Music genre classification: looking for the Perfect Network*

Daniel Kostrzewa[0000−0003−2781−3709], Piotr Kaminski, and Robert Brzeski[0000−0001−7127−0989]

Department of Applied Informatics, Silesian University of Technology, Gliwice, Poland {daniel.kostrzewa,robert.brzeski}@polsl.pl

Abstract. This paper presents research on music genre recognition. It is a crucial task because there are millions of songs in the online databases. Classifying them by a human being is impossible or extremely expensive. As a result, it is desirable to create methods that can assign a given track to a music genre. Here, the classification of music tracks is carried out by deep learning models. The Free Music Archive dataset was used to perform experiments. The tests were executed with the usage of Convolutional Neural Network, Convolutional Recurrent Neural Networks with 1D and 2D convolutions, and Recurrent Neural Network with Long Short-Term Memory cells. In order to combine the advantages of different deep neural network architectures, a few types of ensembles were proposed with two types of results mixing methods. The best results obtained in this paper, which are equal to state-of-the-art methods, were achieved by one of the proposed ensembles. The solution described in the paper can help to make the auto-tagging of songs much faster and more accurate in the context of assigning them to particular musical genres.

Keywords: Music Information Retrieval · Music Genre Recognition · Classification · Deep Learning · Convolutional Neural Network · Recurrent Neural Networks · Long Short-Term Memory · Ensemble · Free Music Archive Dataset

1 Introduction

One of the most vital aspects of music information retrieval is the music genre recognition. This task aims at classifying pieces of music to their musical genre by different methods. The issue itself is not trivial because of the similarity of some genres, e.g., pop and rock or experimental and instrumental genres. Moreover, the given music track can be assigned to many quite different categories. It turns out that even the distinction of songs made by a human being is not always obvious [7]. However, it can be assumed that music can be divided into individual musical genres. Various machine learning methods can be used for automatic

* This work was supported by Statutory Research funds of Department of Applied Informatics, Silesian University of Technology, Gliwice, Poland (BKM21 – DK, BK 02/100/BK_21/0008 – RB).

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_6

recognition. The classification can be carried out with different classifiers, including those based on deep neural networks. In this research, convolutional and recurrent neural networks, were used for the genre classification task. In the first part of the presented research, different architectures of Convolutional Neural Networks (CNN) [11,13,19], and Convolutional Recurrent Neural Networks with 1D and 2D convolutions (1-D CRNN, 2-D CRNN) [15, 27], as well as Recurrent Neural Network with Long Short-Term Memory cells (LSTM) [8, 37] were used.

Moreover, to combine the advantages of different deep neural network models, the ensembles were created. Three types of ensembles were proposed, with two methods of results mixing. The first one is merging outcomes by the usage of a single fully connected layer (FCL). The input of meta-classifier is the output of first-level models, transformed by softmax function and combined together into 3-D matrix. The other one is voting (Vote), where the final classification results from majority voting of base models' predictions. All experiments were performed on Free Music Archive Dataset (FMA) [6] – one of the most popular, publicly available databases.

The issue of automatic genre recognition has practical implications in everyday life of many people. It frequently happens that a person's musical taste prefers only specific music genres. Often, one would like to listen to music from a precisely defined genre range. Considering that nowadays, people have access to millions of songs through various music services, it is impossible for a human being to divide them manually. The only solution is the automation of this process. The described practical applications are the reasons why the authors of this paper have researched in this particular direction.

1.1 Related Work

In recent years, the subject of music genre recognition [4, 20, 29, 35] become a vital field of research. The classification of songs can be carried out by various methods, e.g., using classical classifiers [1,14,31]. However, a particular intensity of work in this area can be observed in the deep learning domain [3, 10, 26]. The most popular solutions employ Convolutional Neural Networks [12, 17, 18] and Convolutional Recurrent Neural Networks [2, 9]. There is also research in which the ensembles of various deep learning models can be found (five CNNs pretrained on ImageNet or Places365 [25], combined outputs from acoustic and visual features [24], and ensemble of one CNN and one RCNN [8]).

Unfortunately, the classification of music genre, performed by other authors, has been made on various databases, with a different number of musical genres, music tracks, and various lengths. The quality of classification [23] also can be measured by many parameters for various criteria and conditions. Therefore, the comparison between the obtained results is difficult, and the conclusions could be ambiguous. Additionally, some of the obtained results, particularly high-performance results, can be deceiving and result from inadequate, flawed experimentation [36,38]. However, the parameters used in this article are one of the most popular and potentially give the best possibilities for comparison with other works.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_6

1.2 Contribution and Paper Structure

The contribution of this article is twofold. Firstly, classification of the musical tracks by the genre using several deep learning models (CNN, CRNN, LSTM) and comparing the obtained results. Secondly, proposing and implementing the classification using several ensembles – various combinations of deep learning models with two types of results' mixing methods. To the best of the knowledge of the authors of this research, there were no created such ensembles yet (with the usage of 2-dimensional convolutions and LSTM cells as well as with different types of results' mixing methods). The comparison of all outcomes, their analysis, and conclusions were made.

The deep neural network architectures created for the purpose of this work are presented in Section 2. The used dataset, description of the conducted research, and the obtained results are provided in Section 3, while the summary and the final conclusions have been included in Section 4.

2 Deep Learning Models

2.1 Convolutional Neural Network

The convolutional neural network (CNN), based on [18], consists of four convolutional layers (Fig. 1). Layer 1 and 2 have both 64 kernels each, whereas layers 3 and 4 have 128 kernels. The kernel size of all layers is equal to 5. After each layer, there is 2-D max pooling applied with kernel size and stride equal 2. In every convolutional layer, $ReLU$ is used as an activation function. Batch normalization is performed afterward. The convolutional layers are followed by one fully connected linear layer with linear activation function and the final output of 8 nodes. The mel-spectrogram is the input of the described architecture. The mel-spectrogram [21] is a type of spectrogram with the Mel scale as its vertical axis. The Mel scale is a result of a non-linear transformation of the frequency scale. The Mel scale is constructed in such a way that sounds at equal distances from each other also for people sound as if they are equidistant from each other. Converting to Mel scale, divides the frequency scale into parts and converts each part into a corresponding Mel scale. To create the mel-spectrogram the song is divided into time windows, then each window is transformed by discrete Fourier transform, from the time domain to frequency domain, then the Mel scale is created using overlapped triangular windows. Finally, the spectrogram can be created by decomposition for each window the magnitude of the signal into its components, corresponding to the frequencies in the Mel scale. In this work, mel-spectrograms with size 128 × 128 were used.

2.2 1-Dimensional Convolutional Recurrent Neural Network

1-Dimensional Convolutional Recurrent Neural Network (1-D CRNN) (Fig. 2) is based on [8], and was created to extract time-dependent features from mel-spectrograms in this model. In this architecture, three 1-D convolutional layers

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_6

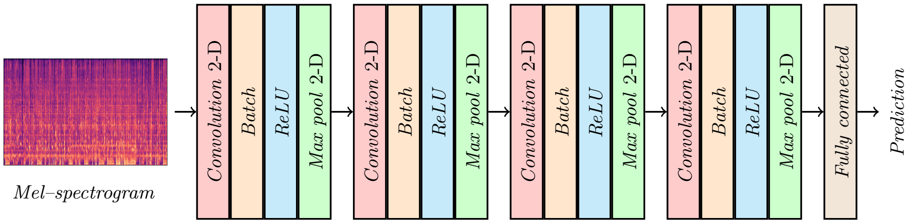

Fig. 1. The architecture of proposed convolutional neural network.

are used. The quantities of kernels are 128, 128, and 64, respectively. In every layer, kernel size equals 5, and stride size equals 1. 1-D batch normalization and *ReLU* are applied after each convolution layer. Following the convolutional layers, there are two stacked Long Short Term Memory (LSTM) cells with a hidden size equal to 128 each. In LSTM cells, dropout is conducted with a probability equal 0.2 to prevent over-fitting. Final prediction is made by one fully connected layer with linear activation.

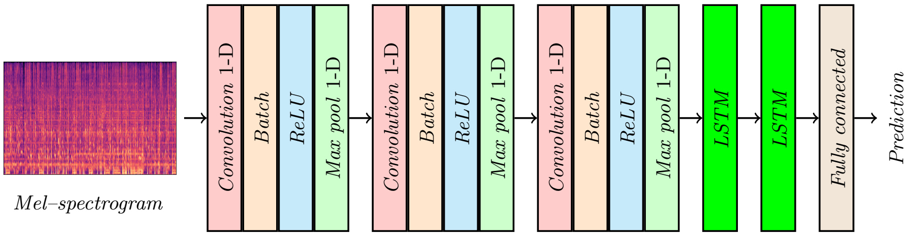

Fig. 2. The architecture of proposed 1-dimensional convolutional recurrent neural network.

2.3 2-Dimensional Convolutional Recurrent Neural Network

2-Dimensional Convolutional Recurrent Neural Network (2-D CRNN) (Fig. 3) in contrary to 1-D CRNN, consists of four 2-D convolutional layers (64, 64, 128, 128 filters respectively) with the kernel size of 3 and stride 1. 2-D batch normalization and *ReLU* are also applied after each convolutional layer. Besides, similar LSTM and fully connected layers are used afterward.

2.4 Recurrent Neural Network with Long Short-Term Memory Cells (LSTM)

Recurrent Neural Network with Long Short-Term Memory cells includes three layers with 256 hidden size, and dropout probability equals 0.3. The fully connected layer with linear activation is stacked afterward. The input is the set of

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_6

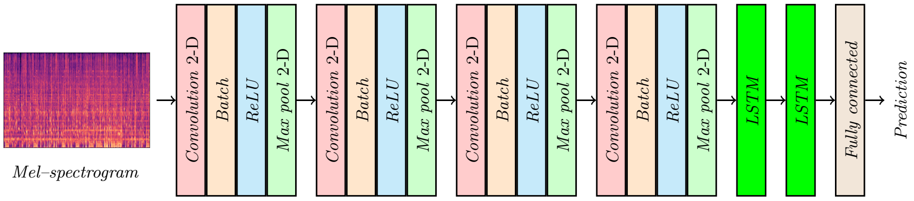

Fig. 3. The architecture of proposed 2-dimensional convolutional recurrent neural network.

MFCC values. The MFCC values are obtained by taking the logarithm of the powers at each of the mel frequencies and do discrete cosine transform [5,22,30]. The MFCC values are the amplitudes of the resulting spectrum [40]. In this work, 13 MFCC values were used for each time window.

2.5 Ensembles

In order to combine the advantages of different deep neural network architectures, three types of ensembles were proposed. All types include three networks.

Ensemble 1: Stacked CNNs and 1-D CRNN. Ensemble 1 consists of three different models. Two of them are CNNs (Section 2.1), and the third is 1-D CRNN (Section 2.2).

Ensemble 2: Stacked CNNs and 2-D CRNN. Similar to ensemble 1, this architecture is built using the two simple CNNs (section 2.1), whereas the third model used is built with 2-D CRNN (section 2.3).

Ensemble 3: Stacked RNN. The third ensemble contains three models with LSTM cells. Each model's cell varies with hyper-parameters such as the cell's hidden size (256, 128, and 256), the number of stacked layers (3, 3, and 2), and dropout probability (0.3, 0.1, and 0.2, respectively).

The output of models in each ensemble is followed by two types of results' mixing methods (to provide the final outcome). The first one is mixing outcomes by usage of the single fully connected layer (FCL). The meta-classifier input is the output results of first-level models transformed by softmax function and combined into a 3-D matrix. Meta-classifier is trained using the validation subset. The other one is voting (Vote), where the final classification results from majority voting of base models' predictions or, if there is no clear winner, meta-classifier's highest score. As a result six different ensembles was proposed (i.e., Ensemble 1 – FCL, Ensemble 1 – Vote, Ensemble 2 – FCL, Ensemble 2 – Vote, Ensemble 3 – FCL, and Ensemble 3 – Vote).

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_6

3 Experiments

3.1 Dataset and Hardware Setup

For the needs of this work, the Free Music Archive Dataset (FMA) was chosen. It was published in December 2016 by an inter-university team from Singapore and Switzerland [6]. The full FMA dataset contains 106,574 tracks and 161 genres (including sub-genres) along with its metadata and audio features. Each file has two channels and is sampled with 44.1 kHz. The biggest advantage of the dataset is that it contains a copy-free audio files, is feature-rich, and includes metadata (e.g., genre), and is already split into three subsets (small, medium, and full), which makes it additionally useful for people with lower computational resources.

The small subset of the FMA dataset contains 8000 tracks, each 30 seconds long, that are categorized into eight top-level genres such as Hip-Hop, Folk, Experimental, International, Instrumental, Electronic, Pop, and Rock. For each genre there, are 1000 tracks assigned, which makes the dataset well-balanced. That is why for training and evaluation purposes in this work, the small FMA dataset has been chosen.

All the experiments were conducted on a notebook with the following hardware configuration: CPU - Intel Core i5-9300HF, memory - 16 GB RAM, GPU - NVIDIA GeForce GTX 1650 4GB GDDR5.

3.2 Additional Settings

Data Normalization

Each of the channel is normalized according to the Equation 1.

$$I_{normalized}[d] = \frac{I[d] - mean}{std}, \tag{1}$$

where d is an index of dimension in image $d = \{0, 1, 2\}$, I is an input mel-spectrogram, $mean$ and std are arbitrary chosen values equal to 0.5. Normalization can accelerate further computing up to one order of magnitude and decrease error rate [33].

Loss Function

During training as a loss function cross entropy loss is used with the Equation 2.

$$loss(\hat{y}, class) = -\hat{y}[class] + log\left(\sum_{j=0}^{n-1} exp(\hat{y}[j])\right), \tag{2}$$

where $\hat{y}$ is a vector consisting classification outputs (in the form of predicted class index of each sample), $class$ is an index of true data class, and n is the size of mini-batch.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_6

Optimizer and Learning Rate

To perform gradient-based optimization in learning phase, Adam (adaptive moment estimation) is used. This algorithm is proven to be memory-efficient. Its another benefit is making rapid progress lowering the cost in the initial stage of training and converging faster than other popular algorithms like AdaGrad or simple stochastic gradient descent (SGD). Learning rate equals 0.001.

Regularization

It is widely believed that nets with a large number of parameters are powerful machine learning systems. Deep networks, however, are very often prone to overfitting. Especially when training set is rather small.

To address this problem dropout regularization method is applied. The term dropout refers to dropping out neurons, both visible and hidden, in a neural network. Dropping a neuron out means temporarily removing it from the network, along with the connections [34]. A single parameter p is passed to a dropout function. It is a fixed probability that a neuron can be abandoned in the training process. For every CNN models p equals 0.2 is used. Values of dropouts for RNN models are presented in Section 2.5. Dropouts are not applied while testing.

3.3 Quantitative Results

Several experiments were conducted to compare the performance and efficiency of individual models. The FMA small dataset was divided into three independent subsets in an 80:10:10 ratio. The largest subset was used for training, another two for validation and testing, respectively. After the training phase, results were measured using the following metrics: accuracy, precision, recall, and F1 score.

Table 1 contains quantitative outcomes obtained by first-level models (i.e., CNN, 1-D CRNN, 2-D CRNN, and LSTM). Evaluation is performed on a test subset unseen by the models. The values marked in bold are the highest ones.

The best values were achieved by CNN and 1-D CRNN models, while 2-D CRNN and LSTM were significantly worse. This is surprising because LSTM cells should consider the sequence of the acoustic signal and use it to their advantage.

Table 1. The results obtained by the first-level models.

Model	Accuracy [%]	Precision [%]	Recall [%]	F1 Score [%]
CNN	**51.63**	**51.81**	**50.47**	48.59
1-D CRNN	49.88	49.63	50.09	**49.43**
2-D CRNN	44.36	45.22	43.84	42.1
LSTM	45.11	45.33	44.59	42.53

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_6

Table 2. The results obtained by the ensembles.

	Model	Accuracy [%]	Precision [%]	Recall [%]	F1 Score [%]
Ensemble 1	CNN 1	51.63	51.81	50.47	48.59
	CNN 2	50.38	49.61	**55.05**	48.24
	1-D CRNN	49.88	49.63	50.09	49.43
	Ensemble 1 – FCL	55.14	55.22	54.50	53.41
	Ensemble 1 – Vote	**56.39**	**56.25**	54.80	**54.91**
Ensemble 2	CNN 1	50.13	50.62	52.96	48.94
	CNN 2	52.63	52.80	54.09	51.63
	2-D CRNN	44.36	45.22	43.84	42.10
	Ensemble 2 – FCL	53.01	52.92	52.47	50.10
	Ensemble 2 – Vote	53.13	53.28	51.97	51.76
Ensemble 3	LSTM 1	42.98	43.90	43.76	41.20
	LSTM 2	43.86	43.75	43.58	42.39
	LSTM 3	45.11	45.33	44.59	42.53
	Ensemble 3 – FCL	43.36	43.20	42.00	40.84
	Ensemble 3 – Vote	44.74	44.77	43.07	42.86

Table 2 gathers quantitative results for prepared ensembles and their base models. Underlined values are the best outcomes in a particular ensemble, and values marked in bold are the highest values at all.

The analysis of Table 2 leads to some interesting conclusions. As mentioned before, the performance of the 2-D CRNN and LSTM is noticeably lower than the CNNs and 1-D CRNN models. Ensembles 1 and 2 provide significantly better accuracy, precision, and F1 score than the first-level models' values. Moreover, the outcomes show that the voting result's mixing method is slightly better than the fully connected layer. However, in most cases, an ensemble with a fully connected layer can still provide better results than any base model.

3.4 Qualitative Results

Table 3 shows the confusion matrix obtained for Ensemble 1 with the voting model, the best of all developed models. From the results obtained, it can be concluded that not all musical genres are classified with similar effectiveness. For Hip-Hop, almost 80% accuracy was achieved, and for the next four genres (i.e., Rock, Folk, Electronic, and International), the accuracy was between 60% and 70%. This seems to be a satisfactory result. The worst results were for Pop and Experimental, almost 20% and 40% respectively. This is due to the fact that both of these genres combine the features of other genres. Depending on the particular Pop track, it sometimes has Rock, Hip-Hop, Electronic, or other genres' characteristics.

The assignment of a song to a particular musical genre is often very challenging, even for a skilled listener and music expert. This difficulty increases significantly in vaguely defined genres, such as Pop, Experimental, etc.

Table 3. Confusion matrix for the best model – Ensemble 1 with Voting [in %].

		Predicted label							
		Hip-Hop	Folk	Experimental	International	Instrumental	Electronic	Pop	Rock
True label	Hip-Hop	79.4	2.1	1.0	2.1	2.1	9.3	3.1	1.0
	Folk	2.0	65.0	5.0	5.0	13.0	2.0	6.0	2.0
	Experimental	5.6	6.5	39.3	4.7	16.8	10.3	5.6	11.2
	International	15.3	4.7	2.4	61.2	1.2	4.7	7.1	3.5
	Instrumental	3.0	10.9	14.9	1.0	56.4	2.0	2.0	9.9
	Electronic	15.6	0.0	3.7	0.9	11.0	62.4	4.6	1.8
	Pop	11.8	15.1	3.2	18.3	7.5	11.8	19.4	12.9
	Rock	0.9	10.4	5.7	4.7	0.0	3.8	7.6	67.0

3.5 Comparison of the Outcomes

As part of comparing the obtained results with other works' outcomes, a cumulative table was created (Table 4). It contains the results of the works [16, 28, 32, 39, 41, 42], and, at the bottom, in bold, the best results of this research, for Ensemble 1 with voting.

Table 4. Comparison of different models classifying FMA small with the proposed Ensemble 1 with voting. Recall and F1 score were provided in [39] only. All values are in %.

No.	Model	Accuracy	No.	Model	Accuracy	Recall	F1 Score
1	K-Nearest Neighbors [42]	36.4	12	MoER [41]	55.9	–	–
2	Logistic Regression [42]	42.3	13	FCN [39]	63.9	43.0	40.3
3	Multilayer Perceptron [42]	44.9	14	TimbreCNN [39]	61.7	36.4	35.0
4	Support Vector Machine [42]	46.4	15	End-to-end [39]	61.4	38.4	34.5
5	Original spectrogram [41]	49.4	16	CRNN [39]	63.4	40.7	40.2
6	Harmonic spectrogram [41]	43.4	17	CRNN-TF [39]	64.7	43.5	42.3
7	Percussive spectrogram [41]	50.9	18	CRNN [32]	53.5	–	–
8	Modulation spectrogram [41]	55.6	19	CNN-RNN [32]	56.4	–	–
9	MFCC [41]	47.1	20	CNN TL [16]	51.5	–	–
10	MoEB [41]	54.1	21	CNN TL [28]	56.8	–	–
11	MoEC [41]	55.6	22	C-RNN [42]	65.2	–	–
23	**Ensemble 1 - Vote**	**56.4**	**23**	**Ensemble 1 - Vote**	**56.4**	**54.8**	**54.9**

When analyzing Table 4, it can be concluded, that in general, the values obtained by the authors are within the range of those obtained in other similar studies. Lines 1–4 shows (in Table 4) the values achieved by classical classification methods [42], which are lower by 10–20% than those obtained by Ensemble 1. Lines 5–8 presents results of CNNs fed by different spectrograms, line 9 – was treated as traditionally obtained baseline, while lines 10–12 reveals values (all slightly worse than Ensemble 1) achieved by mixtures of CNNs fed by different spectrograms [41]. Lines 13–17 shows the quantitative outcomes of Fully Convolutional Neural Network, Timbre CNN, End-to-end approach for music auto-tagging, CRNN, and CRNN with Time and Frequency dimensions, respectively [39]. All these methods achieved higher accuracy than Ensemble 1 (up to 8.3%). On the other hand, recall and F1 score are way lower than in presented research (11.3% and 12.6%, respectively). Lines 18–21 presents outcomes for different CRNN models [32] as well as CNN models with the use of transfer learning [16, 28]. These results are very close to Ensemble 1 with voting. Moreover, one of them (line 21) is slightly higher (by 0.4%). The last result (line 22) is much higher than obtained by the authors of this work. However, there is a conceptual defect (data leakage from the training of deep learning model to the test data), which leads to overoptimistic outcomes. Summarizing all the results, it can be noticed that the results achieved by the authors of this research are in the range of, or even above the state-of-the-art solutions.

4 Conclusions

We introduced new types of ensembles, consisting of several different deep neural networks as base models, followed by two different results mixing methods, to solve the classification of songs according to musical genres. We have conducted a series of varied experiments for both the developed deep neural network base models (i.e., CNN, 1-D CRNN, 2-D CRNN, and RNN with LSTM cells) and the elaborated ensembles. The quantitative and qualitative studies have shown that the outcomes achieved easily match and in most cases even beat state-of-the-art methods. Moreover, in this research, only mel-spectrogram and MFCCs were used as the input data. Precise enlargement of the number of input signals will likely allow for a further increase in classification quality.

Future work will be centered around conducting experiments on other deep network architectures (possibly with attention mechanism) and other ensemble configurations. A next step, completely different and more demanding research, would be to classify on the assumption that a given work can belong to several musical genres, i.e., a multi-label approach.

The main advantage of using our approach for music genre classification is the novel, automatic application of an ensemble of different deep learning architectures. In case of finding other good base models, they could be easily applied in the ensemble and thus increase the overall quality of the classification.

It has to be remembered that the creating of the ensemble is more labor-intensive, and the classification itself is longer. However, if the main goal is the quality of the assigning song to the proper musical genre process, the possibility of its improving is always valuable.

References

1. Basili, R., Serafini, A., Stellato, A.: Classification of musical genre: a machine learning approach. In: ISMIR (2004)
2. Choi, K., Fazekas, G., Sandler, M., Cho, K.: Convolutional recurrent neural networks for music classification. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2392–2396. IEEE (2017)

3. Choi, K., Fazekas, G., Sandler, M., Cho, K.: Transfer learning for music classification and regression tasks. arXiv preprint arXiv:1703.09179 (2017)
4. Costa, Y.M., Oliveira, L.S., Silla Jr, C.N.: An evaluation of convolutional neural networks for music classification using spectrograms. Applied soft computing **52**, 28–38 (2017)
5. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE transactions on acoustics, speech, and signal processing **28**(4), 357–366 (1980)
6. Defferrard, M., Benzi, K., Vandergheynst, P., Bresson, X.: Fma: A dataset for music analysis. arXiv preprint arXiv:1612.01840 (2016)
7. Dong, M.: Convolutional neural network achieves human-level accuracy in music genre classification. arXiv preprint arXiv:1802.09697 (2018)
8. Ghosal, D., Kolekar, M.H.: Music genre recognition using deep neural networks and transfer learning. In: Interspeech. pp. 2087–2091 (2018)
9. Gunawan, A.A., Suhartono, D., et al.: Music recommender system based on genre using convolutional recurrent neural networks. Procedia Computer Science **157**, 99–109 (2019)
10. Kereliuk, C., Sturm, B.L., Larsen, J.: Deep learning and music adversaries. IEEE Transactions on Multimedia **17**(11), 2059–2071 (2015)
11. Khan, A., Sohail, A., Zahoora, U., Qureshi, A.S.: A survey of the recent architectures of deep convolutional neural networks. Artificial Intelligence Review **53**(8), 5455–5516 (2020)
12. Kim, T., Lee, J., Nam, J.: Sample-level cnn architectures for music auto-tagging using raw waveforms. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 366–370. IEEE (2018)
13. Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., Inman, D.J.: 1d convolutional neural networks and applications: A survey. arXiv preprint arXiv:1905.03554 (2019)
14. Kostrzewa, D., Brzeski, R., Kubanski, M.: The classification of music by the genre using the knn classifier. In: International Conference: Beyond Databases, Architectures and Structures. pp. 233–242. Springer (2018)
15. Labach, A., Salehinejad, H., Valaee, S.: Survey of dropout methods for deep neural networks. arXiv preprint arXiv:1904.13310 (2019)
16. Lee, D., Lee, J., Park, J., Lee, K.: Enhancing music features by knowledge transfer from user-item log data. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 386–390. IEEE (2019)
17. Lee, J., Nam, J.: Multi-level and multi-scale feature aggregation using pretrained convolutional neural networks for music auto-tagging. IEEE signal processing letters **24**(8), 1208–1212 (2017)
18. Lim, M., Lee, D., Park, H., Kang, Y., Oh, J., Park, J.S., Jang, G.J., Kim, J.H.: Convolutional neural network based audio event classification. KSII Transactions on Internet & Information Systems **12**(6) (2018)
19. Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., Alsaadi, F.E.: A survey of deep neural network architectures and their applications. Neurocomputing **234**, 11–26 (2017)
20. McKay, C., Fujinaga, I.: Musical genre classification: Is it worth pursuing and how can it be improved? In: ISMIR. pp. 101–106 (2006)
21. Mermelstein, P.: Distance measures for speech recognition, psychological and instrumental. Pattern recognition and artificial intelligence **116**, 374–388 (1976)

22. Mogran, N., Bourlard, H., Hermansky, H.: Automatic speech recognition: An auditory perspective. In: Speech processing in the auditory system, pp. 309–338. Springer (2004)
23. Moska, B., Kostrzewa, D., Brzeski, R.: Influence of the applied outlier detection methods on the quality of classification. In: International Conference on Man–Machine Interactions. pp. 77–88. Springer (2019)
24. Nanni, L., Costa, Y.M., Aguiar, R.L., Silla Jr, C.N., Brahnam, S.: Ensemble of deep learning, visual and acoustic features for music genre classification. Journal of New Music Research **47**(4), 383–397 (2018)
25. Nanni, L., Maguolo, G., Brahnam, S., Paci, M.: An ensemble of convolutional neural networks for audio classification. arXiv preprint arXiv:2007.07966 (2020)
26. Oramas, S., Nieto, O., Barbieri, F., Serra, X.: Multi-label music genre classification from audio, text, and images using deep features. arXiv preprint arXiv:1707.04916 (2017)
27. Pamina, J., Raja, B.: Survey on deep learning algorithms. International Journal of Emerging Technology and Innovative Engineering **5**(1) (2019)
28. Park, J., Lee, J., Park, J., Ha, J.W., Nam, J.: Representation learning of music using artist labels. arXiv preprint arXiv:1710.06648 (2017)
29. Pons, J., Serra, X.: Randomly weighted cnns for (music) audio classification. In: ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 336–340. IEEE (2019)
30. Sahidullah, M., Saha, G.: Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition. Speech communication **54**(4), 543–565 (2012)
31. Silla, C.N., Koerich, A.L., Kaestner, C.A.: A machine learning approach to automatic music genre classification. Journal of the Brazilian Computer Society **14**(3), 7–18 (2008)
32. Snigdha, C., Kavitha, A.S., Shwetha, A.N., Shreya, H., Vidyullatha, K.S.: Music genre classification using machine learning algorithms: A comparison. International Research Journal of Engineering and Technology **6**(5), 851–858 (2019)
33. Sola, J., Sevilla, J.: Importance of input data normalization for the application of neural networks to complex industrial problems. IEEE Transactions on nuclear science **44**(3), 1464–1468 (1997)
34. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)
35. Sturm, B.L.: A survey of evaluation in music genre recognition. In: International Workshop on Adaptive Multimedia Retrieval. pp. 29–66. Springer (2012)
36. Sturm, B.L.: The state of the art ten years after a state of the art: Future research in music information retrieval. Journal of New Music Research **43**(2), 147–172 (2014)
37. Tang, C.P., Chui, K.L., Yu, Y.K., Zeng, Z., Wong, K.H.: Music genre classification using a hierarchical long short term memory (lstm) model. In: Third International Workshop on Pattern Recognition. vol. 10828, p. 108281B. International Society for Optics and Photonics (2018)
38. Urbano, J., Schedl, M., Serra, X.: Evaluation in music information retrieval. Journal of Intelligent Information Systems **41**(3), 345–369 (2013)
39. Wang, Z., Muknahallipatna, S., Fan, M., Okray, A., Lan, C.: Music classification using an improved crnn with multi-directional spatial dependencies in both time and frequency dimensions. In: 2019 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2019)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_6

40. Xu, M., Maddage, N.C., Xu, C., Kankanhalli, M., Tian, Q.: Creating audio keywords for event detection in soccer video. In: 2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698). vol. 2, pp. II–281. IEEE (2003)
41. Yi, Y., Chen, K.Y., Gu, H.Y.: Mixture of cnn experts from multiple acoustic feature domain for music genre classification. In: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). pp. 1250–1255. IEEE (2019)
42. Zhang, C., Zhang, Y., Chen, C.: Songnet: Real-time music classification. Stanford University Press (2019)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_6