenres; Gomez Camara (2017) achieves 74.6% considering only six selected subgenres from the website Juno. Our classifier obtains an accuracy of 59%, but it operates on the complete set of 23 subgenres of Beatport in Set 1 (and 48.2% in Set 2 with 29 subgenres). In its actual state, our approach provides enough accuracy to be used by artists or producers as an *assistant tool* to suggest the subgenre of their songs before publishing them.

Interestingly, the classifier performance varies across subgenres. In some of them, it is poor, which could indicate a flawed subgenre definition. The study of the musicological properties of the subgenres would help to identify those really ill-defined and would also clarify whether the subgenre confusions are due to an imprecise definition or to limitations of our approach. That study is out of the scope of the present work, but our findings could provide relevant hints for such research.

Some of our findings about the relevance of the features for classification coincide with some previous works in EDM and were predictable: rhythm patterns, and in particular, BPM are the most discriminating properties. We even used two BPM estimations provided by different feature extraction tools such as pyAudioAnalisis or Essentia. Far from being redundant, the experiments show that the classifiers can benefit from both features and improve their accuracy. This result suggests the interest of studying the performance of different BPM measures in EDM subgenres and to further study the relationship among these measures. In any case, it seems promising to investigate the inclusion of even more features related to rhythm and tempo to improve the classification results.

Moreover, melody, harmony and other structural properties of songs could be also exploited to obtain additional features, such as tonality and keys of a song (Faraldo, Gómez, Jordà, & Herrera, 2016), structural and temporal development (Knees et al., 2015), or even patterns of "drops" and climax sections of the songs (Yadati, Larson, Liem, & Hanjalic, 2014). The analysis across subgenres of structural song segmentation and the subsequent analysis of timbral features of the resulting segments may also provide interesting insights (Rocha, Bogaards, & Honingh, 2013).

In an attempt to explain subgenre confusion for the classifier, we have represented the confusion matrix as a graph and interpret it as a taxonomic map of EDM. Confusion between subgenres translates into nearness in the map, and subgenres with good performance metrics translate into big nodes. The resulting graph can help to identify inconsistencies or problems within the taxonomy, changes when comparing two graphs from different periods, or the position of new emerging subgenres. This kind of taxonomic maps could be further explored by musicologists when analyzing subgenre relations in EDM or other music styles.

Finally, our work confirms the volatile nature of EDM subgenres over time. While automatic classification can help to cope with the annotation of the new songs produced each week, classifiers need to be updated regularly to avoid performance deterioration and to catch up with changes in the taxonomy and the music trends. Our work also brings to light the need for more classification studies on EDM informed by musicological perspective to delve into the topics explored in this work.

6. Acknowledgments

We would like to thank the anonymous reviewers for helping us improve this paper with their suggestions. This work was partially supported by the Spanish project TIN2017-86217-R (CAVI-ART-2) and Madrid regional project S2018/TCS-4339 (BLOQUES-CM) co-funded by EIE Funds of the EU.

7. Declaration of interest statement

The authors declare no conflict of interest.

References

Aucouturier, J.-J., & Pachet, F. (2003). Representing musical genre: A state of the art. *Journal of New Music Research*, *32*(1), 83-93.

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., … Serra, X. (2013). Essentia: An audio analysis library for music information retrieval. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013)* (pp. 493–498). Curitiba, Brazil: International Society for Music Information Retrieval.

Boot, P., Volk, A., & de Haas, W. B. (2016). Evaluating the role of repeated patterns in folk song classification and compression. *Journal of New Music Research*, *45*(3), 223-238.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* Monterey, CA: Wadsworth Brooks/Cole Advanced Books & Software.

Chai, W., & Vercoe, B. (2001). Folk music classification using Hidden Markov Models. In *Proceedings of International Conference on Artificial Intelligence: IC-AI' 2001* (p. 216-221). Las Vegas, Nevada, USA: CSREA Press.

Cole, D. (2016). *Beatport is taking a new approach to genres.* `https://djtechtools.com/2016/08/12/beatports-re-approach-to-genre-tagging/`. (Online; accessed 26-February-2019)

Collins, N. (2012). Influence in early electronic dance music: An audio content analysis investigation. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012)* (p. 1-6). Porto, Portugal: International Society for Music Information Retrieval.

Defferrard, M., Benzi, K., Vandergheynst, P., & Bresson, X. (2017). FMA: A dataset for music analysis. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017)* (p. 316-323). Suzhou, China: International Society for Music Information Retrieval.

Defferrard, M., Mohanty, S. P., Carroll, S. F., & Salathé, M. (2018). Learning to recognize musical genre from audio: Challenge overview. In *Companion Proceedings of the The Web Conference 2018* (pp. 1921–1922). Lyon, France: International World Wide Web Conferences Steering Committee.

Diakopoulos, D., Vallis, O., Hochenbaum, J., Murphy, J. W., & Kapur, A. (2009). 21st Century Electronica: MIR Techniques for Classification and Performance. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)* (p. 465-469). Kobe, Japan: International Society for Music Information Retrieval.

Dixon, S., Pampalk, E., & Widmer, G. (2004). Evaluating rhythmic descriptors for musical genre classification. In *Audio Engineering Society Conference: 25th International Conference: Metadata for Audio.* London, UK: Audio Engineering Society.

Faraldo, Á., Gómez, E., Jordà, S., & Herrera, P. (2016). Key estimation in electronic dance music. In N. Ferro et al. (Eds.), *ECIR 2016: Advances in information retrieval* (pp. 335–347). Padua, Italy: Springer.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*(63), 3–42.

Giannakopoulos, T. (2015). pyAudioAnalysis: An open-source Python library for audio signal analysis. *PloS One*, *10*(12).

Gomez Camara, L. (2017). *Feature analysis and subgenre classification of electronic dance music* (Master's thesis, Queen Mary University of London, United Kingdom). Retrieved from `https://doi.org/10.13140/RG.2.2.28129.40802`

Gouyon, F., Klapuri, A., Dixon, S., Alonso, M., Tzanetakis, G., Uhle, C., & Cano, P. (2006). An experimental comparison of audio tempo induction algorithms. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(5), 1832-1844.

Hansen, L. K., & Solomon, P. (1990). Neural network ensembles. *IEEE Transactions of Pattern Analysis and Machine Intelligence*(12), 903–1002.

Hillewaere, R., Manderick, B., & Conklin, D. (2012). String methods for folk tune genre classification. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012)* (p. 217-222). Porto, Portugal: International Society for Music Information Retrieval.

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One*, *9*(6).

Knees, P., Faraldo, Á., Herrera, P., Vogl, R., Böck, S., Hörschläger, F., & Goff, M. L. (2015). Two data sets for tempo estimation and key detection in electronic dance music annotated from user corrections. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)* (p. 364-370). Málaga, Spain: International Society for Music Information Retrieval.

Leimeister, M., Gaertner, D., & Dittmar, C. (2014). Rhythmic classification of electronic dance music. In *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio.* London, UK: Audio Engineering Society.

Li, T., Ogihara, M., & Li, Q. (2003). A comparative study on content-based music genre classification. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval* (pp. 282–289). ACM.

Liu, X., Yang, D., & Chen, X. (2008). New approach to classification of Chinese folk music based on extension of HMM. In *2008 International Conference on Audio, Language and Image Processing* (p. 1172-1179). Shanghai, China: IEEE.

Murauer, B., & Specht, G. (2018). Detecting music genre using extreme gradient boosting. In *Companion Proceedings of the The Web Conference 2018* (pp. 1923–1927). Lyon, France: International World Wide Web Conferences Steering Committee.

Nanni, L., Costa, Y. M. G., Aguiar, R. L., Silla Jr., C. N., & Brahnam, S. (2018). Ensemble of deep learning, visual and acoustic features for music genre classification. *Journal of New Music Research*, *47*(4), 383-397. doi:

Panteli, M., Bogaards, N., & Honingh, A. K. (2014). Modeling rhythm similarity for electronic dance music. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)* (p. 537-542). Taipei, Taiwan: International Society for

Music Information Retrieval.
Panteli, M., Rocha, B., Bogaards, N., & Honingh, A. (2017). A model for rhythm and timbre similarity in electronic dance music. *Musicae Scientiae*, *21*(3), 338-361.
Peeters, G. (2004). *A large set of audio features for sound description (similarity and classification) in the CUIDADO project* (Technical Report). IRCAM. Retrieved from http://recherche.ircam.fr/anasyn/peeters/ARTICLES/Peeters_2003_cuidadoaudiofeatures.pdf
Rocha, B., Bogaards, N., & Honingh, A. (2013). Segmentation and timbre similarity in electronic dance music. In *Proceedings of the Sound and Music Computing Conference 2013* (p. 754-761).
Scaringella, N., Zoia, G., & Mlynek, D. (2006). Automatic genre classification of music content: a survey. *IEEE Signal Processing Magazine*, *23*(2), 133–141.
Silla, C. N., Koerich, A. L., & Kaestner, C. A. A. (2008). The Latin Music Database. In *Proceedings of the 9th International Society for Music Information Retrieval Conference (ISMIR 2008)* (p. 451-456). Philadelphia, Pennsylvania, USA: International Society for Music Information Retrieval.
Sridharan, A., Moh, M., & Moh, T. (2018). Similarity estimation for classical indian music. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (p. 814-819). Orlando, Florida, USA: IEEE.
Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, *10*(5), 293-302.
Tzanetakis, G., Essl, G., & Cook, P. (2001). Automatic musical genre classification of audio signals. In *Proceedings of the 2nd Annual International Symposium on Music Information Retrieval (ISMIR 2001)* (pp. 205–210). Bloomington, Indiana, USA: International Society for Music Information Retrieval.
Vitos, B. (2014). *Experiencing electronic dance floors: a comparative research of techno and psytrance in melbourne* (Doctoral dissertation, Monash University, Melbourne). Retrieved from https://doi.org/10.4225/03/58ae26c48115e
Williams, H. (2017a). *Beatport expands platform by adding the afro house category.* https://mixmag.net/read/beatport-expands-platform-by-adding-the-afro-house-category-news. (Online; accessed 26-February-2019)
Williams, H. (2017b). *Beatport updates platform by adding three new genre categories.* https://mixmag.net/read/beatport-updates-platform-by-adding-three-new-genre-categories-news. (Online; accessed 26-February-2019)
Yadati, K., Larson, M., Liem, C. C. S., & Hanjalic, A. (2014). Detecting drops in electronic dance music: Content based approaches to a socially significant music event. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)* (p. 143-148). Taipe, Taiwan: International Society for Music Information Retrieval.

Set 1 subgenre	Set 2 subgenre	Number of songs
Dance	**Funk House**	2
Deep House	**Afro House**	1
Dubstep	**Garage**	1
Electro House	Big Room	2
Funk R&B	Hip Hop	1
Funk R&B	**Funk Disco**	15
Hip Hop	**Trap**	5
House	**Funk House**	2
House	**Afro House**	1
Indie Dance / Nu Disco	**Funk Disco**	3

Table 1. Number of songs that changed the subgenre in Set 1 and 2 (subgenres that are present in only one set in bold letters)

pyAudioAnalisis	Variable index		
Audio Feature	Mean	Std. Dev	Global
Zero Crossing Rate	1	35	
Energy	2	36	
Entropy of Energy	3	37	
Spectral Centroid	4	38	
Spectral Spread	5	39	
Spectral Entropy	6	40	
Spectral Flux	7	41	
Spectral Rolloff	8	42	
MFCCs	9-21	43-55	
Chroma Vector	22-33	56-67	
Chroma Deviation	34	68	
BPM			69
BPM confidence			70

Essentia	Variable index		
Audio Feature	Mean	Std. Dev	Global
BPM			71
First/Second Peak BPM			72,74
Second Peak Spread			75
First/Second Peak Weight			73,76
Danceability			77
Beats Loudness	78	79	
Onset Rate			80
Beats Loudness Band Ratio	81-86	87-92	

Table 2. Audio features extracted for each song with the index used to identify the input variables.

Algorithm	Set 1	Set 2
Decision Tree	$0.422^{a}\pm 0.027$	$0.314^{c}\pm 0.018$
Random Forest	$0.572^{b}\pm 0.038$	$0.473^{d}\pm 0.019$
Extremely Randomized Trees	$0.584^{b}\pm 0.037$	$0.482^{d}\pm 0.024$
Gradient Tree Boosting	$0.590^{b}\pm 0.026$	$0.464^{d}\pm 0.013$

Table 3. Mean and standard deviation of the accuracy in the $k = 10$ validation subsets. Means with the same letter are not significantly different at the $P = 0.05$ level (One-way ANOVA with pairwise comparisons adjusted with Bonferroni.)

Comparison	Set 1	Set 2
RF-ERT	0.8438	0.7928
RF-GB	0.738	0.7096
ERT-GB	0.7453	0.6882

Table 4. Kappa coefficient for the best classifiers.

	Set 1			Set 2		
Subgenre	Precision	Recall	F1 score	Precision	Recall	F1 score
Afro House				0.45	0.46	0.45
Big Room	0.51	0.61	0.56	0.50	0.60	0.54
Breaks	0.75	0.72	0.73	0.53	0.56	0.54
Dance	0.45	0.38	0.41	0.32	0.22	0.26
Deep House	0.39	0.40	0.39	0.26	0.26	0.26
Drum And Bass	0.84	0.88	0.86	0.72	0.79	0.76
Dubstep	0.64	0.72	0.68	0.50	0.48	0.49
Electro House	0.46	0.43	0.45	0.39	0.47	0.42
Electronica Downtempo	0.41	0.46	0.43	0.54	0.44	0.48
Funk R&B	0.68	0.63	0.66			
Funk Disco				0.38	0.51	0.43
Funk House				0.42	0.55	0.48
Future House	0.51	0.45	0.48	0.29	0.46	0.36
Garage				0.49	0.46	0.48
Glitch Hop	0.70	0.74	0.72	0.47	0.55	0.51
Hard Dance	0.79	0.85	0.82	0.78	0.82	0.80
Hardcore Hard Techno	0.89	0.82	0.85	0.67	0.72	0.69
Hip Hop	0.49	0.43	0.46	0.47	0.28	0.35
House	0.40	0.40	0.40	0.21	0.16	0.18
Indie Dance / Nu disco	0.52	0.54	0.53	0.34	0.32	0.33
Left Bass				0.50	0.29	0.37
Left House Techno				0.49	0.40	0.44
Minimal	0.55	0.59	0.57	0.42	0.52	0.47
Progressive House	0.44	0.43	0.44	0.34	0.28	0.31
Psy Trance	0.88	0.94	0.91	0.78	0.86	0.82
Reggae Dub	0.62	0.56	0.59	0.52	0.33	0.40
Tech House	0.41	0.41	0.41	0.48	0.48	0.48
Techno	0.39	0.37	0.38	0.48	0.61	0.54
Trance	0.78	0.82	0.80	0.76	0.72	0.74
Trap				0.50	0.38	0.43

Table 5. Precision, recall and F1 score for each subgenre in Set 1 and Set 2.

Subgenre	Precision	Recall	F1 score	F1 Difference
BigRoom	0.6	0.33	0.43	-0.11
Breaks	0.55	0.6	0.57	0.03
Dance	0.35	0.24	0.29	0.03
Deep House	0.29	0.23	0.26	0
Drum & Bass	0.85	0.63	0.72	-0.04
Dubstep	0.70	0.52	0.59	0.10
Electro House	0.48	0.30	0.37	-0.05
Electronica Downtempo	0.34	0.49	0.40	-0.08
Future House	0.49	0.19	0.28	-0.08
Glitch Hop	0.57	0.43	0.49	-0.02
Hardcore Hard Techno	0.78	0.69	0.73	0.04
Hard Dance	0.64	0.60	0.62	-0.18
Hip Hop	0.28	0.12	0.17	-0.18
House	0.12	0.09	0.10	-0.08
Indie Dance Nu Disco	0.31	0.32	0.32	-0.01
Minimal	0.19	0.80	0.31	-0.16
Progressive House	0.38	0.08	0.13	-0.18
Psy Trance	0.64	0.88	0.74	-0.08
Reggae Dub	0.38	0.56	0.45	0.05
Tech House	0.33	0.49	0.40	-0.08
Techno	0.28	0.17	0.21	-0.33
Trance	0.29	0.1	0.15	-0.59

Table 6. Performance of the classifier trained with Set 1 with the songs of Set 2.

List of Figures

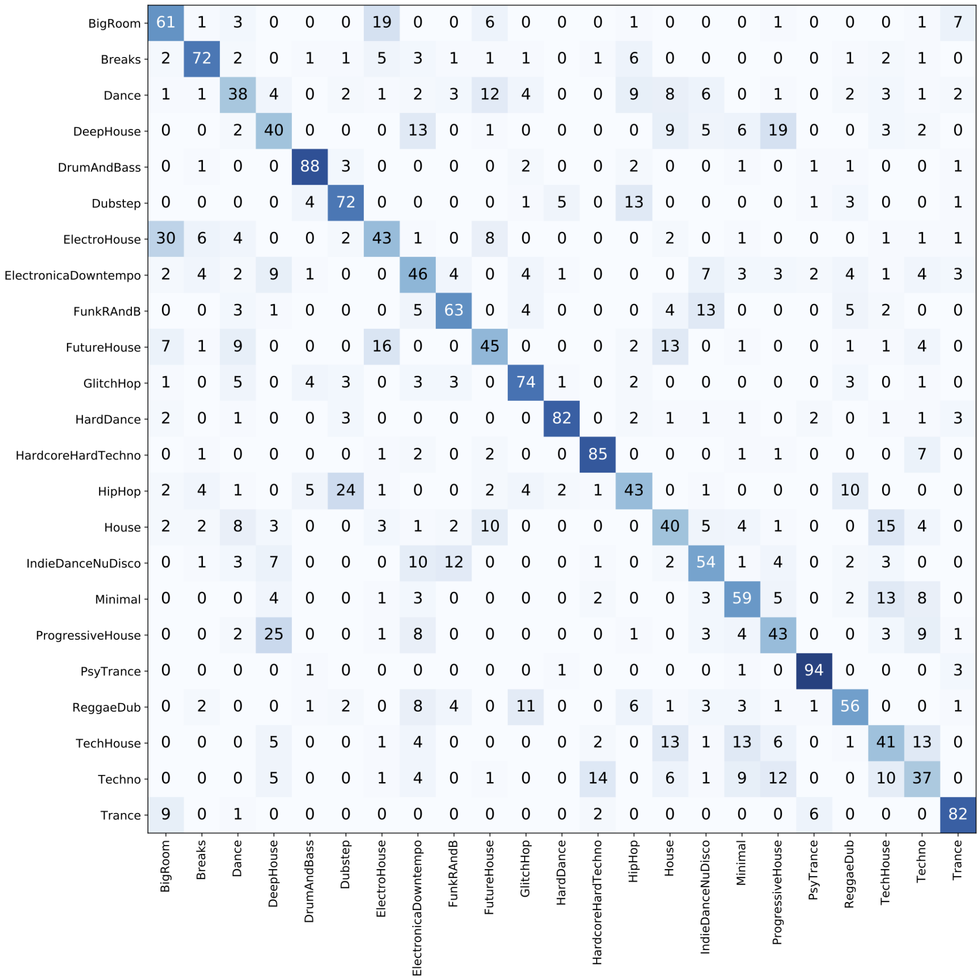

Figure 1. Confusion matrix of the Set 1 (true classes in rows, predicted classes in columns)

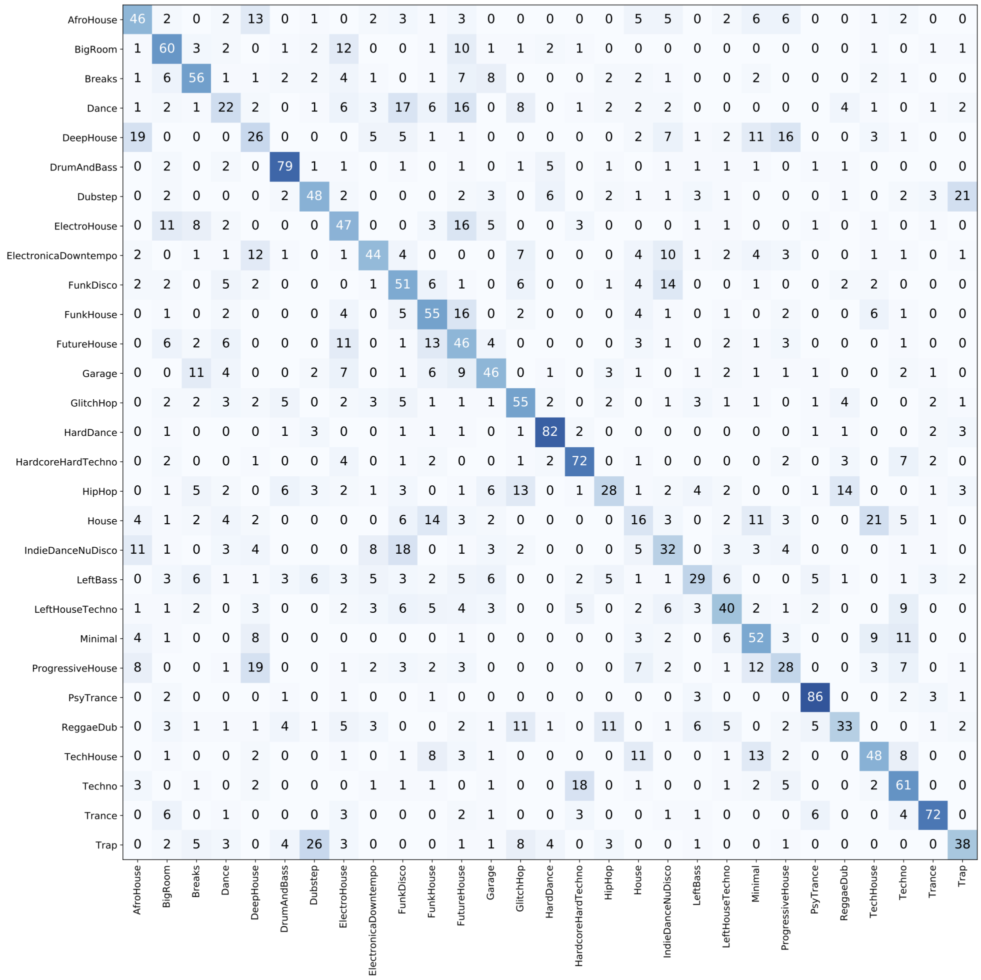

Figure 2. Confusion matrix of the Set 2 (true classes in rows, predicted classes in columns)

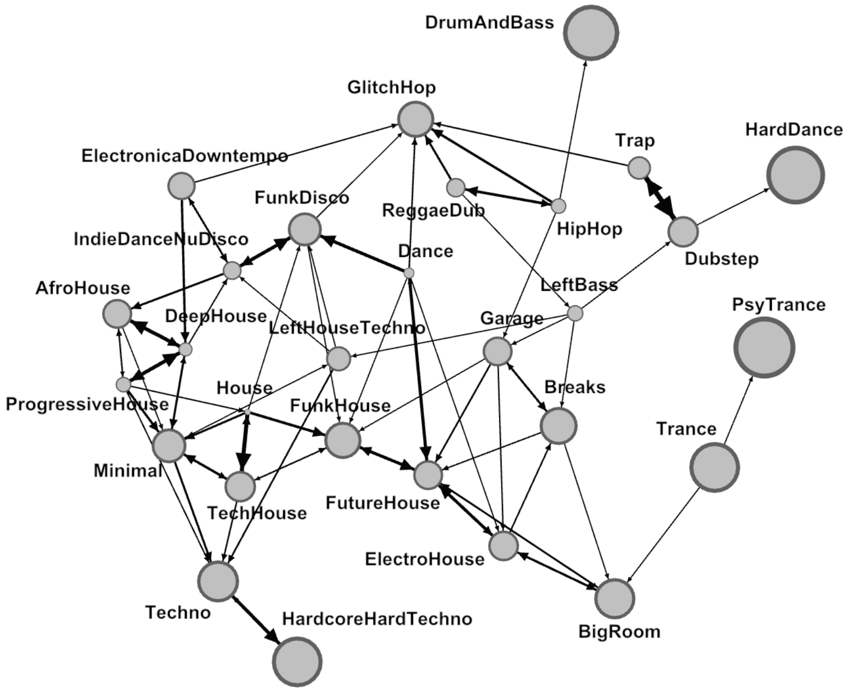

Figure 3. Confusion graph of the Set 2 (edges $weight \geq 5$)

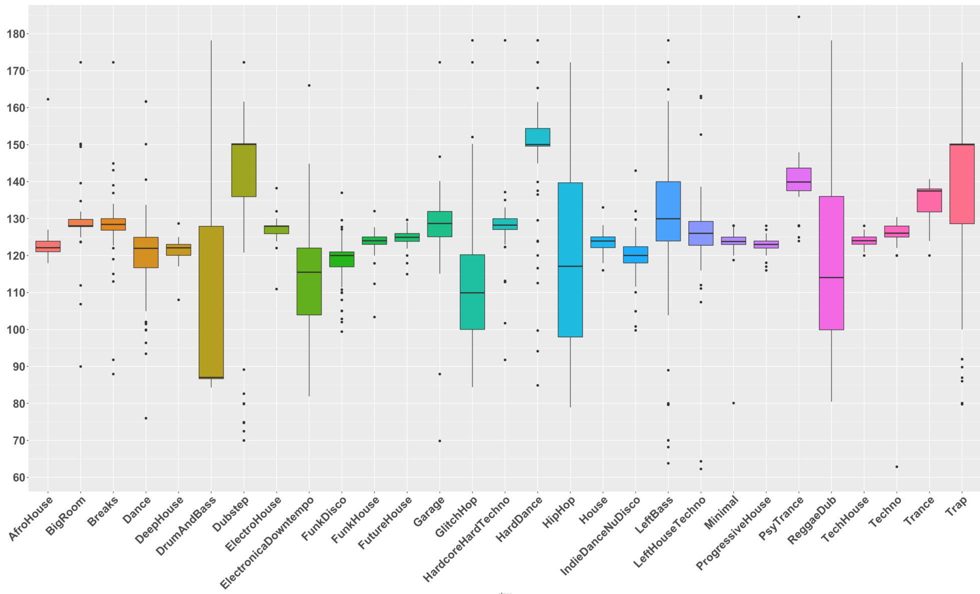

Figure 4. BPM boxplots by subgenre

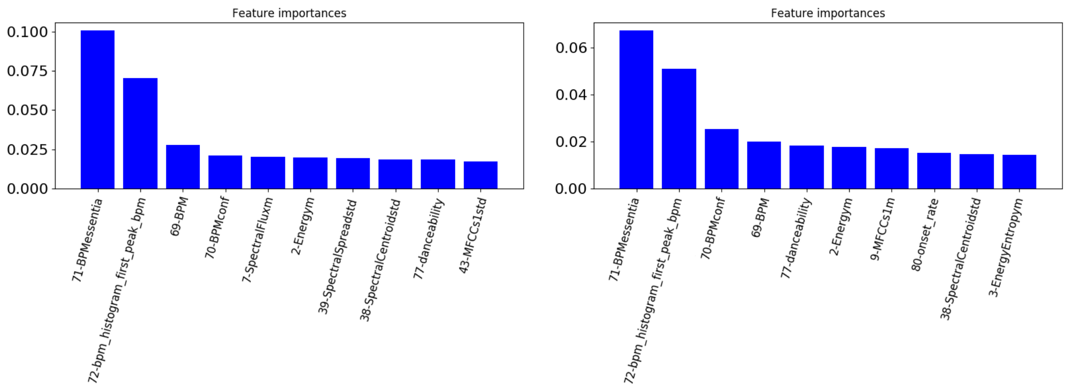

Figure 5. Feature importance of the best classifiers for Set 1 (left) and Set 2 (right)

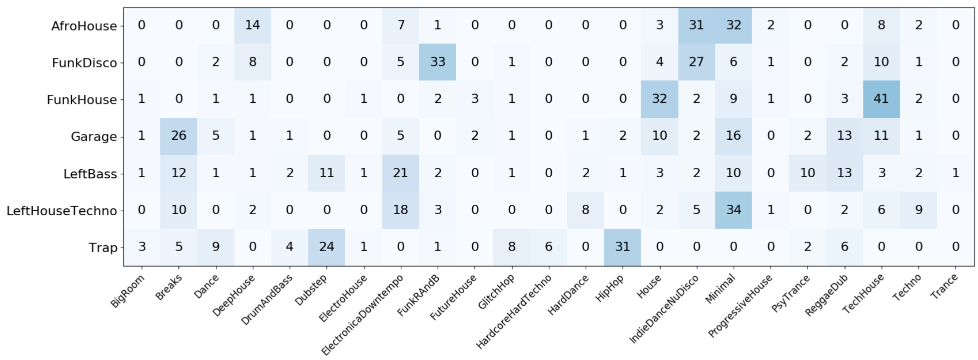

Figure 6. Confusion matrix to relate new subgenres in Set 2 with older ones