entropy

MDPI

Article

Statistical and Visual Analysis of Audio, Text, and Image Features for Multi-Modal Music Genre Recognition

Ben Wilkes, Igor Vatolkin * and Heinrich Müller

Department of Computer Science, Technische Universität Dortmund, 44227 Dortmund, Germany; ben.wilkes@tu-dortmund.de (B.W.); heinrich.mueller@tu-dortmund.de (H.M.)
* Correspondence: igor.vatolkin@tu-dortmund.de

Abstract: We present a multi-modal genre recognition framework that considers the modalities audio, text, and image by features extracted from audio signals, album cover images, and lyrics of music tracks. In contrast to pure learning of features by a neural network as done in the related work, handcrafted features designed for a respective modality are also integrated, allowing for higher interpretability of created models and further theoretical analysis of the impact of individual features on genre prediction. Genre recognition is performed by binary classification of a music track with respect to each genre based on combinations of elementary features. For feature combination a two-level technique is used, which combines aggregation into fixed-length feature vectors with confidence-based fusion of classification results. Extensive experiments have been conducted for three classifier models (Naïve Bayes, Support Vector Machine, and Random Forest) and numerous feature combinations. The results are presented visually, with data reduction for improved perceptibility achieved by multi-objective analysis and restriction to non-dominated data. Feature- and classifier-related hypotheses are formulated based on the data, and their statistical significance is formally analyzed. The statistical analysis shows that the combination of two modalities almost always leads to a significant increase of performance and the combination of three modalities in several cases.

Keywords: music genre recognition; multi-modal classification; feature evaluation; audio signal features; album cover images; lyrics

Citation: Wilkes, B.; Vatolkin, I.; Müller, H. Statistical and Visual Analysis of Audio, Text, and Image Features for Multi-Modal Music Genre Recognition. *Entropy* **2021**, *23*, 1502. https://doi.org/10.3390/e23111502

Academic Editor: Gholamreza Anbarjafari

Received: 1 October 2021
Accepted: 9 November 2021
Published: 12 November 2021

1. Introduction

Music genre recognition is one of the most common classification tasks in music information retrieval, with several hundreds of published studies mentioned by Sturm [1]. Traditional approaches are usually based on an individual feature source, mainly the audio signal. Because different modalities beyond audio, such as text, images, or symbolic representations, may contain complementary information, multi-modal approaches bear great opportunities to improve the classification performance. In this work, we present a multi-modal genre recognition framework that considers audio, text and image features of a music track by features of audio tracks, album cover images, and lyrics.

Because, in the field of image and text classification, artificial neural networks achieved comparatively good classification results to date [2,3], a group of text- and image-based features computed by artificial neural networks is taken into account in our framework. However, the features automatically learned by a neural network are often less interpretable and can also have a poor generalization ability because of a typically very large number of parameters of a trained neural network. Therefore, a further group of handcrafted text- and image-based features is additionally employed, which have been successfully used for image or text classification tasks in the past. For audio, several groups of features related to harmony, rhythm and tempo, timbre, and musically meaningful semantic properties from previous work predicted by supervised classification models are considered.

Genre recognition is performed based on binary classification of a music track with respect to each genre. From the results of the genre classifiers, the membership to one of the

genres is predicted, and a confidence value for this prediction is given. Besides combination of features into fixed-length feature vectors, a second approach of feature combination in form of confidence-based fusion of predictions obtained from several feature vector-based predictions is employed. This allows a detailed representation of longer audio tracks by a length-dependent number of feature values. Combinations of text features and combinations of image features, as well as combinations of text and image features are handled by feature vectors, whereas combinations of audio features and of audio, text, and image features are handled by confidence-based fusion.

A special focus is placed on performance analysis. To assess the influence of classifiers and features on the quality of genre recognition, we have determined the balanced classification error experimentally for three classifiers and numerous feature combinations. The resulting error values are presented visually. To improve perceptibility, data reduction techniques based on multi-objective analysis and restriction to non-dominated data are proposed and applied. Based on these data, feature- and classifier-related hypotheses are formulated and their significance is statistically tested. The global finding is that the combination of features from two modalities yields a significant reduction of the classification error for the majority of use cases. The extension to three modalities leads in several cases to further significant improvements.

The following Section 2, presents a review of related work and the contributions of the paper in this context. Section 3 describes our approach in detail. Section 4 deals with the experimental evaluation of the proposed system. Section 5 provides a summary of results and an outlook on possible future work.

2. Related Work

Common approaches to genre classification focus on audio features and their combinations. This is motivated by the situation that audio features describe or correlate to many different musically meaningful properties of a music piece and can be extracted when the digital score is not available. In one of the first related studies, Tzanetakis and Cook [4] introduced features to represent pitch, rhythmic structure, and timbre. Based on these features and their combinations, a Gaussian classifier for music genre classification was trained. Lidy and Rauber [5] presented different rhythm characteristics and compared their performance when used for music genre classification. In addition, the influence of psychoacoustic transformations on rhythm features was considered to improve the classification performance. Scaringella et al. [6] also provided an overview of various audio features describing timbre, harmony, and rhythm, and examined the impact of using different classifiers.

One possibility for applying image classification methods for genre classification is the use of direct visual representation of music. Bainbridge and Bell [7] and Burgoyne et al. [8] extract musical notes and lyrics from images of scores. Another concept is to convert the audio signal into a two-dimensional image representation (e.g., a spectrogram) and to apply further image processing methods. For instance, Ke et al. [9] used spectrograms to identify related music pieces. Another option is to use image-based information, which is often associated with music, especially album covers, photographs, and videos. Dorochowicz and Kostek [10] conducted a study with the aim to find out whether there exists a relationship between typographic, compositional, and coloristic elements of the music album cover design and genre of the music contained in the album. Le [11] measured the color similarities of album covers based on various genres and presents a study to verify whether the average listener can determine the genre of contemporary albums based on the graphics displayed on album covers. Schindler [12] discussed the role of visual information for music information retrieval and music genre classification, presents methods for the use of image information, analyzes them on the basis of images from music videos, and draws conclusions about their significance for album covers, as well. Oramas et al. [2,3] used album covers as image component for multi-modal genre classification from audio, text, and images. This work will be discussed in more detail later in this section. Libeks and

Turnbull [13] presented an image classification system that is able to estimate the similarity of music artists or to determine related genres based on album covers and photos of the artists. A data set of artists was built along with genre annotations and their most popular album covers and photos. The classification system calculates, for each photo and cover of a given artist, the most similar image from the data set. For each resulting image, the genre annotations of the associated artist were collected and then averaged over the data set, leading to genre prediction for a given artist.

Lyrics are more commonly used as an information source than album cover art. Logan et al. [14] estimated the similarity of artists based on their lyrics and compared the results with an audio-based approach, which achieved better results. The authors suggested to combine audio and text features to get better results. Other studies applied lyrics features for mood prediction [15,16].

The combinations of features from different sources for music classification are until now not very thoroughly explored. In the following, we provide some references. For a recent overview, we refer to Simonetta et al. [17].

Most studies on multi-modal music classification combine two sources. Audio, together with lyrics, seems to be the most frequent case. These sources were applied for genre recognition [18–21], mood and emotion recognition [22–25], artist identification [26], hit song prediction [27], and playlist prediction [28]. Audio and symbolic features were used for genre recognition [29,30]. Audio and images were employed for mood prediction [31] and genre recognition [12].

Rather few studies addressed three and more feature sources. Audio, cultural, lyrics, and symbolic descriptors were combined for genre recognition by McKay et al. [32] and audio, symbolic, and lyrics descriptors for mood detection by Panda et al. [33].

To our knowledge, the papers by Oramas et al. [2,3] are the only published works that deal with music genre classification on the basis of image, text, and audio-based features. Three separate artificial neural networks were trained on album covers, audio tracks, and album reviews. As inputs, the audio signals were converted into spectrograms and the album reviews into a bag-of-words representation. After the training, the three resulting networks were combined into a new network by reconnecting some layers and re-training. This network is used in our work for the extraction of image features.

Although a general concept of our framework is inspired by Oramas et al. [3], there exist several important differences. First, we also take handcrafted features into account, but apply classification methods for genre prediction, which have significantly fewer parameters than deep neural networks. This can help to create more interpretable models based on semantic features and has a further advantage in that the models can be trained with very small data sets. For example, when a listener defines a new category based on only a few representative tracks, models with many parameters, such as neural networks, will tend to overfit in that real-world application scenario. Second, we estimate text features only from lyrics and not album reviews. Although it is argued by Oramas et al. [3] that relevant genre information must not be captured in reviews, and, thus, reviews "will unlikely comply with the current taxonomy of the collection to be classified", it is a safer way to consider only lyrics. Third, audio tracks in reference [3] are always represented with 15-s frames, as the convolutional networks expect a fixed-size input. However, particularly for more complex genres and styles with very different segments, the analysis of complete music tracks may be useful and important information may be omitted when a frame of a fixed size is used for each track independently of its length. We handle this issue by considering multiple fixed-length frames by confidence-based feature combination.

Further differences between our work and [3] include (a) the approach to fuse the results of classification models for each modality based on the confidence level, which is estimated differently for individual modalities (Section 3.5), (b) the method for visualizing experimentally collected performance data for comparing the influence of different feature combinations and classifier models (Section 4.2), and (c) rigorous statistical testing of hy-

potheses, which underlines that, in some cases, the combination of several modalities does not necessarily lead to a significant improvement of the classification quality (Section 4.3).

3. A Multi-Modal Approach to Music Genre Recognition

In the following, we present the backgrounds and the details of our framework. Section 3.1 starts with a brief discussion on music genres and a description of our data set. Section 3.2 provides an overview of our approach. Section 3.3 describes audio-, text-, and image-based features used in our study. Section 3.4 briefly summarizes the classification algorithms used. The fusion of classification models trained separately for individual modalities is introduced in Section 3.5.

3.1. Data Set

Moore [34] refers to a *music genre* as a set of musical events, the scope of which is determined by specific generally accepted rules. Often, music pieces of the same genre have similar characteristics in instrumentation, rhythm structure, and pitch content [4]. Music genres, however, have never been formally defined [35], so that the assignment of music pieces to genres is often a matter of personal interpretation. In particular, music pieces could be assigned to different genres at the same time.

Therefore, the genre annotations used in our work are subjective and represent only one possible scenario. Because not all modalities can always be automatically extracted, we have created an in-house multi-modal data set of 446 tracks compiled from several music collections: 1000 songs, 1517-artists, SALAMI, SLAC, and an album collection of TU Dortmund. Appendix A provides details about these collections. The genres to predict are *Rock Rap/Hip-Hop*, *Electronic*, *Folk/World/Country*, *Blues*, *R&B*, *Jazz*, *Pop*, *Classical*, and *Reggae* (sorted by the number of corresponding tracks). Each music track is assigned to exactly one genre. Appendix B provides the details of the reassignment of the music genres of the original data sets to the newly created data set. Figure 1 shows the distribution of the genres.

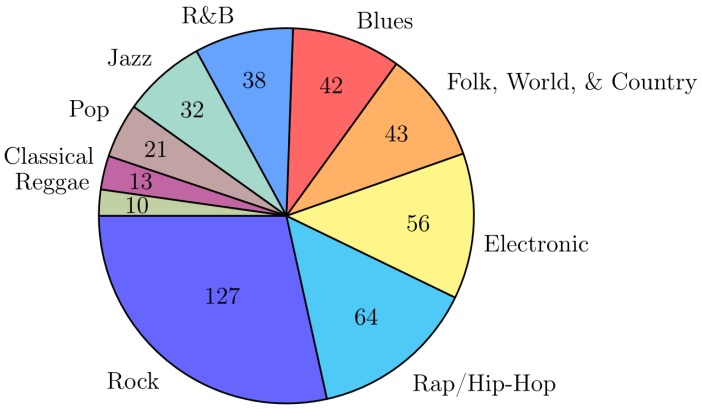


Figure 1. Composition of the genres of the created data set.

To find album covers, music texts, and genres, the Internet databases of Discogs [36] and MusicBrainz [37] for album covers, the Internet databases of MetroLyrics [38], LyricWiki [39], CajunLyrics [40], Lololyrics [41], and Apiseeds Lyrics [42] for lyrics and the database of Discogs for genres were used, in that order.

3.2. General Approach

We treat music genre recognition as a classification problem, which maps objects (here, music tracks) to classes (here, genres). We adopt the two-step approach of classification, which first assigns features to the objects and then uses them to perform the classification. Parametrized statistical models are employed as classifiers, which are trained in a preprocessing step by supervised learning. The training procedure adjusts the parameter

values, so that the objects of a given training set, whose classes are known, are classified as correctly as possible.

We use two types of features. The first type are features that have proven to be particularly useful in the field of audio, text, and image classification. The second type of features result from classifying artificial neural networks. Such neural networks combine feature assignment and classification. The features depend on parameters, whose values are determined simultaneously with the parameters of the classification step by training. We use features computed in this way analogous to the features of the first type. Section 3.3 provides an overview of the features employed in this paper.

Genre recognition is performed by binary classification of a music track with respect to each genre based on combinations of elementary features. A binary classifier is assigned to each genre, which decides whether a music track belongs to the genre (a positive prediction) or does not belong to the genre (a negative prediction). We employ three classifier models, Naïve Bayes, Support Vector Machine, and Random Forest, which are briefly recalled in Section 3.4.

For feature combination a two-level technique is used. The first level is feature aggregation into fixed-length feature vectors. Combinations of text features and combinations of image feature, as well as combinations of text and image features are handled in this way. From the results of applying all genre classifiers to such a feature vector of a piece of music, the membership to one of the genres is predicted, and a confidence value for this prediction is given. The second level of feature combination is confidence-based fusion of predictions obtained from several feature vector-based predictions. Combinations of audio features and of audio, text, and image features are handled in this way. Section 3.5 presents the details of this approach.

3.3. Features

In the following, we present audio, text, and image features estimated from audio tracks, lyrics, and album covers.

3.3.1. Audio Features

Audio features are calculated from 22,050 Hz mono wave files converted from original mp3 tracks. The description and grouping of the audio features described below is based on previous work [43].

Tempo and Rhythm

A typical characteristic of the temporal progress of a music piece is the number of beats per minute, where the beat events correspond to perceived sound pulses with highly repetitive structure. The rhythm is described by the special arrangement of the note lengths and accentuation in a music piece [44]. To describe the rhythm of a music piece, for example, the change in the loudness of certain sub-frequency bands can be examined, such as fluctuation patterns [45]. Rhythm must be differentiated from tempo because a particular rhythm pattern can be played in different tempo; therefore, they are not firmly connected. However, rhythm and tempo are strongly related, as they describe the temporal aspect of the music piece. They often consist of autocorrelation (the correlation of the audio signal with itself after an additional time lag). Appendix C.1 lists all tempo and rhythm features, together with their dimensionality and related references.

Timbre

Timbre can be defined as the part of the auditory sensation that allows the listener to distinguish between two sounds that have the same loudness and pitch [46]. The timbre depends, for instance, on the instrument used or the way it is played. Features that describe the timbre can be grouped by their extraction domains, such as time domain (e.g., the root mean square energy), spectrum (spectral centroid), cepstrum (MFCCs), or phase domain (angles in the phase domain). Appendix C.2 lists all timbre features used in our study.

Harmony

Harmony can be defined as the relationship between simultaneously played notes and the way how these relationships change over time, cf. reference [47]. The difference in tone frequencies between two notes played at the same time is called an "interval". Intervals may be consonant or dissonant, i.e., sounding pleasing/perfect or unpleasing/tense to listeners, which, however, can be perceived subjectively. The ratio of consonant to dissonant intervals is central to the study of the harmony of music. In addition, the transform of the frequency amplitudes to the halftones (chromagram or pitch class profile) can be treated as a harmonic feature because it serves as a base feature for more complex properties, such as chords or keys. Appendix C.3 shows the harmony features.

Semantic Features

Semantic features describe characteristics of the piece of music, which are related to music theory, such as the instrumentation, characteristics of the voices in the song, or the mood expressed. To capture semantic features from digitally represented music, various classifiers have been trained on a set of audio features, using multi-objective feature selection and ensembles of classifiers, with some semantic features derived or also predicted from other semantic features as introduced in reference [43]. The corresponding descriptors are listed in Appendix C.4.

3.3.2. Text Features

Two text feature groups are used, which are induced by the multidimensional Bag-of-Words feature and the doc2vec feature described below.

Bag-of-Words (BoW) Feature

In its simplest version, the BoW text feature [48] measures the occurrence frequencies of words from a given domain of words. The result is a real vector whose components correspond to the words of the domain. Before the feature estimation some preprocessing procedures are typically applied [49] (p. 242). In this work, stop words, such as "is", "to", or "with", are removed, and words are substituted with their stems, such as "lov" for "lover" and "loving". Furthermore, the frequency of a word is measured with the *Term Frequency-Inverse Document Frequency* (TF-IDF). TF-IDF is the product of the relative frequency and the *Inverse Document Frequency* (IDF). IDF is the inverse of the relative frequency of occurrence of a word in a document from the document collection under consideration. The motivation for using TF-IDF is that infrequent words could describe important text properties.

Advantages of BoW features are the fast estimation and high comprehensibility. Disadvantages are the loss of information about the order of the words, as well as a possible high dimension of the feature vectors, which depends on the number of different words used.

Doc2vec Feature

As in the BoW feature, the doc2vec feature [50] comes from the discipline of document classification. The doc2vec feature extends the idea of the word2vec-feature [51].

Word2vec and doc2vec are methods for the numerical representation of words or documents in a vector space. One simple option for such a representation is *one-hot-coding*. This means that, for every possible word or document, respectively, there is exactly one vector component whose value is 1 for the represented word, and 0 otherwise. In contrast to this, word2vec methods represent the words of a vocabulary in a latent space, which has a lower dimension than one-hot-encoding and aims to store context information of words. Doc2vec expands the latent space by a representation of documents by low-dimensional vectors, which store the context information document-specifically. In both cases, the vectors result from weights of neural networks.

One approach of word2vec uses fully connected neural networks with one hidden layer that outputs for two input words the probability of all words to occur as the middle

word in the context of the input words. The hidden layer calculates a feature vector, which is used by the output layer to determine the context probabilities. The input layer, such as the output layer, has a neuron for each word of the vocabulary, for one-hot coding. Each input neuron is connected to all hidden layer neurons. The weights on these connections form the representing vector of the word. Its dimension and, thus, the dimension of latent space is the number of neurons in the hidden layer. A simple introduction is given by Skansi [52] (Chap. 9).

For doc2vec, input neurons are added for the documents, whose vectors are then constructed accordingly.

Compared to BoW features, doc2vec features have the advantage that the information about the contextual relation of the words is included in the feature calculation. While the number of lyrics increases, the dimensionality of the feature vector does not increase, as for BoW features, because it is not dependent on the diversity of the words in the lyrics, but is an adjustable parameter. By using artificial neural networks, however, the interpretability and the explainability suffer because the semantics of doc2vec features is hardly comprehensible for humans.

3.3.3. Image Features

Two image features groups are used, which are induced by the Bag-of-Features with SIFT descriptors and deep convolutional neural network features.

Bag-of-Features (BoF) with SIFT Descriptors

SIFT stands for "*Scale -Invariant Feature Transform*" [53]. SIFT features are local image pixel descriptors and are invariant against rotation, scaling, and displacement. A SIFT descriptor is a 128-dimensional vector, which encodes properties of a pixel and its local environment in the image: a size, a position, an orientation, and further characteristics of its environment.

The BoF feature is an extension of the principle of the BoW feature to other data types than text. In the case of images, a *visual vocabulary* of so-called *visual words* is constructed. A BoF feature produces a real-valued vector, which measures frequencies of each visual word. Frequency measures can be, such as for BoW features, the absolute and relative frequency or TF-IDF. The dimension of the BoF feature depends on the size of the visual vocabulary.

To construct the visual vocabulary, local image features are determined for the images of the training set at first. The resulting set of image features is assigned to k clusters by applying a clustering algorithm, where k is the desired size of the visual vocabulary. In this work, we apply the *k-means algorithm* by Lloyd [54], which also estimates a cluster center for each cluster. The set of cluster centers forms the visual alphabet.

Being related to the BoW feature, the BoF feature has similar weaknesses. It also loses contextual information because information about the locations of the local image features on the images is neglected. In contrast to BoW features, the size of the vocabulary is a freely adjustable parameter. Nevertheless, it can be assumed that, as the number of images increases, so also does the diversity of descriptors and more cluster centers should be used.

Features of Deep Convolutional Neural Networks

Deep convolutional neural networks have a high number of hidden layers and are particularly successful for image classification. For this purpose, various network architectures are known, e.g., the ResNet [55]. To estimate a feature vector for the given image, it is used as input in an image classification network. The output values of the last hidden layer, which are also the input values of the classifier section, build the feature vector. In this work, the network by Oramas et al. [2] based on *ResNet101* by He et al. [55] is used to classify the album covers. The obtained features are in the following called "*DNN features*" or "*DNNF*".

3.3.4. Reduction of Dimension by Principal Component Analysis

The text and image features presented can be high-dimensional, depending on their extraction parameters. This may cause the problem that the number of available pieces of music (see Section 3.1) is not sufficient to train the classifiers with acceptable general performance. For this reason, *Principal Component Analysis* (PCA) [56] is additionally applied to reduce the dimensionality of the corresponding feature vectors.

3.4. Classifiers

The supervised classification algorithms listed below operate on numeric feature vectors of fixed dimension.

Naïve Bayes

The Naïve Bayes classifier was originally designed by Maron [57] for the classification of text documents. According to Qiang [58], it is very efficient and provides good results in many applications. However, if the structure of the feature vectors deviates strongly from the assumption of independence, the classification quality suffers. It can be assumed that some of the features in our study are dependent, so that this circumstance is a weak point of the classifier in the application of this work.

Linear Support Vector Machine

The linear *Support Vector Machine* (SVM) [59] classifies items by placing hyperplanes in feature space and determining the class membership of an item to be classified by the location of its feature vector relative to the hyperplane. The location of the hyperplane is determined by training the SVM. The details are described by Cristianini and Shawe-Taylor [60]. Linear SVMs are known to provide good classification results even for high-dimensional feature vectors and comparatively little training data. Linear SVMs can additionally be calculated very efficiently. However, if the separation of the data of the problem by its position in the feature space cannot be approximated by a linear hyperplane, then, linear SVMs show high error rates.

Random Forest

The Random Forest classifier [61] is based on a set of decision trees that vote by majority over the class of a feature vector. The Random Forest uses the *Classification and Regression Trees* (CARTs) [62]. CARTs have many applications in machine learning because they are invariant in scaling and many other transforms of feature vectors. Furthermore, they are also robust against inserting irrelevant data and create models that can be read and understood by humans. However, their classification performances are seldom good (reference [63], p. 352), as they tend to overfit [64]. To counteract this property, the Random Forest classifier uses modified CARTs along with the bagging developed by Breiman [65]. For further information, we refer to the remarks of Au [64].

3.5. Fusion of Binary Models Trained for Individual Genres and Modalities

The multi-modal genre recognition in this study is based on binary decisions. This means that, for each genre $g \in \{1, \ldots, G\}$, an individually trained classification model indicates whether a given music piece belongs to this genre.

For the fusion of binary models, which predict genres based on individual modalities, we distinguish between three cases to estimate confidences for genre predictions: (1) *audio features only*, (2) *a combination of text and image features only*, and (3) *a combination of audio, text, and image features*. The final decision based on all modalities takes into account the confidences of predictions of the cases (1) and (2), as described below.

In the *subcase* (1), genre predictions are first done on time intervals (classification frames) of 4 s length with 2 s overlap. The aggregation of features along the complete music track would decrease the classification performance because, even for tracks of the same genre, each music piece typically contains several different segments with respect to

instrumentation, harmonic, and rhythmic properties. Let W_m be the number of classification frames in the music track m, which is represented with feature vectors $\vec{x}_1(m), \ldots, \vec{x}_{W_m}(m)$. Let $\hat{y}_w(m,g) \in \{0,1\}$ be the prediction for the w-th classification frame (equal to 1 when this frame is predicted to belong to the genre g and 0 otherwise). The assignment of complete tracks to genres is done by majority voting (index "a" stands for audio):

$$\hat{y}_{\mathrm{a}}(m,g) := \hat{y}_{\mathrm{a}}(\vec{x}_1(m), \ldots, \vec{x}_{W_m}(m), g) = \left\lceil \frac{1}{W_m} \cdot \sum_{w=1}^{W_m} \hat{y}_w(m,g) - \frac{1}{2} \right\rceil, \tag{1}$$

and the *confidence of the prediction* based on audio features is given as:

$$c_{\mathrm{a}(m,g)} = \begin{cases} \frac{1}{W_m} \cdot \sum_{w=1}^{W_m} \hat{y}_w(m,g) & \text{if } \hat{y}_{\mathrm{a}}(m,g) = 1 \\ 1 - \frac{1}{W_m} \cdot \sum_{w=1}^{W_m} \hat{y}_w(m,g) & \text{otherwise.} \end{cases} \tag{2}$$

In the *subcase* (2), vectors of text and image features can be simply concatenated because they have the same length for all music pieces. The confidence of the prediction $\hat{y}_{\mathrm{it}}(m,g)$ (index "it" stands for image and text) for music piece m and genre g depends on the number of positive predictions for all other genres $i \in \{1, \ldots, G\} \setminus \{g\}$:

$$c_{\mathrm{it}}(m,g) = \begin{cases} 1 - \frac{1}{G-1} \cdot \sum_{i \in \{1,\ldots,G\} \setminus \{g\}} \hat{y}_{\mathrm{it}}(m,i) & \text{if } \hat{y}_{\mathrm{it}}(m,g) = 1 \\ \frac{1}{G-1} \cdot \sum_{i \in \{1,\ldots,G\} \setminus \{g\}} \hat{y}_{\mathrm{it}}(m,i) & \text{otherwise.} \end{cases} \tag{3}$$

Thus, the highest possible confidence $c_{\mathrm{it}}(m,g) = 1$ is given only if the music piece m is assigned to genre g by the binary classification model, which predicts this genre and is assigned as not belonging to all other genres by the related classification models.

In the *subcase* (3), the final prediction is made with respect to predictions and confidences of decisions done in subcases (1) and (2):

$$\hat{y}_{\mathrm{ait}}(m,g) = \left\lceil \frac{1}{2} \cdot (\hat{y}_{\mathrm{a}}(m,g) \cdot c_{\mathrm{a}}(m,g) + \hat{y}_{\mathrm{it}}(m,g) \cdot c_{\mathrm{it}}(m,g)) - \frac{1}{2} \right\rceil. \tag{4}$$

Training sets for each classifier are balanced, i.e., they contain the same number of positive (belonging to the genre to predict) music tracks and negative (not belonging to this genre) tracks, in order to avoid a bias of one of the classes. For this purpose, the set of initially available tracks for training is selected as follows. Let $V(g)$ be the number of tracks available, which belong to the genre g, and $\overline{V}(g)$ the number of tracks not belonging to this genre. For the data set described in Section 3.1, $V(g) < \overline{V}(g)$ holds for all genres. The number of negative training tracks is reduced to approximately $\overline{V}(g)$ by first sorting those songs according to their genres and then retaining only every $\lfloor \overline{V}(g)/V(g) \rfloor$-th element.

4. Evaluation

The main goal of the evaluation is to understand the influence of modalities and feature groups on music genre recognition. In addition, insights into the performance of different tested classifiers (Naïve Bayes, SVM, and Random Forest), in absolute terms and in comparison, should be gained. For this purpose, we formulate several hypotheses. *Feature-related hypotheses* are addressed in Section 4.3 and *classifier-related hypotheses* in Section 4.4. The focus of the evaluation of feature-related hypotheses is on statements on the effect of feature combinations. Their significance is assessed by statistical tests. Statements on the classifier-related hypotheses are based on the visual analysis of the data, which is discussed in Section 4.2. The configuration of experiments is provided in Section 4.1.

4.1. Configuration Of Experiments

The configurations of the text BoW and doc2vec features are summarized in Tables 1 and 2.

Table 1. Configurations of BoW-features.

	Vocabulary Line Size	Stop Word Line Removal	Stemming	TF-IDF	PCA
Configuration 1	400	yes	yes	yes	no
Configuration 2	400	yes	yes	yes	32

Table 2. Configurations of doc2vec-features.

	Size of the Hidden Layer	PCA
Configuration 1	100	no
Configuration 2	100	16

The configurations of the image BoF SIFT and DNN features are provided in Tables 3 and 4.

Table 3. Configurations of the BoF-features with SIFT-descriptors (SIFT_BOF).

	Vocabulary Size	PCA
Configuration 1	400	no
Configuration 2	400	16
Configuration 3	400	64

Table 4. Configurations of the features from deep neural networks (DNNF).

	PCA
Configuration 1	32
Configuration 2	64

The parameters for text and image features were determined experimentally in random samples. For this purpose, a grid search was executed on a strongly reduced version of the training data set. We studied vocabulary sizes of 25, 50, 100, 200, and 400 for doc2vec, BoF, and BoW features, and for all those features PCA parameters of 16, 32, and 64 dimensions. There is further optimization potential here. The audio features were calculated using the software AMUSE [66].

The linear SVM, the Random Forest with 100 trees, and the Naïve Bayes classifier were employed as basic classifiers. The models were validated based on the balanced classification error estimated during stratified cross-validation with $k = 5$ partitions (see Section 3.2). The balanced error is estimated from applications of a classifier on a test data set, which is independent from the training data set, and is defined as

$$e_{bal} = \frac{1}{2} \cdot \left(\frac{c_{1,2}}{c_{1,1} + c_{1,2}} + \frac{c_{2,1}}{c_{2,1} + c_{2,2}} \right), \tag{5}$$

where the parameters $c_{i,j}$, $i, j \in \{1, 2\}$ are the entries of the *confusion matrix*, which summarizes the numbers of positive and negative predictions (Figure 2).

Stratified cross validation divides the available data set into $k \geq 2$ non-overlapping partitions [67]. In k runs, each partition, in turn, is used as test set and the other $k - 1$ partitions form the training set, and the mean balanced test error across all runs is reported. Stratification ensures that the ratio of the different classes to predict in the partitions is approximately the same as in the given data set.

		annotation	
		positive	negative
prediction	positive	$c_{1,1}$	$c_{1,2}$
	negative	$c_{2,1}$	$c_{2,2}$

Figure 2. Confusion matrix of a binary classification problem.

4.2. Visual Data Analysis

The results are presented as *heat maps* (Figure 3). The horizontal axis corresponds to genres and the vertical axis to feature combinations. The entries of the resulting matrix contain the balanced error rates, additionally visualized with colors. In each column, the minimum with the best configuration per genre is marked with a frame. The genres are sorted in ascending order by the minimum of the related column. The vertical axis is grouped into blocks of feature combinations of the same modality and combinations of several modalities. The blocks are sorted in ascending order based on the number of modalities; the first block contains only combinations of audio features, the second one—image features, the third one—text features, the fourth one—combinations of audio and image features, etc.

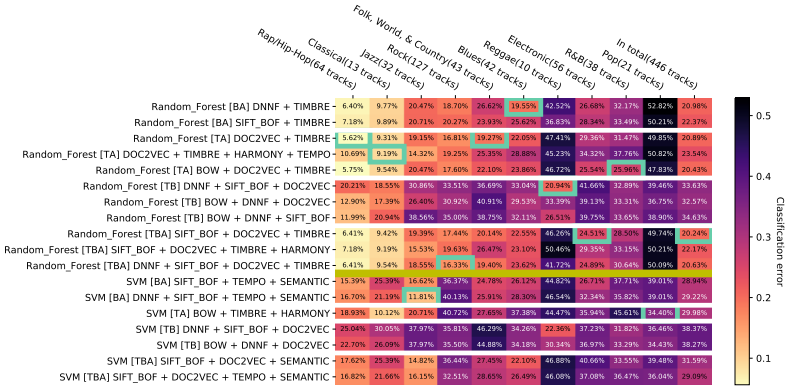


Figure 3. Visualization example.

All results are visualized in the Appendix D.1. It is very difficult to provide general recommendations because of the large number of configurations and feature combinations. In order to reduce this effect, we propose three steps presented below: *aggregation of the same combinations of features*, *removal of dominated results*, and *filtering of less relevant results*.

4.2.1. Aggregation of the Same Combinations of Features

To reduce the number of lines in the visualizations, the results are aggregated by combinations of features that use the same features but different configurations for them. For example, the combinations

$$\begin{aligned} &\textit{SIFT\_BOF } (v = 400, \textit{pca} = \textit{no}) + \textit{TIMBRE},\\ &\textit{SIFT\_BOF } (v = 400, \textit{pca} = 16) + \textit{TIMBRE},\\ &\textit{SIFT\_BOF } (v = 400, \textit{pca} = 64) + \textit{TIMBRE}, \end{aligned}$$

aggregated as

$$\textit{SIFT\_BOF} + \textit{TIMBRE},$$

correspond to a vector whose components are the minimal errors across all aggregated combinations of features for each classifier and each genre. This aggregation is in the following called *minimum accumulation*. Appendix D.2 shows the minimum accumulation for the individual classifiers.

4.2.2. Removal of Dominated Results

It is desired to achieve the lowest possible error rates for each genre. The selection of features and their configurations can, therefore, be interpreted as a multi-objective minimization problem with G optimization criteria (errors for each genre). According to Zitzler et al. [68], a solution K_1 (feature configuration) *dominates* a solution K_2 if and only if the configuration K_1 has a better error rate e_{K_1} than e_{K_2} in at least one genre and no worse one in any other genre. Dominated feature configurations are not relevant for the investigation of certain hypotheses and can be removed from the views. The application of this method after minimum accumulation described in the previous section leads to Appendices D.3–D.5. In Appendix D.6, the results of all classifiers have been compiled, then the same configurations of features have been aggregated, and, finally, the dominated configurations of features have been removed.

4.2.3. Filtering of Less Relevant Results

As described in the previous section, the identification of the best feature groups can be understood as a multi-objective minimization problem with G objectives.

Let r be a *reference point* in the multi-objective space, which indicates the worst possible solution (all errors are equal to 1). When considering a solution K (selected feature group) in the objective function range, a volume exists with respect to r that is dominated by K (Figure 4a). All arbitrary solutions K_{dom} within this volume are dominated by K. This volume is called the "dominated hypervolume of solution K". Likewise, a set $\mathcal{K}$ of solutions has a dominated hypervolume (Figure 4b). It is the volume, in which all objective function values of all arbitrary solutions are dominated by at least one solution in $\mathcal{K}$.

Each non-dominated solution K from $\mathcal{K}$ contributes a part to the total dominated hypervolume of $\mathcal{K}$, which is dominated exclusively by K (Figure 4c). This volume can be calculated. To every solution, i.e., every combination K of features, a share v_K of the contribution to the total dominated hypervolume of $\mathcal{K}$ can, therefore, be assigned. A small v_K is an indicator that there are further solutions near K in the objective function range.

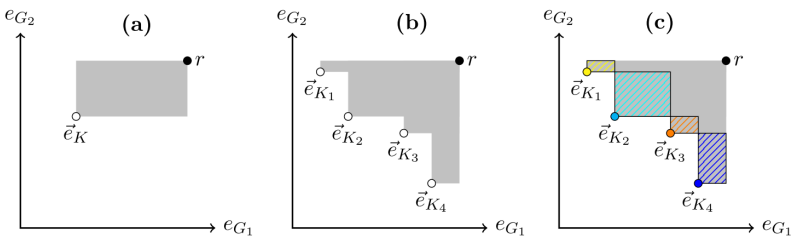


Figure 4. Visualizations of dominated hypervolumes with respect to the reference point r with a two-objective minimization problem. (**a**) Dominated hypervolume of a solution. (**b**) Dominated hypervolume of a set of solutions. (**c**) Individual contributions of solutions to the dominated hypervolume.

Feature combinations with small v_K-values may be less interesting when examining the hypotheses, since there are other combinations of features whose classification error rates are similar to that of K. In order to further reduce the visualization of the test results, feature combinations K are removed, for which $v_K < t \cdot \max(v_{K'} \mid K' \in \mathcal{K})$, where $t \in [0, 1]$. The application of this approach with $t = 0.01$ and $t = 0.05$ after the removal of non-dominated results leads to Appendices D.7–D.14.

4.3. Feature-Related Hypotheses

The feature-related hypotheses are as follows:

M_1: The classification with audio-based features achieves a better error rate than the classification with non-audio-based features. Feature combinations are not examined here.

M_2: The combination of features of different modalities leads to a better error rate. More specifically:

- $M_{2,1}$: The combination of any features of two modalities results in a better error rate compared to using any features of one of the two modalities.
- $M_{2,2}$: The combination of any features of three modalities results in a better error rate compared to using any features of two of the three modalities.

M_3: Non-audio-based features achieve a better error rate for certain genres whose error rate is high when classified via audio features.

M_4: The use of principal component analysis for text and image features does not degrade the results with the respect to the classification error.

All hypotheses are examined via *Wilcoxon Signed Rank Tests* [69], checking whether the values of two paired samples are different. For this purpose, a *null hypothesis* H_0 and an *alternative hypothesis* H_1 are first set up. H_0 is an assertion about the observed error rates that the test is intended to refute. H_1 is the opposite of H_0, i.e., either H_0 or H_1 is true. The samples represent two observed error rates of different configurations corresponding to H_0. Then, a *significance level* α, $0 < \alpha \leq 1$, is chosen. It describes the probability of H_0 being incorrectly rejected by the test. Finally, the test is carried out. The result is a so-called p-value. If the p-value is below α, the test rejects H_0. The error rates examined differ significantly in this case. However, if the test does not reject H_0, this does not mean that H_0 is approved; rather, the null hypothesis is simply not rejected.

All hypotheses are examined with a commonly used significance level of $\alpha = 5\%$. All tests are performed on the error rates of the individual classifiers to check whether some hypotheses can only be confirmed or rejected by using certain classifiers. Since all hypotheses are analyzed by multiple tests, the significance level for individual tests is further lowered by the Bonferroni correction, as described in reference [70] (p. 247).

Details of the procedure are described in the following analysis of hypothesis M_4. This is done before the analysis of the other hypotheses because the Bonferroni correction can be explained well on the basis of this hypothesis. The data basis for hypothesis M_4 is the error rates shown in Appendix D.1. The null hypothesis H_0 of the test is that the error rate remains the same when using PCA. To test H_0, sub-hypotheses are set up comparing configurations with and without PCA. Examples are:

$H_{0,1}$: The use of BoW features without PCA achieves the same error rate as the use of BoW features with a PCA with dimensionality reduction to 64 dimensions.

$H_{0,2}$: The use of BoW features without PCA achieves the same error rate as the use of BoW features with a PCA with dimensionality reduction to 32 dimensions.

$H_{0,3}$: The use of BoF features without PCA achieves the same error rate as the use of BoF features with a PCA with dimensionality reduction to 64 dimensions.

H_0 must be rejected as soon as at least one of $H_{0,1}, H_{0,2}, \ldots, H_{0,k}$ is rejected. Let α_k be the level of significance, with which the tests on the hypotheses $H_{0,1}, H_{0,2}, \ldots, H_{0,k}$ are performed. Then, there is the likelihood of falsely rejecting one of these hypotheses at $1 - (1 - \alpha_k)^k$. If we want to test α on H_0 with a significance level α, then, $\alpha_k = \alpha/k$ can be chosen because $1 - (1 - \alpha/k)^k < \alpha$ for $k > 1$. The Bonferroni correction describes this procedure. Instead of changing the significance level $\alpha_k = \alpha/k$, the p-value obtained by the test can be equivalently also adapted to $p_k = p \cdot k$. The Bonferroni correction is used in all subsequent tests.

To test *hypothesis* M_4, the error rates are tested against each other using different feature configurations K_1 and K_2. K_1 will be tested against K_2 if all of the following conditions are true:

- K_1 and K_2 are no combinations of individual features groups.
- K_1 and K_2 are only features of type BoW, doc2vec, or BoF.

- K_1 and K_2 are the same feature type.
- K_1 does not use PCA, K_2 uses PCA.

The results of the tests can be found in Tables A10–A12. For all tests, the null hypothesis is retained for all classifiers. Thus, the results using PCA do not differ significantly from results that did not use PCA. Hypothesis M_4 is, therefore, not rejected, meaning that the number of features can be significantly reduced without a decrease of the classification performance.

For the analysis of *hypothesis* M_1, only those feature configurations are considered from Appendix D.2, which do not consist of feature combinations. These are then partitioned by modality. The individual partitions are summarized by minimum accumulation (see Section 4.2.1). This results in three vectors $\mathbf{e}_a$, $\mathbf{e}_i$, $\mathbf{e}_t$ of error rates for audio, image, and text features. Then, using the Bonferroni correction $\mathbf{e}_a$ against $\mathbf{e}_i$ and $\mathbf{e}_a$ against $\mathbf{e}_t$ are tested. The results of the tests can be found in Tables A1–A3.

For Random Forest and the Naïve Bayes classifier, at least one of the null hypotheses is rejected. Audio features in this case provide results that are significantly different to text or image features. Since, in any case, the median error rate is lower when using audio features, we can agree with hypothesis M_1, at least when using Random Forest or Naïve Bayes classifier.

Hypothesis $M_{2,1}$ is checked by partitioning all error rates of the feature configurations from Appendix D.2 to modality combination. This results in the partitions *audio*, *text*, *image*, *audio* + *text*, etc., which are combined by minimum accumulation to form error rate vectors $\mathbf{e}_a$, $\mathbf{e}_t$, $\mathbf{e}_i$, $\mathbf{e}_{at}$, etc. Now, all error rates $\mathbf{e}_m$ are tested against $\mathbf{e}_n$, for which

- $\mathbf{e}_m$ belongs to a partition of exactly one modality (e.g., *audio*),
- $\mathbf{e}_n$ belongs to a partition of two modalities (e.g., *audio* + *text*),
- the modalities of the partition of $\mathbf{e}_n$ include the modality of the partition of $\mathbf{e}_m$.

The results of the tests adjusted by the Bonferroni correction can be found in Tables A4–A6. For all classifiers, at least two of the six resulting null hypotheses are rejected. Thus, there are partially significant differences in the error rates when using features of different modalities compared to the error rates when using features that belong to only one modality. For each rejected null hypothesis, the error rates using features of two modalities show a lower median, so the error is significantly better here. Hypothesis $M_{2,1}$ cannot always be approved, as not all null hypotheses are rejected. However, if we restrict ourselves to certain classifiers and modalities, such as Random Forest with audio and text features, the hypothesis can be approved. Accordingly, it seems to apply only in certain scenarios.

Considering the tests individually without using the Bonferroni correction, it is worth noting that almost every null hypothesis is rejected in favor of the combination of features of different modalities. Exceptions are the null hypotheses of the tests, which test the use of text features against the use of text and image features and the use of audio features against the use of audio and text features in classification via an SVM. It is also noticeable that the median error rate is 5% to 17% lower compared to using a non-audio feature when the non-audio feature is combined with an audio feature. Overall, it is apparent that the combination of features of two modalities almost invariably leads to an improvement in the error rate, whereby the inclusion of audio features in the feature combination seems to lead to the greatest improvement in error.

Hypothesis $M_{2,2}$ is investigated analogously to hypothesis $M_{2,1}$. Partitions by modality combination are created and again summarized by minimum accumulation. All error rates $\mathbf{e}_m$ against $\mathbf{e}_n$ are tested, for which

- $\mathbf{e}_m$ belongs to a partition of exactly two modalities (e.g., *audio* + *text*),
- $\mathbf{e}_n$ belongs to a partition of exactly three modalities (e.g., *audio* + *text* + *image*),
- the modalities of $\mathbf{e}_n$'s partition include all modalities of the partition of $\mathbf{e}_m$.

The results of the tests adjusted by the Bonferroni correction can be found in Tables A7–A9. For all tests, the null hypothesis is retained for all classifiers. Neglecting the Bonferroni correction, it turns out that taking image or audio features into the feature combination

with SVM as a classifier and including text or audio features in the feature combination using Naïve Bayes shows a statistically significant improvement in the classification quality. Altogether, contrary to the observations of hypothesis $M_{2,1}$, hypothesis $M_{2,2}$, therefore, cannot be generally confirmed. Nevertheless, the combination of features of three modalities in the cases mentioned brings an improvement in the error rate. For hypothesis M_2 as a whole, the combination of features of two modalities certainly brings an improvement in the error rate. However, adding more modalities does not necessarily improve the classification performance significantly.

For *hypothesis* M_3, from Appendix D.2 only error rates of the genres R&B, Reggae, Pop, and Electronic are considered. These genres were chosen because none of the classifiers are able to achieve error rates below 25% using only audio features. Since there are only four observations per potential test, no tests can be used. For this reason, this hypothesis is assessed using Appendix D.2. Considering the error rates of the classifiers using image or audio features only for the selected genres, it is easy to see that the use of non-audio-based features does not effect any noticeable improvement. Although some combinations of image features may bring an improvement to the reggae genre, this seems to be an exception, so that hypothesis M_3 is generally unconfirmed.

4.4. Classifier-Related Hypotheses

The classifier-related hypotheses are as follows:

M_5: The different classification methods have different error rates for the same features.

M_6: There are genres, for which certain classifiers achieve a better error rate for the same features than other classifiers.

To study *hypothesis* M_5, we first consider Appendix D.2. Here, the classification performances of the three classifiers with all feature combinations are shown. A first visual impression conveyed by the color coding is that the Naïve Bayes classifier delivers results that differ significantly from the results of the other classifiers. On a majority of genres, the classification error appears to be higher than the error of SVM and Random Forest. This is also evident from the absence of the yellow-orange block in the left-hand part of the diagram, which arises in the charts of SVM and Random Forest in that certain genres can be better classified almost independently of the feature selection. Looking at Appendix D.6, which summarizes the data in Appendix D.2 and outlined feature combinations, this assumption is confirmed. Most of the results of the Naïve Bayes classifier are dominated by other results. In Appendix D.14, feature combinations are removed that are less than $t = 0.05$ contributing to the dominated hypervolume of the total set. In this figure, no result of the Naïve Bayes classifier is listed. Therefore, on the features studied here, this classifier generally appears to provide higher error rates compared to SVM and Random Forest, so hypothesis M_5 can be agreed.

Hypothesis M_6 is checked using Appendix D.2. The sorting of the genres on the horizontal axis of the three visualizations is different, so the classifiers have different best error rates per genre. It is striking that the genres Rap/Hip-Hop, Classical, and Jazz are among the three genres of all classifiers that can be classified with the lowest error rate. The genres Pop, R&B, and Electronic are among the genres with the highest classification error rate for all classifiers. Thus, there seem to be tendencies of classification quality per genre, which are independent of the used classifier. However, there are also strong differences in the error rates of the individual classifiers. Random Forest provides noticeably better error rates on the Rock genre than the Naïve Bayes classifier and SVM when using feature combinations that include audio features.

The Naïve Bayes classifier, on the other hand, tends to achieve an error rate of approximately 50% in many genres when feature combinations with audio features are used. Further investigations show that this error rate arises because the classifier always classifies tested music pieces as not belonging to the genre to predict. This may be explained by the fact that audio features may have correlations with each other, which the Naïve Bayes

classifier cannot handle. It seems, therefore, that the Naïve Bayes classifier with audio features is a non-recommendable configuration for a genre recognition system. Overall, hypothesis M_6 can be approved.

5. Conclusions and Future Work

We have proposed a multi-modal genre recognition framework that considers the modalities audio, text, and image by features extracted from audio signals, album cover images, and lyrics of music tracks. The basis of recognition is binary classification, and the well-known and proven classifier methods (Naïve Bayes, Support Vector Machine, and Random Forest) were chosen for this purpose. Features were selected that are known to be particularly powerful in the domains of audio signal, text, and image, and an approach to their combination that meets the requirements of the features of the different modalities was presented.

Extensive experiments have been conducted for the three classifier models and numerous feature combinations. As no suitable data collection was available, an in-house multi-modal data set was compiled from several music collections. Determining the feature values required some effort, but it should be noted that the feature values are reusable. On the other hand, the training and application of the classifiers required comparatively little time. The training runtimes for the three classifiers used are low, compared to those often observed for end-to-end classifiers, such as deep neural networks.

The influence of the classifiers and the features on the classification quality was assessed by using the balanced classification error. The error values were presented visually by tables with color coding. Three approaches to data reduction were applied: aggregation of the combinations of the same, but differently configured features, removal of dominated results based on multi-objective non-dominated sorting of selected combinations of features and classifiers, and removal of less relevant results with small hypervolume contributions. The approach has proven successful for comparative visual analysis by allowing the range from a heatmap-like overview based on the original data to a detailed table-based view based on the reduced data.

The statistical comparison of all combinations of two modalities against individual ones always led to smaller classification errors. Those errors were also significantly smaller for all cases, except for text and audio modality against audio, and text and image against text using SVM. A more general hypothesis that "two modalities are always better than one" was confirmed by adjusted p-values after the Bonferroni correction for multiple tests for half of all combinations. The extension to the third modality further reduced the errors in almost all cases, but the general hypothesis that "three modalities are always better than two" could not be confirmed by adjusted p-values; the advantage rather depends on the classifier and features used.

For more robust genre recognition and music recommendation systems, future work should further extend the number of modalities (e.g., integrating MIDI scores, music videos, meta data), feature groups, and classification methods. To better understand the characteristics of music categories, it is possible to build and compare distinct feature sub-groups based on musical and statistical properties, extraction costs, availability in open-source frameworks, etc. Deep features can be extracted not only from the last hidden layer of the previously trained network but also from other layers, as proposed by Choi et al. [71]. For a more efficient identification of the best classification models, feature selection and systematic tuning of classifiers can be further applied. In addition, the experiments can be repeated using further data sets and genres or also other music categories, such as emotions or personal preferences. Last but not least, the demands on resources (runtime, storage space) can be measured.

Author Contributions: Conceptualization, B.W., I.V. and H.M.; methodology, B.W., I.V. and H.M.; software, B.W.; validation, B.W.; formal analysis, B.W.; investigation, B.W.; resources, B.W. and I.V.; data curation, B.W.; writing—original draft preparation, I.V. and H.M.; writing—review and editing,

I.V. and H.M.; visualization, B.W.; supervision, I.V. and H.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: In case the article will be accepted, we will release the data set with all extracted features.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BoF	Bag-of-Features
BoW	Bag-of-Words
CARTs	Classification and Regression Trees
DNNF	Deep Neural Network Features
MFCCs	Mel Frequency Cepstral Coefficients
SIFT	Scale-Invariant Features Transform
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
PCA	Principal Component Analysis

Appendix A. Data Sets

1517 artist data set:

The data set contains 3180 pieces of music by 1517 artists of varying popularity. Each piece of music is assigned to exactly one of 19 genres. This data set was presented by reference [72]. Album covers and lyrics are not given and had to be collected elsewhere.

1000 songs data set:

The data set compiled by reference [73] consists of 744 pieces of music that have been released under free license. It contains annotations for the emotion expressed for each piece of music, which is why it is often used in the emotion classification on pieces of music. This additional data is not relevant for the further experiments. In addition, missing here are album covers and lyrics, as well as the genres of the pieces of music.

SALAMI data set:

This data set compiled by reference [74] consists of 1383 pieces of music with annotations about the structure of each piece of music. However, this additional information is not needed in this work. Each piece of music is also assigned to one of 5 genres. Album art and lyrics are not given, so they had to be collected additionally.

SLAC data set:

The data set compiled by reference [32] consists of 250 pieces of music from 10 genres, which are grouped into 5 more general genres. There are both audio and MIDI files of the music. In addition, cultural meta data from the Internet (e.g., last.fm user annotations [75]) and song lyrics are stored. In this work, only the audio files, genres, and lyrics are relevant.

CDs data set:

The chair 11 of the Department of Computer Science, TU Dortmund, has its own music collection consisting of 120 albums [76], as well as a second collection, TAS120 [77], of 120 individual pieces of music, which are divided into 6 genres. For these pieces of music, album covers and lyrics had to be searched.

Appendix B. Reassignment of Music Genres

The reassignment of the music genres of the original data sets in the newly created data set is compiled below.

Original Genre	New Genre
Alternative Pop/Rock	-
Alternative and Punk	Rock
Alternative-Rock	Rock
Ambient	Electronic
Avant-Garde	Jazz
Big Band	Jazz
Big Beat	Electronic
Bluegrass	Folk, World, & Country
Blues-Contemporary Blues	Blues
Blues-Country Blues	Blues
Blues-Urban Blues	Blues
Classic	Classical
Classical-Classical	Classical
Country	Folk, World, & Country
Dance	Electronic
Dance Pop	Pop
Deutscher Rock Pop	-
Disco	Electronic
Easy Listening and Vocals	-
Electronic and Dance	Electronic
Electronica	Electronic
Eletronica	Electronic
Euro-Techno	Electronic
Folk	Folk, World, & Country
Funk/Soul	R&B
Grunge	Rock
Heavy Metal	Rock
Hip Hop/Rap	Rap/Hip-Hop
Hip Hop	Rap/Hip-Hop
Hip-Hop	Rap/Hip-Hop
House	Electronic
Humor	-
Indie	Rock
International	Folk, World, & Country
Jazz & Vocal	Jazz
Jazz-Acid Jazz	Jazz
Jazz-Bebop	Jazz
Jazz-Dixieland	Jazz
Jazz-Post-Bop	Jazz
Jazz-Soul Jazz	Jazz
Kölsch-Rock	Rock
Latin	-
Metal	Rock
Modern Folk-Alternative Folk	Folk, World, & Country
Modern Folk-Singer/Songwriter	Folk, World, & Country
Non-Music	-
Oldies	-
Other	-
Pop/Rock	-
PopRock	-

Original Genre	New Genre
Progressive Rock	Rock
R and B and Soul	R&B
R&B-Contemporary R&B	R&B
R&B-Contemporary R&B	R&B
R&B-Funk	R&B
R&B-Gospel	R&B
R&B-Rock & Roll	R&B
R&B-Soul	R&B
Rap	Rap/Hip-Hop
Rave	Electronic
Religious	-
RnB	R&B
Rock & Pop	-
Rock-Alternative Metal/Punk	Rock
Rock-Classic Rock	Rock
Rock-Metal	Rock
Rock-Roots Rock	Rock
Rock Pop	-
Rock and Pop	-
Soul	R&B
Soundtrack	-
Soundtracks and More	-
Stage & Screen	-
Symphonic Metal	Rock
Synthpop	Pop
Trance	Electronic
Trip-Hop	Rap/Hip-Hop
World-African	Folk, World, & Country
World-Calypso	Folk, World, & Country
World-Celtic	Folk, World, & Country
World-Chanson	Folk, World, & Country
World-Cuban	Folk, World, & Country
World-Fusion	Folk, World, & Country
World-Klezmer	Folk, World, & Country
World-U.S. Traditional	Folk, World, & Country
World	Folk, World, & Country

Appendix C. Audio Features

Appendix C.1. Audio Features of the TEMPO Feature Group. For Each Feature, the Average and Standard Deviation per Calculated Time Window Are Calculated

Feature	AMUSE-ID	Dim.	Reference
Duration	400	1	Theimer et al. [78]
Characteristics of fluctuation patterns	410	7	Theimer et al. [78]
Rhythmic clarity	418	1	Lartillot [79]
Estimated onset number per minute	420	1	Theimer et al. [78]
Estimated beat number per minute	421	1	Theimer et al. [78]
Estimated tatum number per minute	422	1	Theimer et al. [78]
Tempo based on onset times	425	1	Lartillot [79]
Five peaks of fluctuation curves summed across all bands	427	5	Lartillot [79]

Appendix C.2. Audio Features of the Feature Group TIMBRE. For Each Feature, the Average and Standard Deviation per Calculated Time Window Are Calculated

Domain	Feature	AMUSE-ID	Dim.	Reference
Time	Root mean square	4	1	Theimer et al. [78]
Time	Low energy	6	1	Theimer et al. [78]
Time	RMS peak number in 3 s	11	1	Lartillot [79]
Time	RMS peak number above mean amplitude in 3 s	12	1	Lartillot [79]
Frequency	Tristimulus	1	2	Theimer et al. [78]
Frequency	Spectral centroid	14	1	Theimer et al. [78]
Frequency	Spectral irregularity	15	1	Lartillot [79]
Frequency	Spectral bandwidth	16	1	Theimer et al. [78]
Frequency	Spectral skewness	17	1	Theimer et al. [78]
Frequency	Spectral kurtosis	18	1	Theimer et al. [78]
Frequency	Spectral crest factor	19	4	Theimer et al. [78]
Frequency	Spectral flatness measure	20	4	Theimer et al. [78]
Frequency	Spectral extent	21	1	Theimer et al. [78]
Frequency	Spectral flux	22	1	Theimer et al. [78]
Frequency	Sub-band energy ratio	25	4	Theimer et al. [78]
Frequency	Spectral slope	29	1	Theimer et al. [78]
Phase	Angles in phase domain	32	1	Theimer et al. [78]
Phase	Distances in phase domain	33	1	Theimer et al. [78]
Cepstral	Mel frequency cepstral coefficients (MIR-Toolbox-Implementation)	39	13	Theimer et al. [78]
Cepstral	Delta MFCCs (MIR-Toolbox-Implementation)	48	13	Lartillot [79]

Appendix C.3. Features of the HARMONY Feature Group. For Each Feature, the Average and Standard Deviation per Calculated Time Window Are Calculated

Feature	AMUSE-ID	Dim.	Reference
Fundamental frequency	200	1	Theimer et al. [78]
Inharmonicity	217	1	Lartillot [79]
Chroma Energy Normalized Statistics	218	12	Müller [80]
Chroma DCT-Reduced log Pitch	219	12	Müller and Ewert [81]
Local tuning (NNLS Implementation)	253	1	Mauch and Dixon [82]
Harmonic change (NNLS Implementation)	254	1	Mauch and Dixon [82]
Consonance (NNLS Implementation)	255	1	Mauch and Dixon [82]
Number of different chords	257	1	Vatolkin [43]
Number of chord changes	258	1	Vatolkin [43]
Shares of the most frequent 20, 40 and 60 percents of chords with regard to their duration	259	3	Vatolkin [43]
Key and its clarity 4096	10202	2	Lartillot [79]
Major/minor alignment 4096	10203	1	Lartillot [79]
Strengths of major keys 4096	10209	12	Lartillot [79]
Tonal centroid vector 4096	10216	6	Lartillot [79]
Harmonic change detection function 4096	10217	1	Lartillot [79]

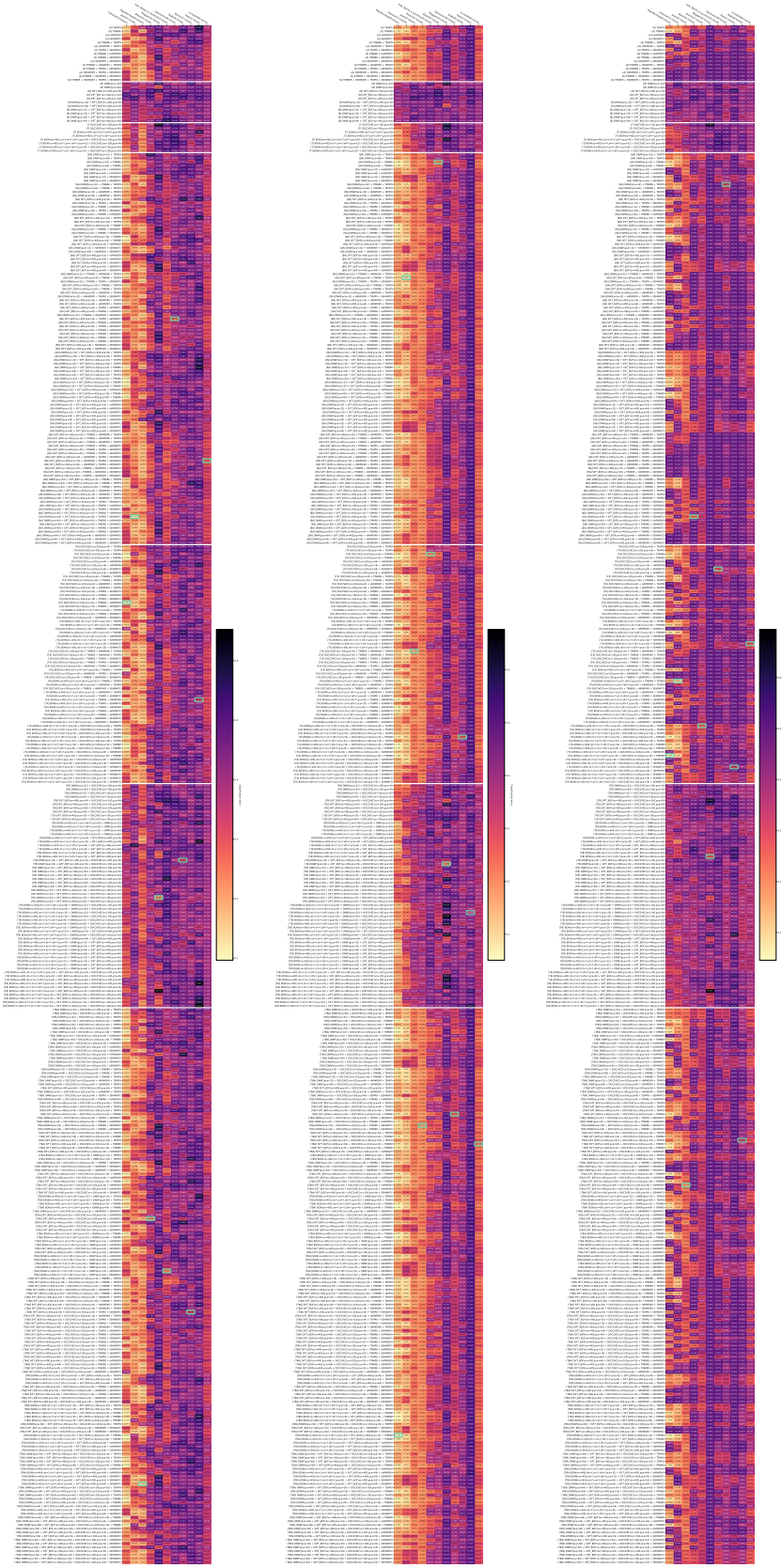
Appendix C.4. Audio Features of the SEMANTIC Feature Group. For Each Feature, the Average and Standard Deviation per Calculated Time Window Are Calculated

Feature	AMUSE-ID	Dim.	Reference
Guitar RF Chord-based	2001	1	Vatolkin [43]
Guitar SVM Chord-based	2003	1	Vatolkin [43]
Piano RF Chord-based	2021	1	Vatolkin [43]
Piano SVM Chord-based	2023	1	Vatolkin [43]
Wind RF Chord-based	2041	1	Vatolkin [43]
Wind SVM Chord-based	2043	1	Vatolkin [43]
Strings RF Chord-based	2061	1	Vatolkin [43]
Strings SVM Chord-based	2063	1	Vatolkin [43]
AMG mood Aggressive best RF model	4002	1	Vatolkin [43]
AMG mood Aggressive best SVM model	4006	1	Vatolkin [43]
AMG mood Energetic best RF model	4062	1	Vatolkin [43]
AMG mood Energetic best SVM model	4066	1	Vatolkin [43]
AMG mood Sentimental best RF model	4122	1	Vatolkin [43]
AMG mood Sentimental best SVM model	4126	1	Vatolkin [43]
AMG mood Stylish best RF model	4142	1	Vatolkin [43]
AMG mood Stylish best SVM model	4146	1	Vatolkin [43]
AMG mood Reflective best RF model	4102	1	Vatolkin [43]
AMG mood Reflective best SVM model	4106	1	Vatolkin [43]
AMG mood Confident best RF model	4022	1	Vatolkin [43]
AMG mood Confident best SVM model	4026	1	Vatolkin [43]
AMG mood Earnest best RF model	4042	1	Vatolkin [43]
AMG mood Earnest best SVM model	4046	1	Vatolkin [43]
AMG mood PartyCelebratory best RF model	4082	1	Vatolkin [43]
AMG mood PartyCelebratory best SVM model	4086	1	Vatolkin [43]
GFKL2011 Activation Level High best RF model	6002	1	Vatolkin [43]
GFKL2011 Activation Level High best SVM model	6006	1	Vatolkin [43]
GFKL2011 Effects Distortion best RF model	6022	1	Vatolkin [43]
GFKL2011 Effects Distortion best SVM model	6026	1	Vatolkin [43]
GFKL2011 Singing clear best RF model	6042	1	Vatolkin [43]
GFKL2011 Singing clear best SVM model	6046	1	Vatolkin [43]
GFKL2011 Singing Range middle best RF model	6062	1	Vatolkin [43]
GFKL2011 Melodic range $>$ octave best RF model	6242	1	Vatolkin [43]
GFKL2011 Melodic range $>$ octave best SVM model	6246	1	Vatolkin [43]
GFKL2011 Melodic range $\leq$ octave best RF model	6262	1	Vatolkin [43]
GFKL2011 Melodic range $\leq$ octave best SVM model	6266	1	Vatolkin [43]
GFKL2011 Melodic range linear best RF model	6282	1	Vatolkin [43]
GFKL2011 Melodic range linear best SVM model	6286	1	Vatolkin [43]
GFKL2011 Melodic range volatile best RF model	6302	1	Vatolkin [43]
GFKL2011 Melodic range volatile best SVM model	6306	1	Vatolkin [43]

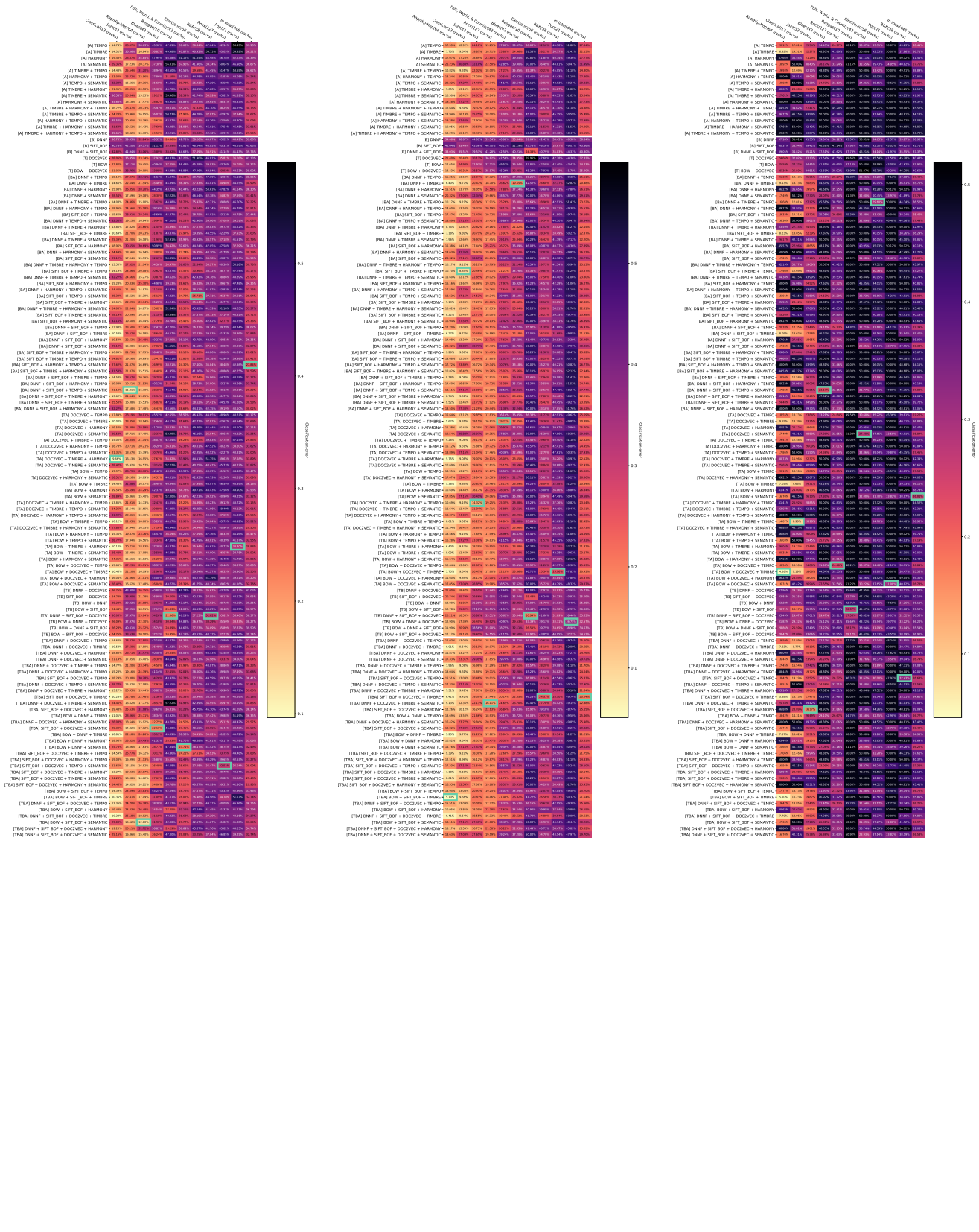
Appendix D. Visualizations of the Test Results

In the following, the results of the experiments from Section 4 of the paper are visualized. The methods for reducing the data volume of Section 4.2 of the paper are used to obtain the various representations. To see details, it is recommended to use the digital version of this work to enlarge the images.

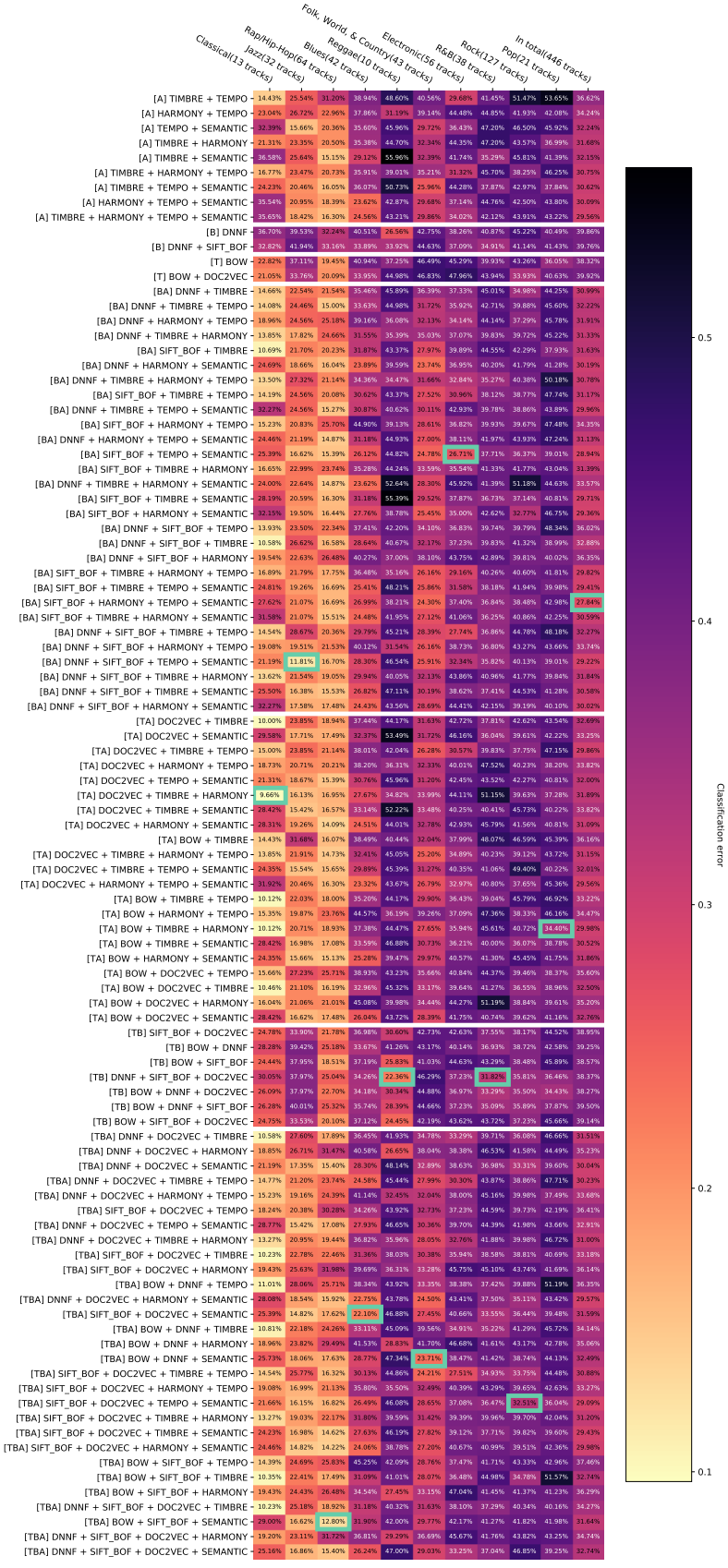
Appendix D.1. Error Rates of the Three Classifiers for the Different Combinations of Features



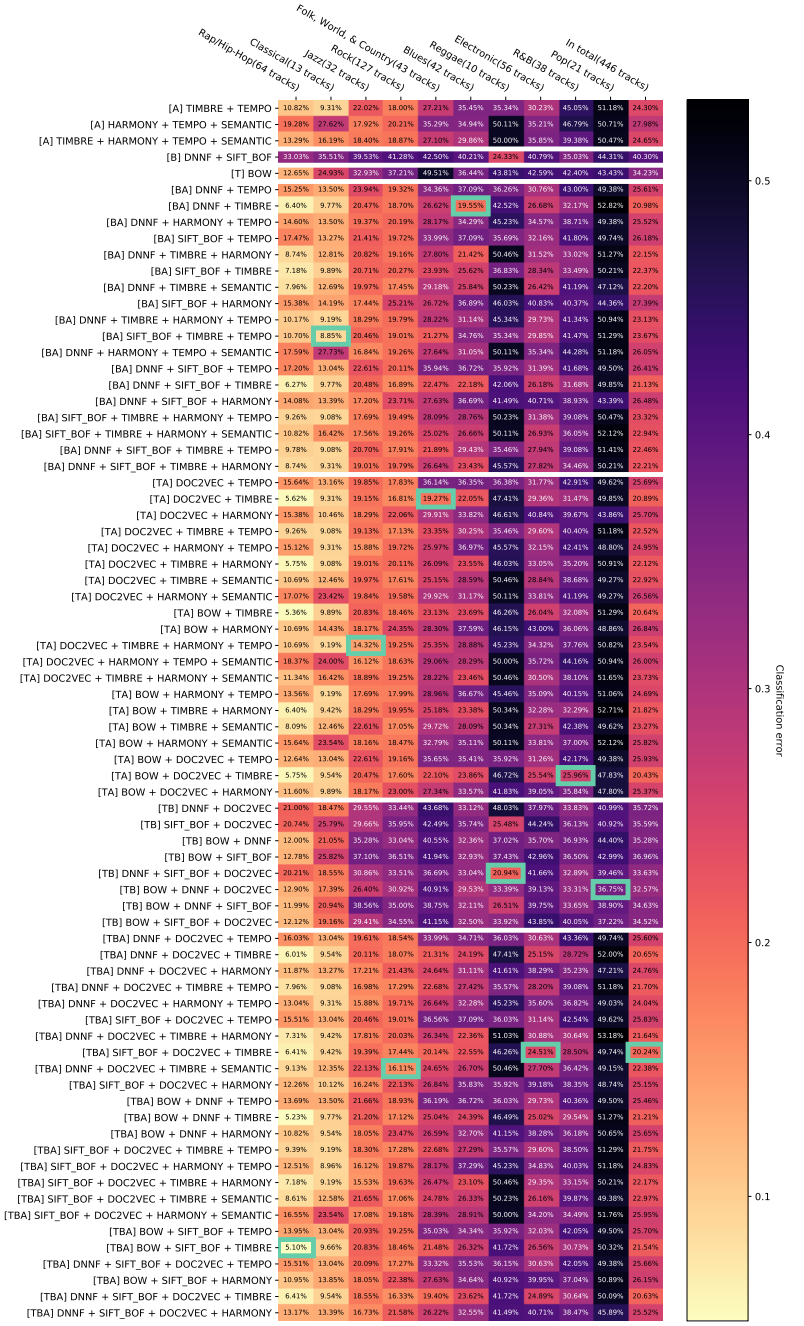
Appendix D.2. Error Rates of the Three Classifiers with Aggregated Feature Combinations



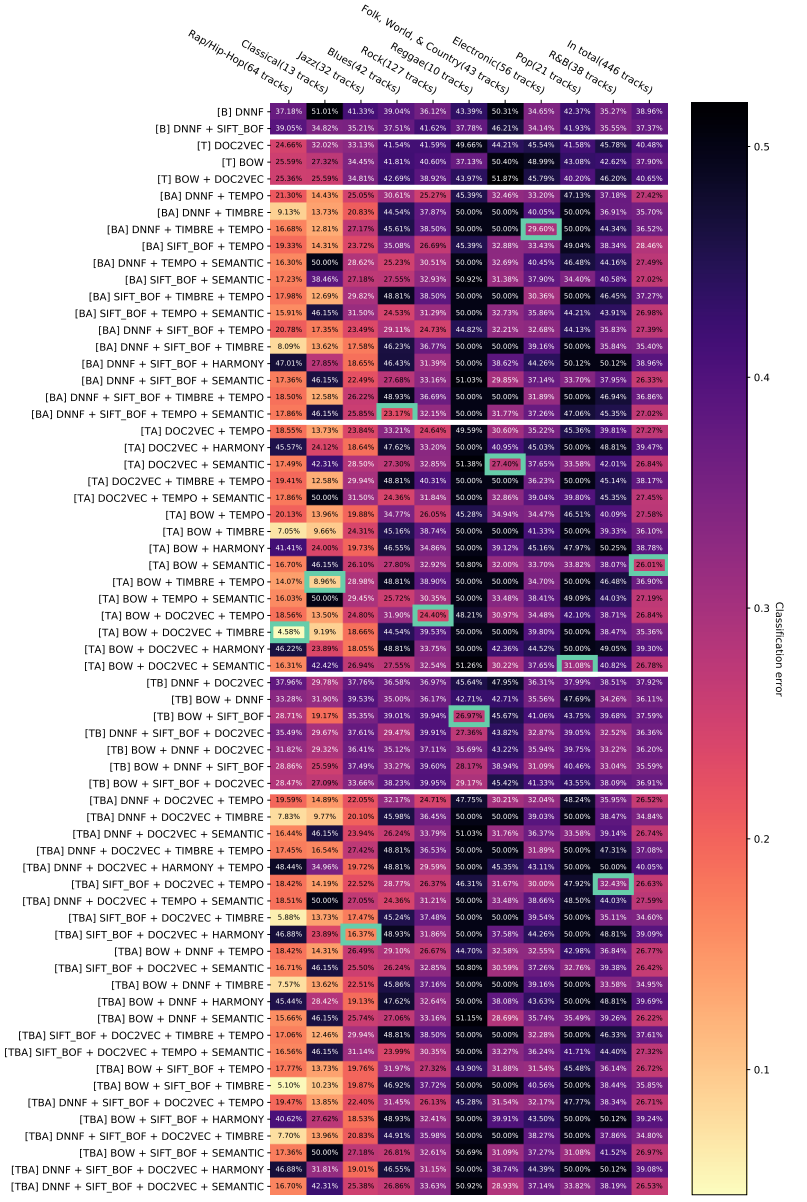
Appendix D.3. Error Rates for Aggregated Feature Combinations with Svm as a Classifier without Dominated Feature Combinations



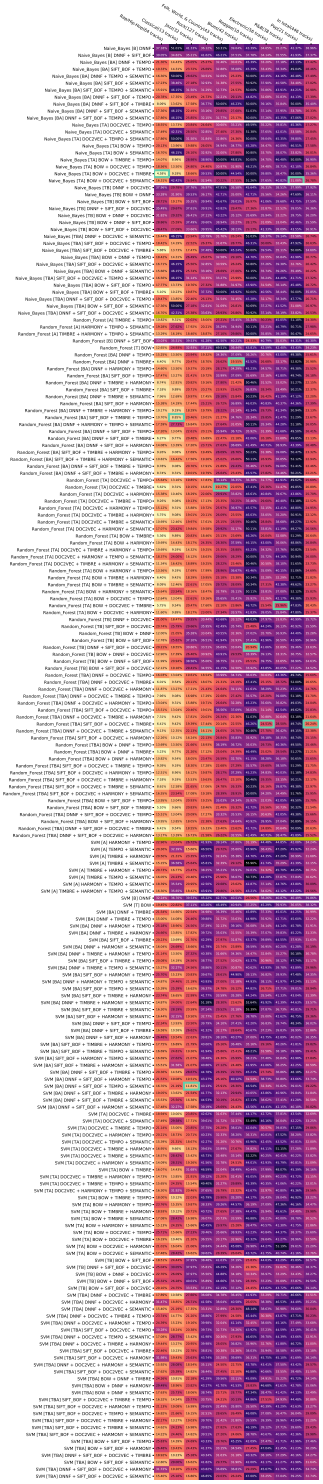
Appendix D.4. Error Rates for Aggregated Feature Combinations with the Random Forest Classifier without Dominated Feature Combinations



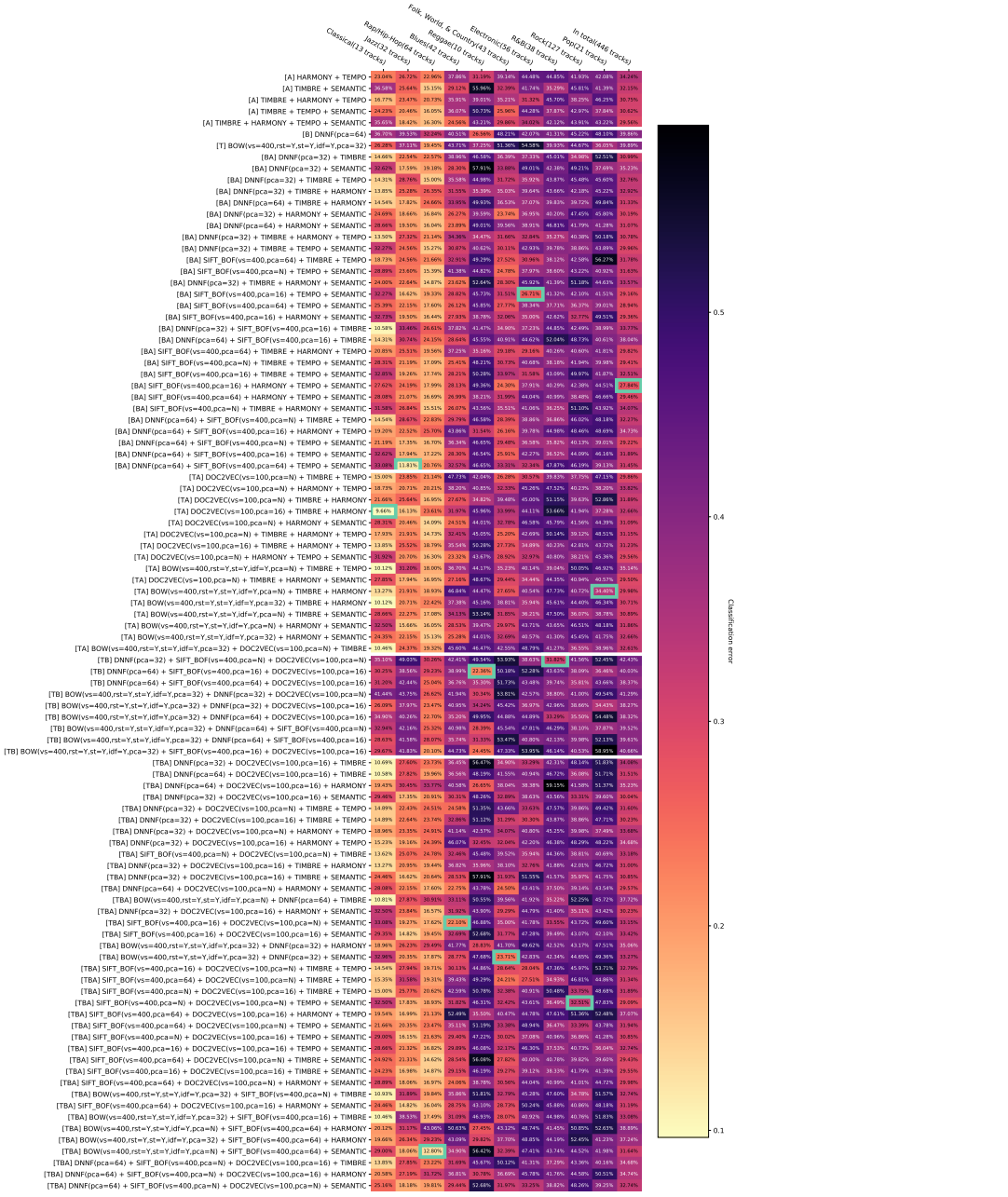
Appendix D.5. Error Rates for Aggregated Feature Combinations with the Naïve Bayes Classifier without Dominated Feature Combinations



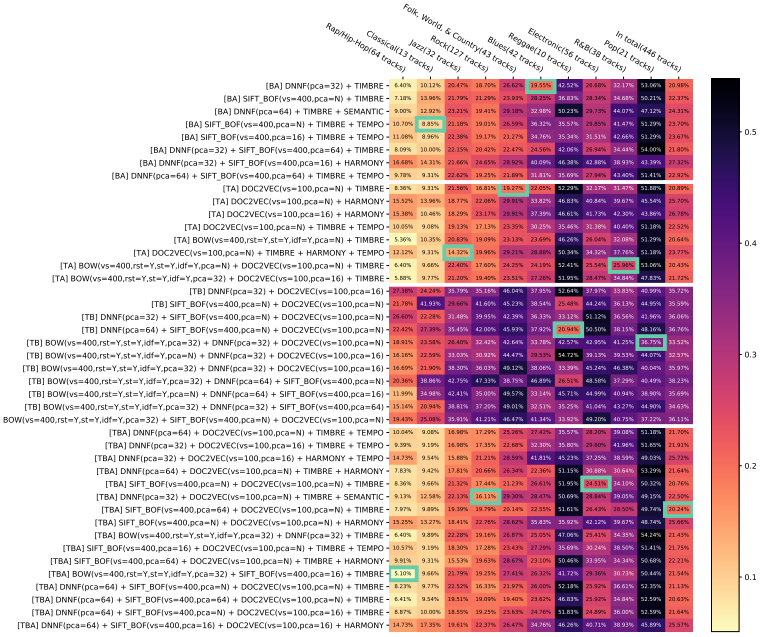
Appendix D.6. Error Rates of All Classifiers for Aggregated Feature Combinations without Dominated Feature Combinations



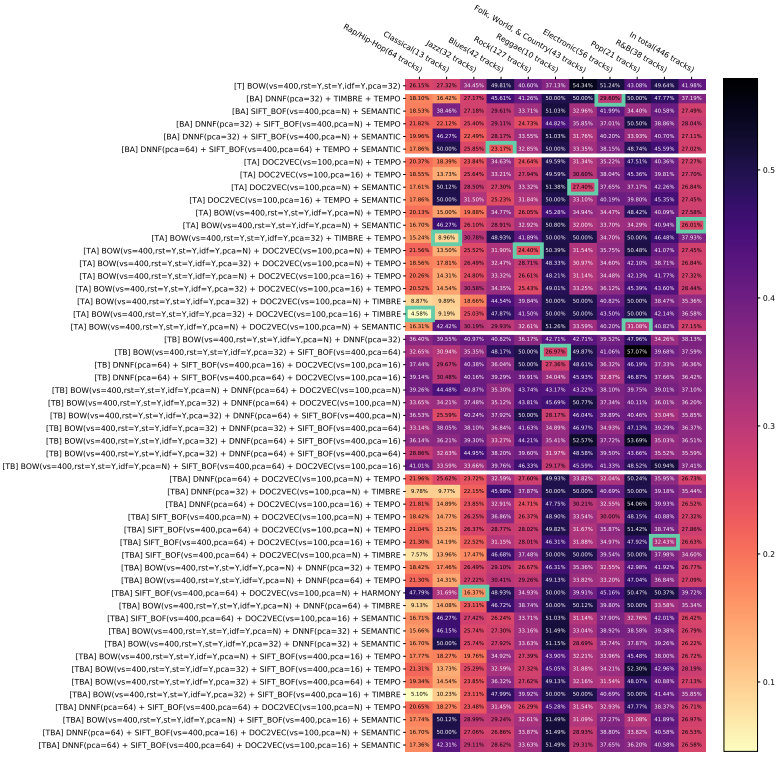
Appendix D.7. Error Rates for Aggregated Feature Combinations with Svm as a Classifier without Feature Combinations Contributing Less than $T = 0.01$ to the Dominated Hypervolume



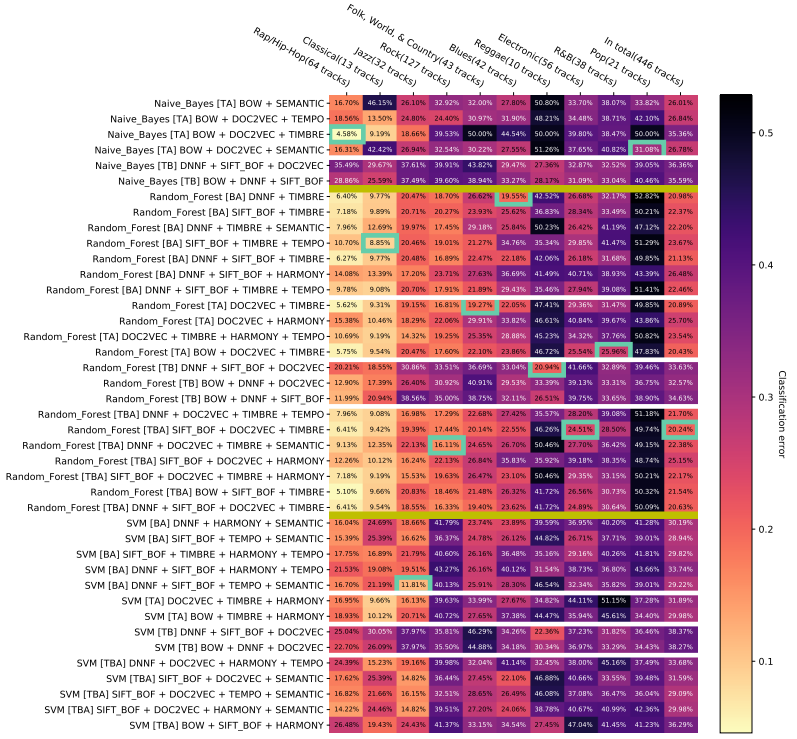
Appendix D.8. Error Rates for Aggregated Feature Combinations with the Random Forest Classifier without Feature Combinations Contributing Less than $T = 0.01$ to the Dominated Hypervolume



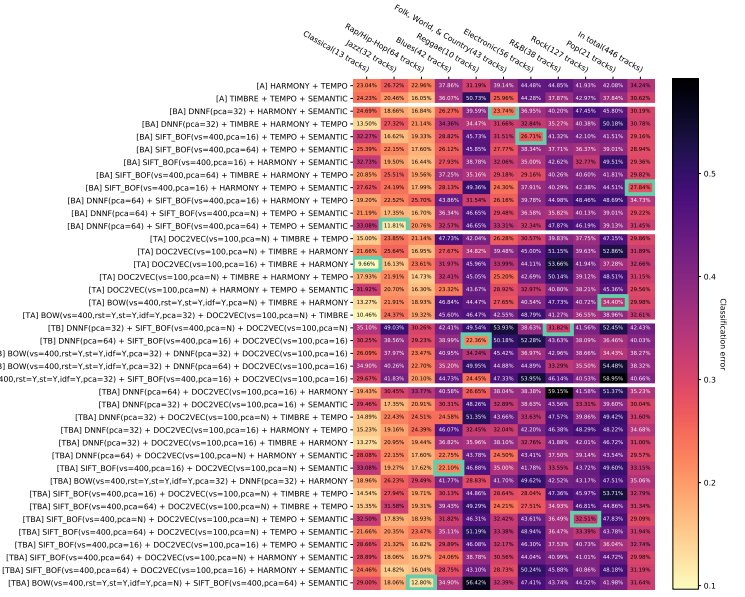
Appendix D.9. Error Rates for Aggregated Feature Combinations with the Na ïve Bayes Classifier without Feature Combinations Contributing Less than $T = 0.01$ to the Dominated Hypervolume



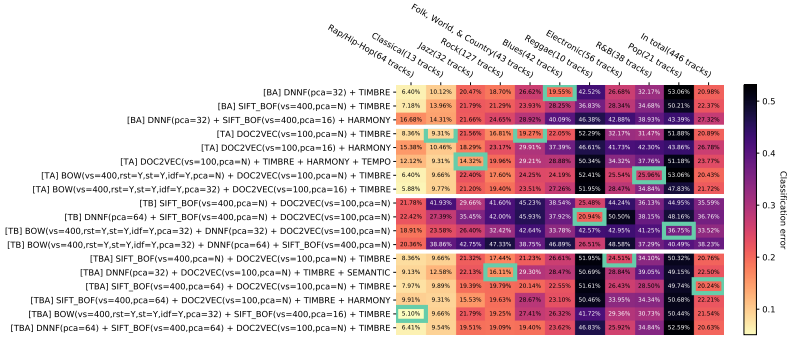
Appendix D.10. Error Rates of All Classifiers for Aggregated Feature Combinations without Feature Combinations Contributing Less than $T = 0.01$ to the Dominated Hypervolume



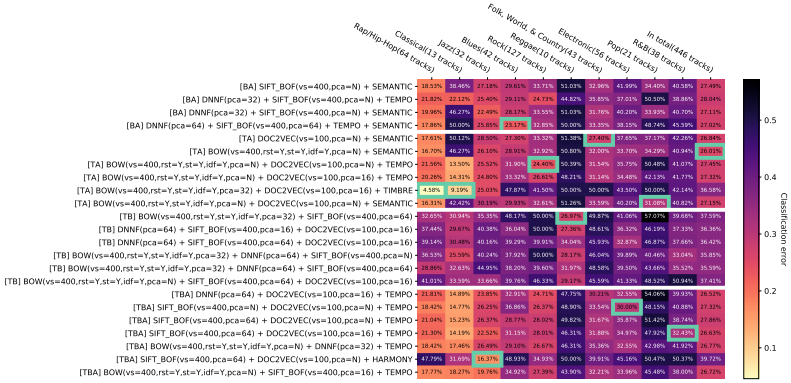
Appendix D.11. Error Rates for Aggregated Feature Combinations with Svm as a Classifier without Feature Combinations Contributing Less than $T = 0.05$ to the Dominated Hypervolume



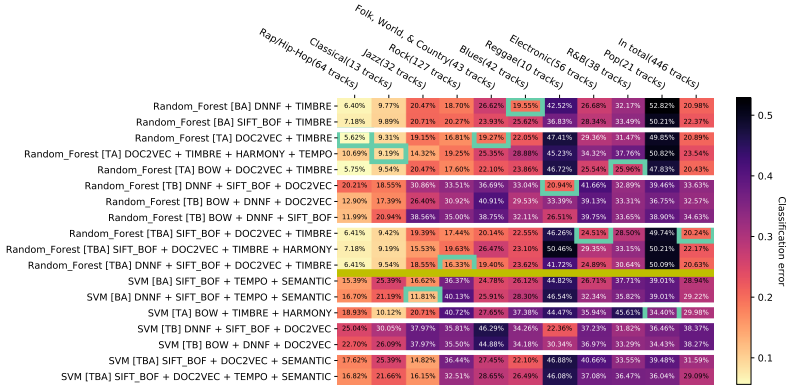
Appendix D.12. Error Rates for Aggregated Feature Combinations with the Random Forest Classifier without Feature Combinations Contributing Less than $T = 0.05$ to the Dominated Hypervolume



Appendix D.13. Error Rates for Aggregated Feature Combinations with the Na ïve Bayes Classifier without Feature Combinations Contributing Less than $T = 0.05$ to the Dominated Hypervolume



Appendix D.14. Error Rates of All Classifiers for Aggregated Feature Combinations without Feature Combinations Contributing Less than $T = 0.05$ to the Dominated Hypervolume



Appendix E. Test Results

In the following, the results of the statistical tests from Section 4.3 of the paper are listed.

Table A1. Results of the tests for hypothesis M_1 based on the error rates of the SVM classifier from Appendix D.2.

Test of Variables Line *A* against *B*	*p*-Value	Adapted *p*-Value	Median *A*	Median *B*	Null Line Hypothesis
Audio against image	0.062	0.124	0.368	0.399	retained
Audio against text	0.328	0.656	0.368	0.372	retained

Table A2. Results of the tests on hypothesis M_1 based on the error rates of the random forest classifier from Appendix D.2.

Test of Variables Line *A* against *B*	*p*-Value	Adapted *p*-Value	Median *A*	Median *B*	Null Line Hypothesis
Audio against image	0.010	0.020	0.250	0.390	rejected
Audio against text	0.006	0.012	0.250	0.368	rejected

Table A3. Results of the tests on hypothesis M_1 based on the error rates of the Naïve Bayesian classifier from Appendix D.2.

Test of Variables Line *A* against *B*	*p*-Value	Adapted *p*-Value	Median *A*	Median *B*	Null Line Hypothesis
Audio against image	0.041	0.082	0.282	0.380	retained
Audio against text	0.021	0.042	0.282	0.406	rejected

Table A4. Results of the tests on hypothesis $M_{2,1}$ based on the error rates of the SVM classifier from Appendix D.2.

Test of Variables Line *A* against *B*	*p*-Value	Adapted *p*-Value	Median *A*	Median *B*	Null Line Hypothesis
Audio against image + audio	0.037	0.220	0.296	0.267	retained
Audio against text + audio	0.248	1.000	0.296	0.295	retained
Image against image + audio	0.010	0.060	0.371	0.267	retained
Image against text + image	0.003	0.020	0.371	0.337	rejected
Text against text + audio	0.006	0.035	0.361	0.295	rejected
Text against text + image	0.075	0.452	0.360	0.337	retained

Table A5. Results of the tests on hypothesis $M_{2,1}$ based on the error rates of the random forest classifier from Appendix D.2.

Test of Variables Line *A* against *B*	*p*-Value	Adapted *p*-Value	Median *A*	Median *B*	Null Line Hypothesis
Audio against image + audio	0.005	0.030	0.250	0.210	rejected
Audio against text + audio	0.004	0.026	0.250	0.204	rejected
Image against image + audio	0.010	0.060	0.388	0.210	retained
Image against text + image	0.003	0.020	0.388	0.309	rejected
Text against text + audio	0.004	0.027	0.352	0.204	rejected
Text against text + image	0.003	0.020	0.352	0.309	rejected

Table A6. Results of the tests on hypothesis $M_{2,1}$ based on the error rates of the Naïve Bayesian classifier from Appendix D.2.

Test of Variables Line *A* against *B*	*p*-Value	Adapted *p*-Value	Median *A*	Median *B*	Null Line Hypothesis
Audio against image + audio	0.010	0.060	0.278	0.263	retained
Audio against text + audio	0.010	0.060	0.278	0.260	retained
Image against image + audio	0.010	0.060	0.372	0.263	retained
Image against text + image	0.004	0.026	0.372	0.325	rejected
Text against text + audio	0.006	0.035	0.389	0.260	rejected
Text against text + image	0.016	0.098	0.389	0.325	retained

Table A7. Results of the tests on hypothesis $M_{2,2}$ based on the error rates of the SVM classifier from Appendix D.2.

Test of Variables Line *A* against *B*	*p*-Value	Adapted *p*-Value	Median *A*	Median *B*	Null Line Hypothesis
Image + audio against text + image + audio	0.213	0.640	0.267	0.267	retained
Text + audio against text + image + audio	0.033	0.100	0.295	0.267	retained
Text + image against text + image + audio	0.021	0.0624	0.337	0.267	retained

Table A8. Results of the tests on hypothesis $M_{2,2}$ based on the error rates of the classifier Random Forest from Appendix D.2.

Test of Variables Line *A* against *B*	*p*-Value	Adapted *p*-Value	Median *A*	Median *B*	Null Line Hypothesis
Image + audio against text + image + audio	0.328	0.984	0.210	0.202	retained
Text + audio against Text + image + audio	0.534	1.000	0.204	0.202	retained
Text + image against text + image + Audio	0.091	0.273	0.310	0.202	retained

Table A9. Results of tests on hypothesis $M_{2,2}$ based on the error rates of the Naïve Bayesian classifier from Appendix D.2.

Test of Variables Line *A* against *B*	*p*-Value	Adapted *p*-Value	Median *A*	Median *B*	Null Line Hypothesis
Image + audio against text + image + audio	0.021	0.062	0.263	0.262	retained
Text + audio against text + image + audio	0.333	0.999	0.260	0.262	retained
Text + image against text + image + audio	0.033	0.099	0.325	0.262	retained

Table A10. Results of tests on hypothesis M_4 based on the error rates of the SVM classifier from Appendix D.2.

Test of Variables Line *A* against *B*	*p*-Value	Adapted *p*-Value	Median *A*	Median *B*	Null Line Hypothesis
[B] SIFT_BOF line (vs = 400, pca = N) against line [B] SIFT_BOF line (vs = 400, pca = 16)	0.213	0.853	0.480	0.501	retained
[B] SIFT_BOF line (vs = 400, pca = N) against line [B] SIFT_BOF line (vs = 400, pca = 64)	0.374	1.000	0.480	0.482	retained
[T] DOC2VEC line (vs = 100, pca = N) against line [T] DOC2VEC line (vs = 100, pca = 16)	0.155	0.619	0.413	0.408	retained
[T] BOW(. . . , pca = N) line against [T] BOW(. . . , pca = 32)	0.657	1.000	0.409	0.399	retained

Table A11. Results of the tests for hypothesis M_4 based on the error rates of the classifier Random Forest from Appendix D.2.

Test of Variables Line *A* against *B*	*p*-Value	Adapted *p*-Value	Median *A*	Median *B*	Null Line Hypothesis
[B] SIFT_BOF line (vs = 400, pca = N) against line [B] SIFT_BOF line (vs = 400, pca = 16)	0.131	0.523	0.461	0.474	retained
[B] SIFT_ line (vs = 400, pca = N) against line [B] SIFT_BOF line (vs = 400, pca = 64)	0.722	1.000	0.461	0.474	retained
[T] DOC2VEC line (vs = 100, pca = N) against line [T] DOC2VEC line (vs = 100, pca = 16)	0.575	1.000	0.393	0.397	retained
[T] BOW(. . . , pca = N) line against [T] BOW(. . . , pca = 32)	0.213	0.853	0.372	0.434	retained

Table A12. Results of the tests on hypothesis M_4 based on the error rates of the Naïve Bayesian classifier from Appendix D.2.

Test of Variables Line *A* against *B*	*p*-Value	Adapted *p*-Value	Median *A*	Median *B*	Null Line Hypothesis
[B] SIFT_BOF line (vs = 400, pca = N) against line [B] SIFT_BOF line (vs = 400, pca = 16)	0.286	1.000	0.432	0.435	retained
[B] SIFT_BOF line (vs = 400, pca = N) against line [B] SIFT_BOF line (vs = 400, pca = 64)	0.721	1.000	0.432	0.447	retained
[T] DOC2VEC line (vs = 100, pca = N) against line [T] DOC2VEC line (vs = 100, pca = 16)	0.062	0.248	0.458	0.416	retained
[T] BOW(. . . , pca = N) line against [T] BOW(. . . , pca = 32)	0.594	1.000	0.426	0.420	retained

References

1. Sturm, B.L. A Survey of Evaluation in Music Genre Recognition. In Proceedings of the 10th International Workshop on Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation (AMR), Copenhagen, Denmark, 24–25 October 2012; pp. 29–66.
2. Oramas, S.; Nieto, O.; Barbieri, F.; Serra, X. Multi-Label Music Genre Classification from Audio, Text and Images Using Deep Features. In Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR), Suzhou, China, 23–27 October 2017; pp. 23–30.
3. Oramas, S.; Barbieri, F.; Nieto, O.; Serra, X. Multimodal Deep Learning for Music Genre Classification. *Trans. Int. Soc. Music Inf. Retr.* **2018**, *1*, 4–21. [CrossRef]
4. Tzanetakis, G.; Cook, P. Musical Genre Classification of Audio Signals. *IEEE Trans. Speech Audio Process.* **2002**, *10*, 293–302. [CrossRef]
5. Lidy, T.; Rauber, A. Evaluation of Feature Extractors and Psycho-Acoustic Transformations for Music Genre Classification. In Proceedings of the 6th International Society for Music Information Retrieval Conference (ISMIR), Montreal, QC, Canada, 11–16 October 2005; pp. 34–41.
6. Scaringella, N.; Zoia, G.; Mlynek, D. Automatic Genre Classification of Music Content: A Survey. *IEEE Signal Process. Mag.* **2006**, *23*, 133–141. [CrossRef]
7. Bainbridge, D.; Bell, T. The Challenge of Optical Music Recognition. *Comput. Humanit.* **2001**, *35*, 95–121. [CrossRef]

8. Burgoyne, J.; Devaney, J.; Ouyang, Y.; Pugin, L.; Himmelman, T.; Fujinaga, I. Lyric Extraction and Recognition on Digital Images of Early Music Sources. In Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR), Kobe, Japan, 26–30 October 2009; pp. 723–728.
9. Ke, Y.; Hoiem, D.; Sukthankar, R. Computer Vision for Music Identification. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–26 June 2005; IEEE Computer Society: Washington, DC, USA, 2005; Volume 1, pp. 597–604.
10. Dorochowicz, A.; Kostek, B. Relationship between Album Cover Design and Music Genres. In Proceedings of the 2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), Poznan, Poland, 18–20 September 2019; pp. 93–98.
11. Le, V. Visual Metaphors on Album Covers: An Analysis into Graphic Design's Effectiveness at Conveying Music Genres. Bachelor's Thesis, Honors College, Oregon State University, Corvallis, OR, USA, 2020.
12. Schindler, A. Multi-Modal Music Information Retrieval: Augmenting Audio-Analysis with Visual Computing for Improved Music Video Analysis. Ph.D. Thesis, Faculty of Informatics, TU Wien, Hong Kong, China, 2019.
13. Libeks, J.; Turnbull, D. You Can Judge an Artist by an Album Cover: Using Images for Music Annotation. *IEEE Multimed.* **2011**, *18*, 30–37. [CrossRef]
14. Logan, B.; Kositsky, A.; Moreno, P. Semantic Analysis of Song Lyrics. In Proceedings of the 2004 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 27–30 June 2004; IEEE Computer Society: Washington, DC, USA, 2004; pp. 827–830.
15. Xia, Y.; Wang, L.; Wong, K. Sentiment Vector Space Model for Lyric-Based Song Sentiment Classification. *Int. J. Comput. Process. Lang.* **2008**, *21*, 309–330. [CrossRef]
16. Tsaptsinos, A. Lyrics-Based Music Genre Classification Using a Hierarchical Attention Network. In Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR), Suzhou, China, 23–27 October 2017; pp. 694–701.
17. Simonetta, F.; Ntalampiras, S.; Avanzini, F. Multimodal Music Information Processing and Retrieval: Survey and Future Challenges. In Proceedings of the 2019 International Workshop on Multilayer Music Representation and Processing (MMRP), Milano, Italy, 24–25 January 2019; pp. 10–18.
18. Neumayer, R.; Rauber, A. Integration of Text and Audio Features for Genre Classification in Music Information Retrieval. In Proceedings of the 29th European Conference on IR Research (ECIR), Rome, Italy, 2–5 April 2007; pp. 724–727.
19. Mayer, R.; Neumayer, R.; Rauber, A. Combination of Audio and Lyrics Features for Genre Classification in Digital Audio Collections. In Proceedings of the 16th ACM International Conference on Multimedia (MM), Vancouver, BC, Canada, 27–31 October 2008; pp. 159–168.
20. Mayer, R.; Rauber, A. Multimodal Aspects of Music Retrieval: Audio, Song Lyrics-and Beyond? In *Advances in Music Information Retrieval*; Ras, Z.W., Wieczorkowska, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 333–363.
21. Mayer, R.; Rauber, A. Music Genre Classification by Ensembles of Audio and Lyrics Features. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR), Miami, FL, USA, 24–28 October 2011; pp. 675–680.
22. Laurier, C.; Grivolla, J.; Herrera, P. Multimodal Music Mood Classification Using Audio and Lyrics. In Proceedings of the 7th International Conference on Machine Learning and Applications, San Diego, CA, USA, 11–13 December 2008; pp. 688–693.
23. Yang, D.; Lee, W.S. Music Emotion Identification from Lyrics. In Proceedings of the 11th IEEE International Symposium on Multimedia (ISM), San Diego, CA, USA, 14–16 December 2009; pp. 624–629.
24. Xiong, Y.; Su, F.; Wang, Q. Automatic Music Mood Classification by Learning Cross-Media Relevance between Audio and Lyrics. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 961–966.
25. Delbouys, R.; Hennequin, R.; Piccoli, F.; Royo-Letelier, J.; Moussallam, M. Music Mood Detection Based on Audio and Lyrics with Deep Neural Net. In Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR), Paris, France, 23–27 September 2018; pp. 370–375.
26. Suzuki, M.; Hosoya, T.; Ito, A.; Makino, S. Music Information Retrieval from a Singing Voice Using Lyrics and Melody Information. *EURASIP J. Appl. Signal Process.* **2007**, *2007*, 38727. [CrossRef]
27. Dhanaraj, R.; Logan, B. Automatic Prediction Of Hit Songs. In Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR), London, UK, 11–15 September 2005; pp. 488–491.
28. Zangerle, E.; Tschuggnall, M.; Wurzinger, S.; Specht, G. ALF-200k: Towards Extensive Multimodal Analyses of Music Tracks and Playlists. In *Advances in Information Retrieval*; Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2018; pp. 584–590.
29. Cataltepe, Z.; Yaslan, Y.; Sonmez, A. Music Genre Classification Using MIDI and Audio Features. *EURASIP J. Appl. Signal Process.* **2007**, *2007*, 36409. [CrossRef]
30. Velarde, G.; Chac'on, C.C.; Meredith, D.; Weyde, T.; Grachten, M. Convolution-based Classification of Audio and Symbolic Representations of Music. *J. New Music Res.* **2018**, *47*, 191–205. [CrossRef]
31. Dunker, P.; Nowak, S.; Begau, A.; Lanz, C. Content-based mood classification for photos and music: A generic multi-modal classification framework and evaluation approach. In Proceedings of the 1st ACM SIGMM International Conference on Multimedia Information Retrieval (MIR), Vancouver, BC, Canada, 30–31 October 2008; pp. 97–104.
32. McKay, C.; Burgoyne, J.A.; Hockman, J.; Smith, J.B.L.; Vigliensoni, G.; Fujinaga, I. Evaluating the Genre Classification Performance of Lyrical Features Relative to Audio, Symbolic and Cultural Features. In Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR), Utrecht, The Netherlands, 9–13 August 2010; pp. 213–218.

33. Panda, R.; Malheiro, R.; Rocha, B.; Oliveira, A.; Paiva, R.P. Multi-Modal Music Emotion Recognition: A New Dataset, Methodology and Comparative Analysis. In Proceedings of the 10th International Symposium on Computer Music Multidisciplinary Research (CMMR), Marseille, France, 15–18 October 2013.
34. Moore, A.F. Categorical Conventions in Music Discourse: Style and Genre. *Music Lett.* **2001**, *82*, 432–442. [CrossRef]
35. Pachet, F.; Cazaly, D. A taxonomy of musical genres. In Proceedings of the 6th International Conference on Content-Based Multimedia Information Access (RIAO), Paris, France, 12–14 Aprial; pp. 1238–1245.
36. Discogs. Available online: https://www.discogs.com (accessed on 30 October 2021).
37. MusicBrainz. Available online: https://musicbrainz.org (accessed on 30 October 2021).
38. MetroLyrics. Available online: https://en.wikipedia.org/wiki/MetroLyrics (accessed on 30 October 2021).
39. LyricWiki. Available online: https://de.wikipedia.org/wiki/LyricWiki (accessed on 30 October 2021).
40. CajunLyrics. Available online: http://www.cajunlyrics.com (accessed on 30 October 2021).
41. Lololyrics. Available online: https://www.lololyrics.com (accessed on 30 October 2021).
42. Apiseeds Lyrics. Available online: https://apiseeds.com/documentation/lyrics (accessed on 30 October 2021).
43. Vatolkin, I. Improving Supervised Music Classification by Means of Multi-Objective Evolutionary Feature Selection. Ph.D. Thesis, Department of Computer Science, TU Dortmund University, Dortmund, Germany, 2013.
44. Kamien, R. *Music: An Appreciation*; McGraw-Hill Education: New York, NY, USA, 2014.
45. Pampalk, E. Computational Models of Music Similarity and their Application in Music Information Retrieval. Ph.D. Thesis, Department of Computer Science, Vienna University of Technology, Vienna, Austria, 2006.
46. American National Standards Institute. *USA Standard Acoustical Terminology*; ANSI: New York, NY, USA, 1960.
47. Randel, D.M. *The Harvard Dictionary of Music*; Belknap Press: Cambridge, MA, USA, 2003.
48. Harris, Z.S. Distributional Structure. *WORD* **1954**, *10*, 146–162. [CrossRef]
49. Bramer, M. *Principles of Data Mining*; Undergraduate Topics in Computer Science; Springer: London, UK, 2013.
50. Le, Q.; Mikolov, T. Distributed Representations of Sentences and Documents. In Proceedings of the 31st International Conference on International Conference on Machine Learning (ICML), Beijing, China, 21–26 June 2014; Volume 32, pp. 1188–1196.
51. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, NV, USA, 5–10 December 2013; Curran Associates Inc.: Red Hook, NY, USA, 2013; pp. 3111–3119.
52. Skansi, S. *Introduction to Deep Learning-From Logical Calculus to Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2018.
53. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
54. Lloyd, S. Least Squares Quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137. [CrossRef]
55. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
56. Pearson, K. LIII. On Lines and Planes of Closest Fit to Systems of Points in Space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1901**, *2*, 559–572. [CrossRef]
57. Maron, M.E. Automatic Indexing: An Experimental Inquiry. *J. Assoc. Comput. Mach.* **1961**, *8*, 404–417. [CrossRef]
58. Qiang, G. An Effective Algorithm for Improving the Performance of Naive Bayes for Text Classification. In Proceedings of the 2nd International Conference on Computer Research and Development (ICCRD), Kuala Lumpur, Malaysia, 7–10 May 2010; pp. 699–701.
59. Vapnik, V.N.; Chervonenkis, A.Y. *Theory of Pattern Recognition*; USSR: Nauka, MA, USA, 1974.
60. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*; Cambridge University Press: Cambridge, UK, 2000.
61. Ho, T.K. Random Decision Forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition (ICDAR), Montreal, QC, Canada, 14–16 August 1995; pp. 278–282.
62. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Wiley: Wadsworth, OH, USA, 1984.
63. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2009.
64. Au, T. Random Forests, Decision Trees, and Categorical Predictors: The "Absent Levels" Problem. *J. Mach. Learn. Res.* **2018**, *19*, 1–30.
65. Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
66. Vatolkin, I.; Theimer, W.; Botteck, M. AMUSE (Advanced Music Explorer)—A Multitool framework for music data analysis. In Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR), Utrecht, The Netherlands, 9–13 August 2010; pp. 33–38.
67. Kohavi, R. A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), Montreal, QC, Canada, 20–25 August 1995; pp. 1137–1143.
68. Zitzler, E.; Knowles, J.; Thiele, L. Quality Assessment of Pareto Set Approximations. In *Multiobjective Optimization: Interactive and Evolutionary Approaches*; Branke, J., Deb, K., Miettinen, K., Słowiński, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2008; pp. 373–404.
69. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biom. Bull.* **1945**, *1*, 80–83. [CrossRef]

70. Weihs, C.; Jannach, D.; Vatolkin, I.; Rudolph, G. *Music Data Analysis: Foundations and Applications*; CRC Press: Boca Raton, FL, USA, 2016.
71. Choi, K.; Fazekas, G.; Sandler, M.B.; Cho, K. Transfer Learning for Music Classification and Regression Tasks. In Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR), Suzhou, China, 23–27 October 2017; pp. 141–149.
72. Seyerlehner, K.; Widmer, G.; Knees, P. Frame Level Audio Similarity-A Codebook Approach. In Proceedings of the 11th International Conference on Digital Audio Effects (DAFx), Espoo, Finland, 1–4 September 2008.
73. Soleymani, M.; Caro, M.N.; Schmidt, E.M.; Sha, C.Y.; Yang, Y.H. 1000 Songs for Emotional Analysis of Music. In Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia (CrowdMM), Barcelona, Spain, 21 October 2013; ACM: New York, NY, USA, 2013; pp. 1–6.
74. Smith, J.B.L.; Burgoyne, J.A.; Fujinaga, I.; Roure, D.D.; Downie, J.S. Design and Creation of a Large-Scale Database of Structural Annotations. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR), Miami, FL, USA, 24–28 October 2011; pp. 555–560.
75. Last.FM. Available online: https://www.last.fm. (accessed on 30 October 2021).
76. TU Dortmund, Department of Computer Science, Chair for Algorithm Engineering Music Collection. Available online: https://ls11-www.cs.tu-dortmund.de/rudolph/mi/albumlist (accessed on 30 October 2021).
77. TU Dortmund, Department of Computer Science, Chair for Algorithm Engineering Music Collection TAS 120. Available online: https://ls11-www.cs.tu-dortmund.de/rudolph/mi/tsai120 (accessed on 30 October 2021).
78. Theimer, W.; Vatolkin, I.; Eronen, A. *Definitions of Audio Features for Music Content Description*; Technical Report TR08-2-001; Department of Computer Science, TU Dortmund University: Dortmund, Germany, 2008.
79. Lartillot, O. MIRtoolbox 1.4 User's Manual. Technical report, Finnish Centre of Excellence in Interdisciplinary Music Research and Swiss Center for Affective Sciences. 2012. Available online: https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/materials/mirtoolbox/MIRtoolbox%20Users%20Guide%201.4/@@download/file/manual1.4.pdf (accessed on 30 October 2021).
80. Müller, M. *Information Retrieval for Music and Motion*; Springer: Berlin/Heidelberg, Germany, 2007.
81. Müller, M.; Ewert, S. Chroma Toolbox: Matlab Implementations for Extracting Variants of Chroma-Based Audio Features. In Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR), Miami, FL, USA, 24–28 October 2011; pp. 215–220.
82. Mauch, M.; Dixon, S. Approximate Note Transcription for the Improved Identification of Difficult Chords. In Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR), Utrecht, The Netherlands, 9–13 August 2010; pp. 135–140.