Musical Genre Recognition based on Deep Descriptors of Harmony, Instrumentation, and Segments

Igor Vatolkin[1 [0000−0002−9454−9402]], Mark Gotham[2 [0000−0003−0722−3074]], Néstor Nápoles López[3 [????????????]], and Fabian Ostermann[1 [0000−0002−8365−3634]]

[1] Department of Computer Science,
TU Dortmund University, Dortmund, Germany
{igor.vatolkin,fabian.ostermann}@tu-dortmund.de
[2] Department of Computer Science,
TU Dortmund University, Dortmund, Germany
mark.gotham@tu-dortmund.de
[3] McGill University, CIRMMT, Montréal, QC, Canada
nestor.napoleslopez@mail.mcgill.ca

Abstract. Deep learning has recently established itself as a cluster of methods of choice for almost all classification tasks in music information retrieval. However, despite very good classification performance, it sometimes brings disadvantages including long training times and higher energy costs, lower interpretability of classification models, or an increased risk of overfitting when applied to small training sets due to a very large number of trainable parameters. In this paper, we investigate the combination of both deep and shallow algorithms for recognition of musical genres using a transfer learning approach. We train deep classification models once to predict harmonic, instrumental, and segment properties from datasets with respective annotations. Their predictions for another dataset with annotated genres are used as features for shallow classification methods. They can be trained over and again for different categories, and are particularly useful when the training sets are small, in a real world scenario when listeners define various musical categories selecting only a few prototype tracks. The experiments show the potential of the proposed approach for genre recognition. In particular, when combined with evolutionary feature selection which identifies the most relevant deep feature dimensions, the classification errors became significantly lower in almost all cases, compared to a baseline based on MFCCs or results reported in the previous work.

Keywords: Musical genre recognition · Deep neural networks · Transfer learning · Interpretable features · Evolutionary feature selection.

1 Introduction

Many music classification tasks in the audio signal domain are nowadays solved with the help of deep neural networks. The price for achieving very high accuracy

of classification models is often a long training time, lower interpretability, and danger of overfitting due to huge number of trainable parameters. Traditional "shallow" methods built upon manually engineered features offer an alternative, although typically leading to a lower classification performance. However, in some scenarios, where a decision for a target class can be theoretically explained using some mid-level, semantic properties, there exists an opportunity to apply jointly deep and shallow classifiers trying to combine their advantages and reduce their individual drawbacks.

Recognition of musical genres or styles presents such a scenario. Genres, sub-genres, personal preferences, or mood-related tags are usually defined either by experts or listeners based on some more or less clear semantic properties, like the instrumentation, applied digital effects, details of harmonic structure, or characteristics of melodic lines. Then, it is possible to train deep neural networks only once for the prediction of these "mid-level" properties, and integrate a simple, fast, and potentially more interpretable classifier for the prediction of "high-level" target categories over and again. This procedure very well describes the situation where listeners define different personal categories selecting only a few tracks that either perfectly match or mismatch the target class.

In this work, we introduce a framework which integrates both deep and traditional classification models implementing a transfer learning approach. In the first step, deep models are trained with convolutional neural networks to predict harmonic, instrument, and segment statistics, using several annotated datasets. Then, these models are applied to extract predictions for another dataset of music pieces with annotated genres which serve as high-level musical categories. Based on these predictions, various statistics for time frames and complete tracks are saved as features. Finally, these features are used to train a shallow classification method (random forest or support vector machine) to predict genres. To estimate particularly useful dimensions for different genres, an evolutionary feature selection is additionally applied. The results show the high potential of the proposed framework. In combination with evolutionary feature selection, deep features achieve lower classification errors for all tested genres using both applied shallow classifiers, compared to the baseline using MFCCs, but also in almost all cases compared to results reported in the previous work.

The remainder of this paper is organized as follows. Section 2 presents some related work on deep learning and musical genre recognition. Section 3 summarizes deep semantic features used in our study. Section 4 describes the setup of our experiments. Section 5 discusses the results. In Section 6, the most relevant findings of the study are outlined and some ideas for future work are provided.

2 Related Work

Deep neural networks have been shown to be effective for many music information retrieval (MIR) classification tasks. Often, the architectures have been adopted from the image recognition domain [10,21]. In particular, convolutional neural networks (CNNs) [11] play an important role, with Mel frequency spectro-

grams as (image-like) 2D-input [5]. For example, the recognition of predominant instruments was addressed in [8] and segment recognition in [7].

Musical genre recognition is one of the most widely explored classification tasks in MIR [23]. However, it has some problematic issues, e.g., genre taxonomies may be very distinct [17], and frequently applied evaluation measures are not optimal [24]. Nevertheless, genres represent examples of high-level musical categories. Also for genre recognition, CNN-based approaches have been proposed and successfully applied [31].

Deep learning on small datasets still may be problematic because of too many trainable parameters, even when techniques like dropout layers or data augmentation may increase the robustness of classification models. Shallow classifiers usually are more suitable to this, if deep learners are not heavily customized [18]. The idea of combining deep and shallow classifiers for genre classification was presented in [27]. In the experimental study, however, only instrument statistics were integrated as deep features for genre recognition, as an alternative to low-level signal descriptors and an evolutionary based approximation of instrumental texture. In [16], deep features were also used for multi-modal genre recognition.

A more general concept to use mid-level predictions as features for the prediction of high-level categories can be found, e.g., in [1], where so called anchors were designed to measure similarities between music pieces, or in [29], where supervised models were trained to predict expert annotations like instrumentation, moods, or vocal characteristics.

Also, transfer learning has been applied to genre and tag prediction earlier [25,4]. It always shows high potential whilst introducing additional difficulties, for instance, when deep features are too specialized for the source task. Nevertheless, transfer learning is a highly promising direction for future machine learning based classification tasks [32].

3 Deep Semantic Features

Figure 1 illustrates a general overview of our approach. CNN models which predict harmonic properties, instrument, and segments are trained using four datasets with respective annotations. Then, their predictions are used to calculate deep semantic properties for complete tracks or shorter time frames (171 harmonic, 328 instrument, and 30 segment properties, 529 dimensions in total) from the 1517-artists dataset [20] with genre annotations. These features are either all combined to create traditional classification models for genre recognition, or the most relevant dimensions are identified using evolutionary multi-objective feature selection, which simultaneously minimizes the classification error and the number of selected feature dimensions.

3.1 Harmonic Properties

Recently, the modeling of complex tonal relationships has received attention from the MIR community, for example, with a surge of models for automatic Roman numeral analysis, which provide key-and-chord information simultaneously

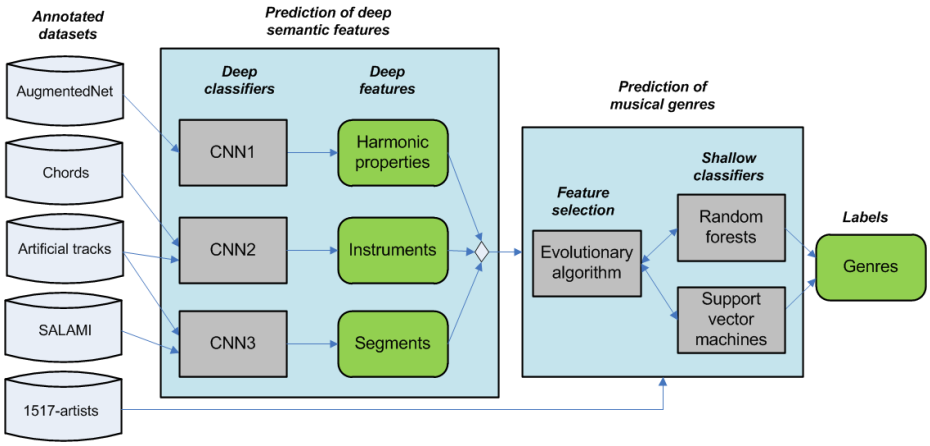

Fig. 1. Overview of data flow in the proposed classification framework.

[15,14,13]. Sometimes, the tonal information is computed concurrently [15,14] and sometimes in a modular fashion [6,13].

Although most of these comprehensive automatic chord recognition models have been trained on symbolic music files, the input representations often resemble the information contained in an audio chromagram extractor [12]. Using this to our advantage, we adapted a recent approach, AugmentedNet [15], to operate with audio chromagram features instead of symbolic ones. The AugmentedNet model provides multitask outputs related to the harmonic rhythm, chord, and key information of the music. More specifically, in the latest version of the network[4] these include one output predicting the segmentation of the chords (harmonic rhythm), two outputs predicting changes of key (local key and tonicized key), and six outputs predicting diverse aspects of the chords (Roman numeral class, pitch class set, and a four-note arrangement of the chord as a bass, tenor, alto, and soprano notes). The four-note arrangement of the chord is a prediction of each individual note in the chord, arranged in ascending order from the bass, modelling each of those notes as a separate output in the multitask layout.

Using this model, we extract mid-level harmonic features to be used for genre classification. Figure 2 shows the architecture of the AugmentedNet.

Table 1 summarizes 171 statistics after AugmentedNet predictions, which are used as deep features for genre recognition. All of them are calculated for time frames of 4s with 2s step size, which are later used as inputs / classification instances in genre recognition.

[4] https://github.com/napulen/AugmentedNet, accessed on 31.01.2023.

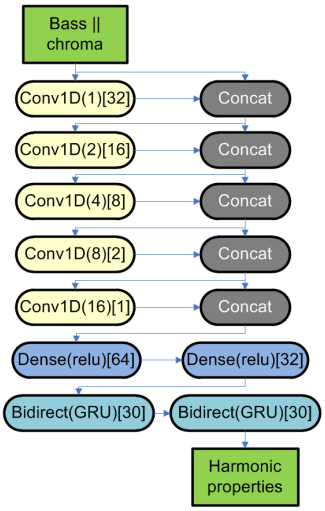

Fig. 2. Architecture of the AugmentedNet. The parameters (filter sizes for convolutional layers, activation functions, GRU layers) are provided in brackets. The numbers of neurons per layer are provided in squared brackets.

3.2 Instrument Predictions

For the prediction of instruments, two datasets were considered to train the neural networks. The first one consists of 5,000 samples and chords generated by mixing of individual instrument samples as described in [27]. 51 different instruments from within and beyond Western music contribute to these examples, many of them represented with several distinct instrument bodies. A CNN after [8] was trained with the Mel spectrograms to output a relative strength of each instrument in the mix (contribution to its overall energy). The second artificial audio multitracks (AAM) dataset[5] contains 3,000 artificially composed music tracks synthesized with real instrument samples using 31 instruments as a subset of instruments from [27]. Here, the Mel spectrograms were estimated only for 2s frames starting with annotated onsets. Figure 3 presents the architecture of the CNN for instrument recognition.

Table 2 summarizes 328 statistics after instrument predictions, calculated for time frames of 4s with 2s step size.

3.3 Segment Statistics

Classification models to predict musical segments are trained using the SALAMI dataset [22] and an artificial track dataset [28] using a CNN after [7]. While SALAMI boundary annotations do not contain additional information, the annotations of the artificial track dataset do list details of the changes in instrumental texture, tempo, and key between the musical segments, so that it was

[5] https://doi.org/10.5281/zenodo.5794629, accessed on 31.01.2023.

Table 1. Deep harmonic properties estimated for classification frames of 4s with 2s step size. The harmonic rhythm features are related to the chord segmentation; the bass, tenor, alto, and soprano features are predictions of each individual note in the chord, arranged in ascending order from the bass; similarly, the roman numeral feature is related to the specific class of Roman numeral of the chord; the local and tonicized key features are related with key predictions (e.g., modulations). Dim.: the number of all individual feature dimensions in the related feature group.

Features	**Dim.**
Predictions trained with AugmentedNet	
Mean and standard deviation of harmonic rhythm	1–2
Relative frequency of specific notes in the alto	3–24
Relative frequency of specific notes in the bass	25–47
Relative frequency of specific roots of local keys	48–71
Relative frequency of specific notes in the soprano	72–92
Relative frequency of specific notes in the tenor	93–112
Relative frequency of specific roots of tonicized keys	113–136
Relative frequency of specific roman numerals	137–160
Relative frequency of modes (major or minor)	161–162
Total number of different symbols	163–171

Table 2. Deep instrument features estimated for classification frames of 4s with 2s step size. Dim.: the number of all individual feature dimensions in the related feature group.

Features	**Dim.**
Predictions trained with chords	
Mean relative strength of 51 predicted instruments (acoustic and electric guitar, organ, piano and electric piano, viola, violin, etc.)	1–51
Standard deviation of the relative strength of 51 predicted instruments	52–102
Minimum relative strength of 51 predicted instruments	103–153
Maximum relative strength of 51 predicted instruments	154–204
Predictions trained with artificial tracks	
Mean relative strength of 31 predicted instruments (subset of 51 instruments)	205–235
Standard deviation of the relative strength of 31 predicted instruments	236–266
Minimum relative strength of 31 predicted instruments	267–297
Maximum relative strength of 31 predicted instruments	298–328

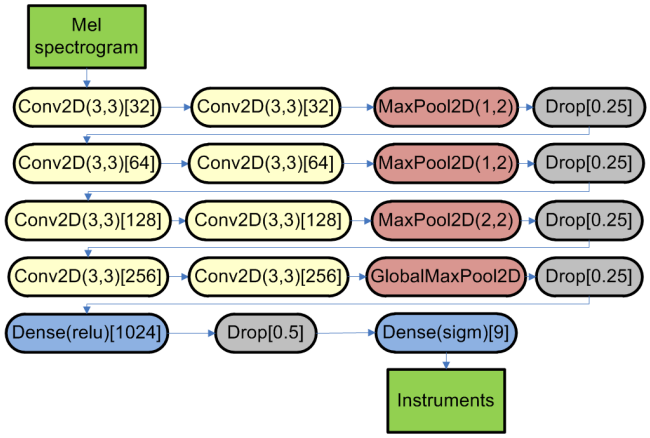

Fig. 3. Architecture of the CNN after [8]. The parameters (filter sizes for convolutional layers, activation functions) are provided in brackets. The numbers of neurons per layer are provided in squared brackets.

possible to train four different CNN models using annotations of either individual boundary types or all segment boundaries. Figure 4 presents an overview of the CNN architecture used for segment boundary prediction.

Table 3. Deep segment statistics estimated for complete audio tracks. Dim.: the number of all individual feature dimensions in the related feature group.

Features	**Dim.**
Predictions trained with SALAMI	
Number of segments	1
Mean segment length	2
Standard deviation of the segment length	3
Maximal segment length	4
Minimal segment length	5
Mean deviation of segment length	6
Predictions trained with artificial tracks	
Segment statistics as for SALAMI, trained to detect all boundaries	7–12
Segment statistics as for SALAMI, trained to detect instrument boundaries	13–18
Segment statistics as for SALAMI, trained to detect key boundaries	19–24
Segment statistics as for SALAMI, trained to detect tempo boundaries	25–30

Table 3 lists 30 segment statistics derived from predicted boundaries. They are estimated for complete audio tracks, and the same values are assigned to all 4s classification frames.

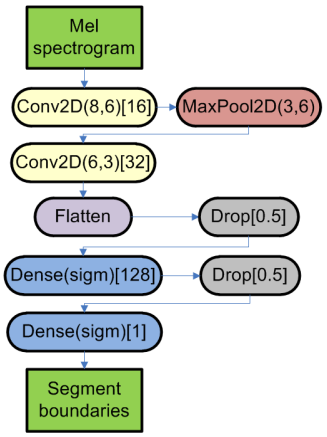

Fig. 4. Architecture of the CNN after [7]. The parameters (filter sizes for convolutional layers, activation functions) are provided in brackets. The numbers of neurons per layer are provided in squared brackets.

4 Setup of Experiments

For the recognition of musical genres, we use a publicly available dataset 1517-artists with 19 annotated genres [20].[6] Each binary genre classification task uses a *training set* of randomly selected 16 "positive" tracks from the selected genre and 18 "negative" tracks from all remaining genres (one track per genre). The number of tracks in the training set is explicitly selected to be rather low, as in the real-world situation, when a listener will try to avoid high efforts selecting many tracks to train an automatic music classification or recommendation system. For the evaluation of feature selection (see below), an *optimization set* of 228 tracks (12 per genre) is used. For the final independent evaluation of feature sets presented in the last iteration of evolutionary feature selection, a *test set* compiled from other 228 tracks (12 per genre) is taken into account. An artist filter was applied before the building of training, optimization, and test sets, so that all of them contain distinct tracks by different artists.

Classification models for genre prediction are trained with random forests [3] (with a default number of 100 trees) and linear support vector machines [30] using RapidMiner [9] integrated into the AMUSE framework [26]. While random forests train an ensemble of pruned decision trees, have only a few hyperparameters to setup, and are typically very robust to overfitting, support vector machines allow for linear separation between classes using transforms to more feature dimensions. Because of uneven distribution of positive and negative tracks in the test set in a binary classification scenario, the evaluation uses balanced relative error e_b which is defined as a mean error for positive and negative tracks:

[6] http://www.seyerlehner.info/joomla/index.php/datasets, accessed on 31.01.2023.

$$e_b = \frac{1}{2}\left(\frac{fn}{tp+fn} + \frac{fp}{tn+fp}\right), \tag{1}$$

where tp is the number of true positives (tracks correctly predicted as belonging to the genre), tn is the number of true negatives (tracks correctly predicted as not belonging to the genre), fp is the number of false positives (tracks wrongly predicted as belonging to the genre), and fn is the number of false negatives (tracks wrongly predicted as not belonging to the genre).

Because of the high overall number of 529 deep feature dimensions, some of them may be too irrelevant or noisy, depending on the particular genre to predict. Therefore, in an additional experiment, evolutionary multi-objective feature selection is applied to select the most relevant features after [29]. Evolutionary algorithms are a good choice here, as they explore a large number of different combinations of feature dimensions, applying a random mutation which selects or deselects individual dimensions to train the classifier. For the multi-objective optimization of two criteria, the minimization of e_b and the minimization of the number of selected feature dimensions, the $\mathcal{S}$ metric selection evolutionary multi-objective algorithm (SMS-EMOA) [2] was adopted. For each genre, 10 statistical repetitions are conducted, each based on 1,000 evolutionary generations, with a population size of 50 (number of feature sets under investigation). For further implementation details, we refer to [29].

5 Discussion of Results

Table 4 provides a summary of results with balanced relative errors for all 19 genres and different feature sets: individual deep feature groups, as well as their combination using all 519 dimensions, and results after evolutionary multi-objective feature selection. The top half of the table contains the test errors achieved by random forests, and the bottom half by support vector machines. As a baseline to compare with, we have also trained classifiers using a set of 13 Mel frequency cepstral coefficients (MFCCs) [19] which were originally developed for speech recognition, but have also been frequently used in music classification tasks. Further, we included in the table the best results from [27]; however, below we explain why it is hard to fairly compare them to the results of our study because of some differences in the experimental setup.

The results show that the performance of deep features varies very strongly and depends on the genre, and some errors are rather high (a random classifier would achieve an expected error of 0.5). However, please note that several challenges do exist in our application scenario. First, the training sets are quite small, as in real world situation where listeners wish to define a category without many attempts, and based on a limited number of selected prototype tracks. Second, the deep features are trained on other datasets, often with limitations. E.g., AugmentedNet is compiled only with classical music; artificial tracks used to train models for the prediction of instruments and segments are composed by an algorithm and thus do not perfectly represent commercial music. So, the

Table 4. Test e_b for 19 musical genre recognition tasks. [27]: the best results reported in that work (however, they are not completely comparable, see the text); MFCCs: Mel frequency cepstral coefficients; Harm: deep harmonic features listed in Table 1; Inst: deep instrument features listed in Table 2; Segm: deep segment features listed in Table 3; All: all deep features; All-FS: the best feature set after evolutionary feature selection. Bolded values are the best (smallest) for each genre in the current study. A bolded value using italic font marks a sole case where an error of [27] was lower than the lowest error in our study.

	Random forests						
Genre	**[27]**	**MFCCs**	**Harm**	**Inst**	**Segm**	**All**	**All-FS**
Alternative and Punk	0.1928	0.2847	0.4861	0.3148	0.4375	0.3218	0.2431
Blues	0.3170	0.4028	0.3727	0.3495	0.4954	0.4259	**0.1921**
Childrens	0.3880	0.5069	0.5116	0.4329	0.3148	0.3102	0.2685
Classical	0.0929	0.1250	0.5231	0.0995	0.2106	0.2083	0.0833
Comedy and Spoken Word	0.2214	0.3333	0.3634	0.3125	0.2894	0.3125	0.2407
Country	0.2350	0.3472	0.4190	0.2199	0.3843	0.3403	**0.1273**
Easy Listening	0.2904	0.2894	0.4537	0.3542	0.3542	0.3866	**0.2245**
Electronic+	0.1487	0.3843	0.2731	0.0926	0.3472	0.2454	**0.0370**
Folk	0.2682	0.3935	0.4236	0.3449	0.5440	0.3264	0.1852
Hip-Hop	0.1240	0.3495	0.4954	0.1065	0.2477	0.2824	0.0880
Jazz	0.3123	0.3889	0.3681	0.3519	0.5231	0.4514	0.2523
Latin	0.3049	0.5069	0.5694	0.4028	0.5231	0.3704	0.2940
New Age	0.2349	0.3056	0.5139	0.2731	0.3773	0.3750	**0.1505**
R'n'B and Soul	0.2534	0.2731	0.4144	0.2500	0.4213	0.2616	**0.1898**
Reggae	0.1941	0.3194	0.5069	0.2454	0.4375	0.3912	0.1875
Religious	0.3759	0.4352	0.3634	0.3912	0.5093	0.3611	0.2523
Rock and Pop	0.2346	0.2870	0.5579	0.2894	0.6273	0.2963	**0.1343**
Soundtracks and More	0.2652	0.2708	0.5926	0.3079	0.4190	0.3750	0.2616
World	0.4059	0.3403	0.4144	0.5069	0.5046	0.4745	**0.2662**
	Support vector machines						
Alternative and Punk	***0.1656***	0.2593	0.4282	0.2546	0.5000	0.2639	**0.2060**
Blues	0.3030	0.4074	0.3449	0.2546	0.5000	0.2847	0.2153
Childrens	0.3366	0.5185	0.5162	0.4769	0.5000	0.5000	**0.1944**
Classical	0.0885	0.0903	0.4190	**0.0810**	0.5000	0.1574	0.0833
Comedy and Spoken Word	0.2360	0.3542	0.3519	0.3426	0.5000	0.2431	**0.1782**
Country	0.2247	0.3565	0.4352	0.2407	0.5000	0.2940	0.1319
Easy Listening	0.2980	0.2315	0.4514	0.4259	0.5000	0.5000	0.2477
Electronic+	0.1448	0.2245	0.3380	0.1412	0.5000	0.1806	0.0532
Folk	0.2621	0.3449	0.4190	0.3495	0.5000	0.4167	**0.1736**
Hip-Hop	0.1201	0.2431	0.5185	0.0671	0.5000	0.5000	**0.0810**
Jazz	0.2680	0.4190	0.2708	0.3588	0.5000	0.3356	**0.2338**
Latin	0.3168	0.4514	0.5509	0.4838	0.5000	0.5602	**0.2593**
New Age	0.2122	0.2685	0.4745	0.2894	0.5000	0.5000	0.1921
R'n'B and Soul	0.2594	0.3380	0.4514	0.2593	0.5000	0.4236	0.2014
Reggae	0.1872	0.2546	0.5301	0.2523	0.5000	0.3449	**0.1690**
Religious	0.3751	0.3981	0.4005	0.3935	0.5000	0.3912	**0.2269**
Rock and Pop	0.2390	0.2014	0.5648	0.2917	0.5000	0.4120	0.1389
Soundtracks and More	0.3108	0.2593	0.5231	0.3773	0.5000	0.3403	**0.2431**
World	0.3604	0.3472	0.4954	0.4306	0.5000	0.3495	0.2731

robustness of deep features is obviously sometimes limited, when they are calculated for other data. Another difficulty is that some genres like Childrens, Rock and Pop, or World are very ambiguously defined. Still, for 14 of 19 categories using random forests and 12 of 19 categories using support vector machines, deep features or their combination are better than models trained with MFCCs. Sometimes, integration of all deep features leads to larger errors compared to the errors of the individual groups, underlining the suggestion that simply using more features is not always the best solution (the curse of dimensionality).

When all features from individual feature groups are compared, instrument statistics seem to be the most relevant; the errors achieved with this group are smaller than from all other groups for 14 of 19 categories using random forests and 18 of 19 using support vector machines. With random forests, harmonic predictions are the best for genres Religious and World, and segment predictions for Childrens and Spoken Word. Segment descriptors did not work sufficiently with support vector machines which predict only one class in all cases. A potential explanation is that the number of dimensions was in that case too low for this classifier.

When all groups are combined (column "All"), the errors are lower than for individual feature groups only for 4 genres for both classifiers. This strengthens the suggestion that too many dimensions are irrelevant for a particular genre, and feature selection may help to identify the most relevant ones. This is confirmed by values in the column "All-FS", which report the smallest test errors achieved after evolutionary feature selection as described in Section 4. Only for Classical, the error achieved with deep instrument features is smaller than after feature selection. Even if this seems to be peculiar at the first glance, the explanation is that feature selection is strictly evaluated using an independent test set for a better measurement of its general performance, but the evaluation of feature sets during the optimization process is done on the optimization set. So, theoretically, if a very informative feature dimension for the test data would not belong to the very best dimensions for the optimization data, it will not contribute to the final output of the algorithm.

Finally, we may compare the results to the best reported errors of the previous work [27]. They show that in all but one case the smaller errors are achieved using our large deep feature set. However, it is important to mention that the comparison is not very fair, because in our study the application of feature selection helped to identify the most relevant feature dimensions, and theoretically some individual timbral or approximative descriptors from [27] could contribute to final feature sets. Further, despite of the same genre labels and similar setup of classifiers, the distribution of data was not the same: in [27], larger training sets of 72 tracks were used, and in our experiments we had to reserve enough artist-independent tracks for a separate optimization set for the evaluation of feature selection.

6 Conclusions

In this paper, we have applied a transfer learning approach, first training several deep convolutional neural networks to predict harmonic, instrumental, and segment characteristics, and then using two traditional shallow classifiers predicting genres based on those deep features. Such a combination of deep and traditional methods can save a lot of resources, as the extensive training of all deep models is done only once, but the prediction of semantic musical properties only one time per track, and the training and classification of genre recognition, which may be repeated over and again for different listeners and application scenarios, can be conducted significantly faster based on classifiers with only a few parameters.

The results showed that deep features are quite successful and contribute to relatively low errors in a challenging application scenario, where the sizes of training sets are very small, as in the typical real-world situation. However, there still exist a large number of noisy and irrelevant dimensions—partly because transfer learning may not always allow for the extraction of very robust characteristics—but also simply because different classification categories have very distinct properties. Thus, the application of feature selection which itself requires some costs, seems to be essential and complements the classification algorithm pipeline, leading to smaller classification errors for all 19 genres compared to MFCC baseline and all deep features, for 18 genres compared to individual deep semantic feature groups, and for 18 genres compared to errors reported in the previous work.

In future work, we plan to integrate more deep features, but also to improve their robustness. A first promising direction is to include more diverse genres into datasets involved to train deep models for prediction of semantic music properties. Additionally, more data augmentation methods can be applied to the annotated data for the training of deep features. Another relevant contribution to justify the application of deep learning would be a strict statistical comparison of "deep" and "shallow" features for the recognition of genres. However, such a comparison is not always straightforward: in preliminary experiments, we have already extracted shallow instrument and segment features using random forests and the same training data as for deep neural networks, but the performance was very poor, potentially, because the Mel spectrograms could not provide enough information for the training of this classification method. Another promising implementation will be to integrate more harmonic analysis features from AugmentedNet, which as a completely novel approach proved solid and successful for genre recognition.

Acknowledgements

The authors gratefully acknowledge the computing time provided on the Linux HPC cluster at Technical University Dortmund (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 271512359.

References

1. Berenzweig, A., Ellis, D.P.W., Lawrence, S.: Anchor space for classification and similarity measurement of music. In: Proceedings of the IEEE International Conference on Multimedia and Expo, ICME. pp. 29–32. IEEE Computer Society (2003)
2. Beume, N., Naujoks, B., Emmerich, M.T.M.: SMS-EMOA: Multiobjective selection based on dominated hypervolume. European Journal of Operational Research **181**(3), 1653–1669 (2007)
3. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
4. Choi, K., Fazekas, G., Sandler, M., Cho, K.: Transfer learning for music classification and regression tasks. In: Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR. pp. 141–149. International Society for Music Information Retrieval (2017)
5. Costa, Y.M., Oliveira, L.S., Silla, C.N.: An evaluation of convolutional neural networks for music classification using spectrograms. Applied Soft Computing **52**, 28–38 (2017)
6. Gotham, M., Kleinertz, R., Weiss, C., Müller, M., Klauk, S.: What if the 'when' implies the 'what'?: Human harmonic analysis datasets clarify the relative role of the separate steps in automatic tonal analysis. In: Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR. pp. 229–236 (2021)
7. Grill, T., Schlüter, J.: Music boundary detection using neural networks on combined features and two-level annotations. In: Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR. pp. 531–537 (2015)
8. Han, Y., Kim, J., Lee, K.: Deep convolutional neural networks for predominant instrument recognition in polyphonic music. IEEE ACM Transactions on Audio, Speech, and Language Processing **25**(1), 208–221 (2017)
9. Hofmann, M., Klinkenberg, R.: RapidMiner: Data Mining Use Cases and Business Analytics Applications. Chapman & Hall/CRC (2013)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of the 26th Annual Conference on Neural Information Processing Systems, NIPS. pp. 1106–1114 (2012)
11. LeCun, Y., Bengio, Y., Hinton, G.E.: Deep learning. Nature **521**(7553), 436–444 (2015)
12. 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. pp. 135–140 (2010)
13. McLeod, A., Rohrmeier, M.A.: A modular system for the harmonic analysis of musical scores using a large vocabulary. In: Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR. pp. 435–442 (2021)
14. Micchi, G., Kosta, K., Medeot, G., Chanquion, P.: A deep learning method for enforcing coherence in automatic chord recognition. In: Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR. pp. 443–451 (2021)
15. Nápoles López, N., Gotham, M., Fujinaga, I.: AugmentedNet: A roman numeral analysis network with synthetic training examples and additional tonal tasks. In: Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR. pp. 404–411 (2021)

16. 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. pp. 23–30 (2017)
17. Pachet, F., Cazaly, D.: A taxonomy of musical genres. In: Proceedings of the 6th International Conference on Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications), RIAO. pp. 1238–1245. CID (2000)
18. Pasupa, K., Sunhem, W.: A comparison between shallow and deep architecture classifiers on small dataset. In: Proceedings of the 8th International Conference on Information Technology and Electrical Engineering, ICITEE. pp. 1–6. IEEE (2016)
19. Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. Prentice Hall, Upper Saddle River (1993)
20. 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 (2008)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 3rd International Conference on Learning Representations, ICLR (2015)
22. Smith, J.B.L., Burgoyne, J.A., Fujinaga, I., De Roure, 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. pp. 555–560. University of Miami (2011)
23. 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. pp. 29–66 (2012)
24. Sturm, B.L.: Classification accuracy is not enough - on the evaluation of music genre recognition systems. Journal of Intelligent Information Systems **41**(3), 371–406 (2013)
25. van den Oord, A. and Dieleman, S. and Schrauwen, B.: Transfer learning by supervised pre-training for audio-based music classification. In: Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR. pp. 29–34 (2014)
26. Vatolkin, I., Ginsel, P., Rudolph, G.: Advancements in the music information retrieval framework AMUSE over the last decade. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR. pp. 2383–2389. ACM (2021)
27. Vatolkin, I., Adrian, B., Kuzmic, J.: A fusion of deep and shallow learning to predict genres based on instrument and timbre features. In: Proceedings of the 10th International Conference on Artificial Intelligence in Music, Sound, Art and Design, EvoMUSART. pp. 313–326. Springer (2021)
28. Vatolkin, I., Ostermann, F., Müller, M.: An evolutionary multi-objective feature selection approach for detecting music segment boundaries of specific types. In: Proceedings of the 2021 Genetic and Evolutionary Computation Conference, GECCO. pp. 1061–1069 (2021)
29. Vatolkin, I., Rudolph, G., Weihs, C.: Evaluation of album effect for feature selection in music genre recognition. In: Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR. pp. 169–175 (2015)
30. Yu, H., Kim, S.: SVM tutorial - classification, regression and ranking. In: Rozenberg, G., Bäck, T., Kok, J.N. (eds.) Handbook of Natural Computing, Volume 1, pp. 479–506. Springer, Berlin Heidelberg (2012)

31. Zhang, W., Lei, W., Xu, X., Xing, X.: Improved music genre classification with convolutional neural networks. In: Proceedings of the 17th Annual Conference of the International Speech Communication Association, Interspeech. pp. 3304–3308. ISCA (2016)
32. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A comprehensive survey on transfer learning. Proceedings of the IEEE **109**(1), 43–76 (2021)