Automatic Music Genre Classification and Its Relation with Music Education

Hasan Can Ceylan[1, *], Naciye Hardalaç[2], Ali Can Kara[1] & Fırat Hardalaç[1]

[1]Department of Electrical & Electronics Engineering, Gazi University, Ankara, Turkey

[2]Department of Music Education, Gazi University, Ankara, Turkey

*Correspondence: Department of Electrical & Electronics Engineering, Gazi University, Ankara, Maltepe, 06570, Turkey. Tel: 90-505-705-5472. E-mail: can_cyln@hotmail.com

Received: February 4, 2021 Accepted: March 24, 2021 Online Published: April 12, 2021

doi:10.5430/wje.v11n2p36 URL: https://doi.org/10.5430/wje.v11n2p36

Abstract

Because the classification saves time in the learning process and enables this process to take place more easily, its contribution to music learning cannot be denied. One of the most valid and effective methods in music classification is music genre classification. Given the rapid progress of music production in the world and the significant increase in the number of data, the process of classifying music genres has now become too complex to be done by humans. Considering the successful results of deep neural networks in this field, the aim is to develop a deep learning algorithm that can classify 10 different music genres. To reveal the efficiency of the model by comparing it with others, we make the classification using the GTZAN dataset, which was previously used in many studies and retains its validity. In this article, we use a convolutional neural network (CNN) to classify music genres, taking into account the previous successful results. Unlike previous studies in which CNN was used as a classifier, we represent music segments in the dataset by mel frequency cepstral coefficients (MFCC) instead of using visual features or representations. We obtain MFCCs by preprocessing the music pieces in the dataset, then train a CNN model with the acquired MFCCs and determine the success of the model with the testing data. As a result of this study, we develop a model that is successful in classifying music genres by using smaller data than previous studies.

Keywords: deep learning, music education, music genre classification, convolutional neural networks (CNN), mel frequency cepstral coefficients (MFCC).

1. Introduction

Since it came into existence, humanity has needed to put complex events or facts into order so as to facilitate perception and understanding processes. One of the first and most basic approaches to meeting this need is classification (Ersoy, 2017). For this reason, it would not be wrong to consider classification, which is a necessity in the learning process, as a prerequisite for obtaining information more easily. The importance of the concept of order, which provides convenience at every stage of life, in the process of acquiring knowledge cannot be denied (Soykan, 2012). From this point of view, the necessity of creating order through classification can be determined in the actions taken to provide understanding and perception.

Music is an area where learning is expected as a result of processes such as perception and discrimination. During the process, the field of music, like all scientific fields, has assumed an intricate structure, so it has become difficult to comprehend and perceive this multifactor construct. There is a need for systematic order in the field of music, which is accumulating and diversifying. As explained by Stowell and Dixon (2011), one of the most valid methods in maintaining this order is music genre classification. This process, however, has become almost impossible to be done by humans, owing to the increasing and diversifying music genres and songs. Therefore, in this study, we aimed to develop a deep learning model that can automatically classify 10 different music genres. Thanks to this model, the music classification process can be done in a shorter time and with a higher success rate than when done by humans.

In section 2, we explain the methods and materials used in the study. In section 3, we examine some of the previous studies on music genre classification. We provide experimental results in section 4. In section 5, we offer a summary of the study and interpret the results.

2. Method

In this section, we explicate features of the dataset and its intended use. Then we explain the mel frequency cepstral coefficients (MFCC) and the preprocessing steps we used to obtain these values. In additon, we expound the concepts of deep learning and convolutional neural networks (CNN). We then give the proposed CNN model and parameters of this model. We provide an overview of the steps followed in the study in Figure 1.

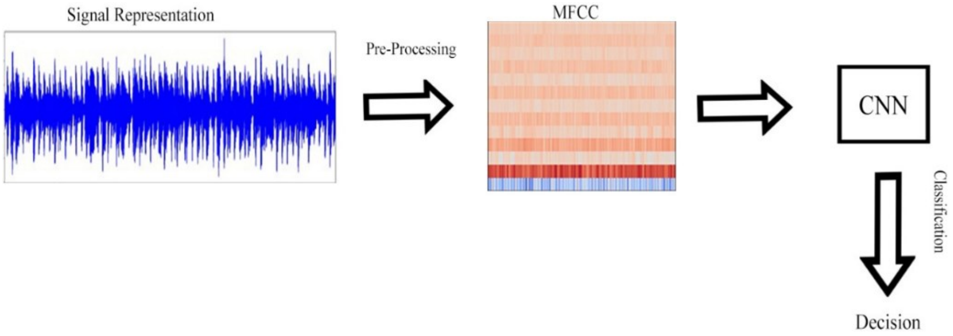


Figure 1. General Classification Scheme

As can be understood from Figure 1, in this study, we first preprocessed the data and transformed each datum into structures represented by MFCCs. Later, we carried out the training of CNN, which was decided as a result of the experiments, using these structures. We decided the parameters of the model as a result of the experiments. While determining the parameters, we took into account the goal of realizing the most successful classification. Then, we determined the classification success of the model with the testing data.

2.1 Dataset

Deep learning models need data when performing operations such as classification, regression, and clustering. Therefore, the distribution and accuracy of the dataset to be selected are of great importance. Also, the dataset to be selected should have been used in previous studies in this field and should maintain its validity because the necessity and validity of one's model are determined by comparing it with previous studies. Considering that it had been used in many previous studies and would enable the model to be evaluated more accurately, we used the GTZAN dataset (Tzanetakis & Cook, 2002), which includes 10 different music genres (blues, classical, country, disco, jazz, hip-hop, metal, pop, reggae, rock) and 100 different samples for each genre. The dataset contains samples lasting 30 seconds, of 22 050 Hz frequency, and 16 bits.

2.2 Preprocessing

The music segments are structures represented as sound signals. These signals, which have direction and intensity, have constantly changing values depending on time. To make the audio signals understandable by the computer, these analog signals should be converted into digital signals by sampling with the specified sampling rate. Signals are represented by graphs created with the amplitude and time information acquired as a result of this process. In Figures 2 and 3, the signal representations of two different data in the GTZAN dataset are given.

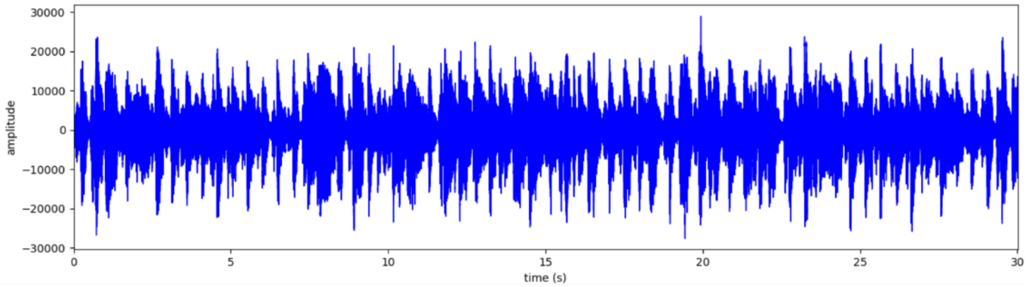


Figure 2. Signal Representation of Blues-0

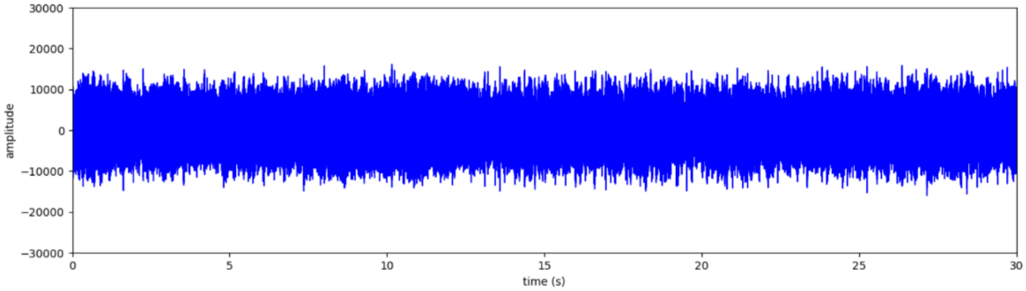


Figure 3. Signal Representation of Metal-0

Audio signals that change constantly depending on time and do not contain order should not be examined as a whole. The reason for this is that they are complex and the learning to be made with the information to be obtained does not have enough efficiency. For this reason, the signals are divided into a certain number of parts, and information is attained from these parts. Two variables called frame length and hop length determine the size of these parts. Frame length represents the number of instances in a frame, while hop length is the variable that determines how many samples each frame will start after the previous frame (Karatana & Yıldız, 2017).

Graphics and amplitude values on the time plane do not contain enough information to represent the properties of an audio signal. Fourier transform is the method used to represent these signals in the frequency domain, based on the idea that these signals are formed by combining more than one sinusoidal signal with different frequency values. With this transformation, the new graphs obtained with frequency and amplitude values by ensuring the transition of the signal from the time plane to the frequency plane are called a spectrum. Figures 4 and 5 show spectrum plots of signals previously represented in the time domain.

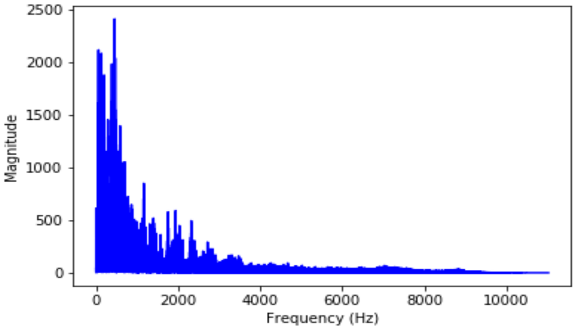


Figure 4. Fast Fourier Transform (FFT) of Blues-0

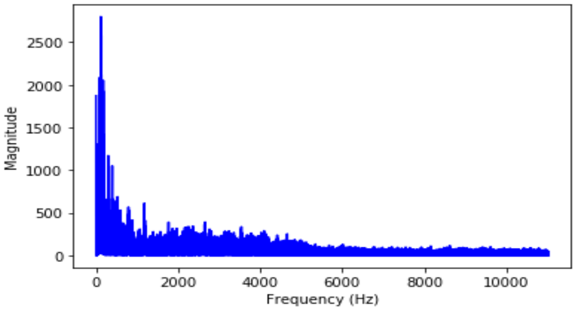


Figure 5. FFT of Metal-0

Graphics resulting from the Fourier transform do not contain time information. To avoid this situation, the audio signal is divided into windows, and Fourier transform is applied to these windows. This application is called short-time Fourier transform, and the graphics that result from the process are called spectrograms (Karatana & Yıldız, 2017). Figures 6 and 7 show spectrogram images of the data.

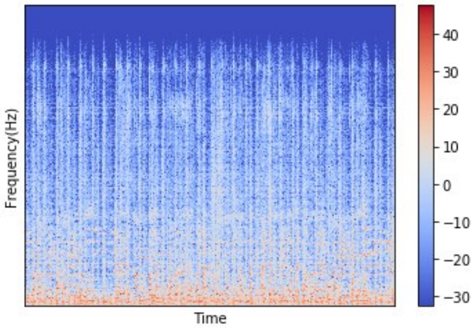


Figure 6. Spectrogram of Blues-0

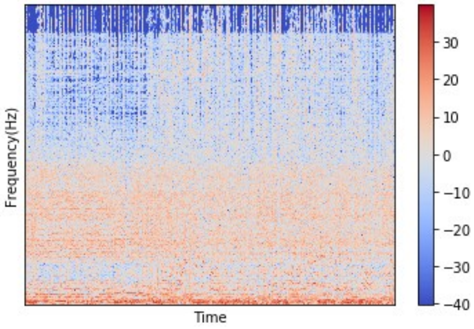


Figure 7. Spectrogram of Metal-0

2.3 Mel Frequency Cepstral Coefficients (MFCC)

MFCCs are features that enable distinction similar to the human voice perception system. In this way, MFCCs have a structure that can achieve high performance in applications such as music classification and voice recognition. Cepstrum is defined as the inverse Fourier transform of the logarithmic Fourier transform of a signal. Before taking the logarithm of the cepstrum, if it is arranged according to the mel filters in Figure 8, which shows similar responses to human hearing nerves, the MFCCs are obtained (Kızrak & Bolat, 2015; Molau et al., 2001).

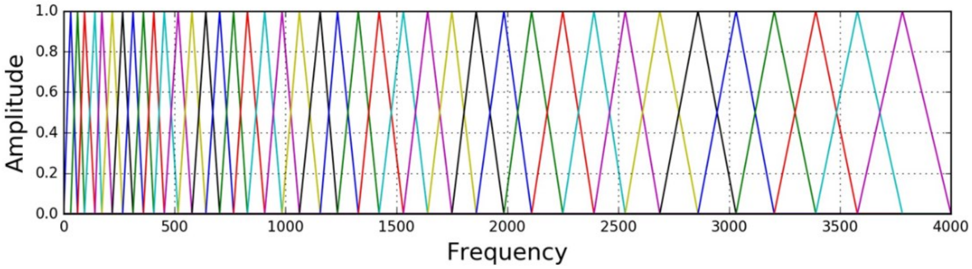


Figure 8. Mel Filters

In our study, we aim to classify music genres. Considering that music is produced for humans, we think it is the right choice to use MFFCs that respond similarly to the human voice perception system. In Figures 9 and 10, we give the MFCC graphics of the data, whose graphics were shown earlier. Graphics can be defined as images that represent the changes of MFCC according to frequency values with a color scale.

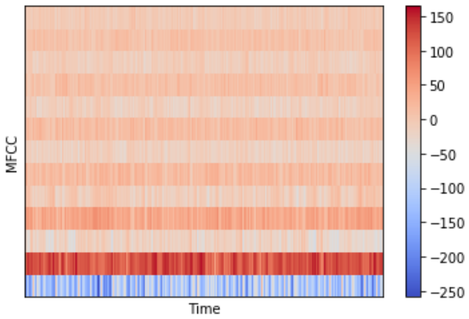


Figure 9. MFCC of Blues-0

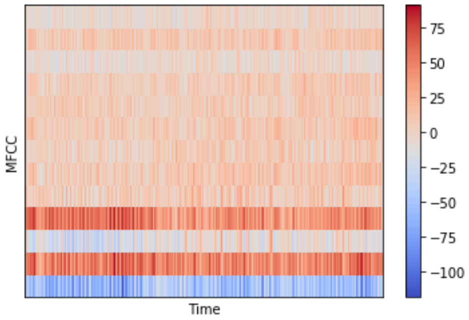


Figure 10. MFCC of Metal-0

As in Zhang et al. (2016), we divided each piece into 3-second pieces, thus ensuring data augmentation. We used the MFCCs of 13x130, which we obtained by selecting the n-mffc, hop-length, and n-fft values as 13, 512, and 2048 respectively (with the use of Python's Librosa library). We divided the 13x130 samples into two as 80% training and 20% testing. As a result of the experiments carried out, we selected all values in such a way that we could make the most efficient classification.

2.4 Deep Learning

Deep learning is a type of machine learning. Machine learning is a learning model in which large datasets are used by the machine. In short, the machine learning method enables artificial intelligence to learn by itself.

Deep learning allows us to train artificial intelligence to predict outputs with a given dataset. Both supervised and unsupervised learning can be used to train artificial intelligence. Supervised learning is a learning method in which control is provided depending on the outputs to be acquired according to the labeled inputs. Unsupervised learning is a type of learning that is used in uncertain situations and where the control is completely left to artificial intelligence.

2.4.1 Artificial Neural Networks

Artificial neural networks are structures similar to human neurons. Neurons are linked to each other and their effect on output is determined by their weight. The artificial neural network consists of three layers: input layer, hidden layer, and output layer. Hidden layers determine the system's response and efficiency based on inputs and outputs. Because the effect of variables such as the number of neurons and the number of hidden layers in artificial neural networks is not known precisely, they are produced in variable structures. This situation can be given as an example of the difficult situations encountered in neural networks.

2.4.2 Deep Neural Networks

"Deep" in deep neural networks refers to having more than one hidden layer. In other words, deep neural networks can be considered as structures in which the number of hidden layers in artificial neural networks is increased. In this way, they make better sense of the relationships among data and enable us to achieve better results.

2.5 Convolutional Neural Networks (CNN)

The CNN is a deep learning algorithm developed for image processing, which has made significant progress in recent years and is widely used in areas such as face recognition, object detection, and speech recognition. It generally has a structure consisting of the layers shown in Figure 11.

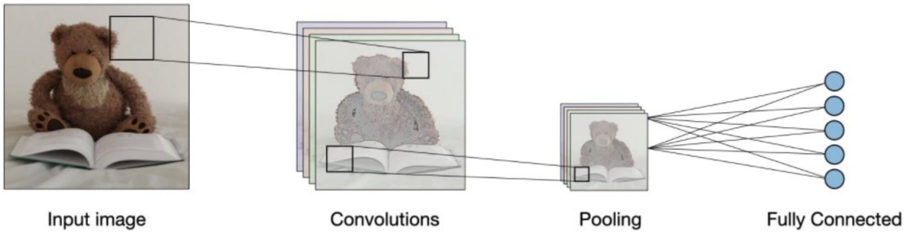


Figure 11. CNN Layers (Khvatova, 2019)

Convolution Layer: In this layer, feature maps are obtained by shifting the filters over the input image. These feature maps can be considered as pixel values representing the image according to the specified filters.

Pooling Layer: In this layer, size reduction is applied to the feature maps that were previously attained in the convolution layer. The purpose of this process is to keep the most important features by reducing the number of parameters.

Fully Connected Layer: In this layer, the final feature vectors are acquired, and the process is completed by determining the number of neurons in the last layer according to the number of classes to be classified.

2.6 Proposed CNN Model

We gave MFCCs resized to 26x65 as input to CNN. The CNN with the most successful results in experiments consists of the following layers:

- Convolution Layer with filters: 256, kernel size: 3x3, activation function: relu, padding: valid,
- Convolution Layer with filters: 256, kernel size: 3x3, activation function: relu, padding: valid,
- Average Pooling with pool size: 3x3, strides: 2x2, padding: same,
- Convolution Layer with filters: 256, kernel size: 3x3, activation function: relu, padding: valid,
- Average Pooling with pool size: 3x3, strides: 2x2, padding: same,
- Convolution Layer with filters: 512, kernel size: 4x4, activation function: relu, padding: valid,
- Global Average Pooling,
- Fully Connected Layer,
- Dense Layer with units: 256, activation function: relu,
- Dense Layer with units: 128, activation function: relu,
- Dense Layer with units: 10, activation function: softmax.

Considering the results we obtained in the experiments, we chose Adam as the optimizer. We carried out the training with 80 epochs using the backpropagation algorithm. We made all the selections to ensure that the model worked more stably and accurately. Deep neural network architecture is given in Figure 12.

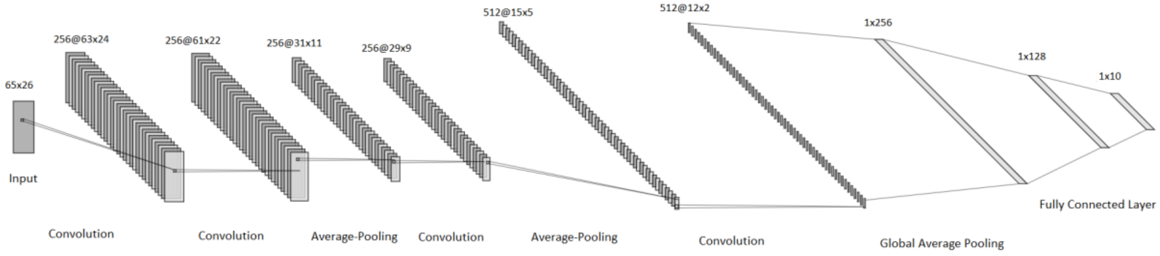


Figure 12. Convolutional Neural Network Architecture

3. Literature Review

Tzanetakis and Cook (2002) were the first to define the classification of musical genres as a pattern recognition problem. These researchers made use of several musical features such as rhythm, timbre, and pitch in the process of classifying music. Besides, the authors created the GTZAN dataset, which includes 10 different music genres and 100 tracks in each genre. This dataset has become one of the most important datasets in music genre classification. It is a frequently used dataset that continues to be valid today.

GTZAN and the different datasets created later have both played a pioneering role in studies of music genre classification and contributed to these studies. In the music genre classification process, which was initially performed using machine learning methods, deep learning methods have become available today, thanks to the developing technology and the use of gpu. CNN, which we used in our study and has a high success rate owing to its results, was used for the first time in 2012 for this process. Humphrey and Bello (2012) stated that by using CNN, which they developed to resolve the difficulties encountered in previous studies and to develop an alternative approach, they achieved higher performance than the models that previously performed the chord recognition process.

Zhang et al. (2016) developed two different deep learning models to increase the performance rate in music genre classification. In the first model, they used a combination of maximum pooling and average pooling, unlike the previous ones. In the second model, in addition to the first, they created the model by adding features from a different layer to the network using shortcut links (He et al., 2016). They achieved successful results compared to previous studies performed using the GTZAN dataset.

Nanni et al. (2018) examined deep learning, acoustic features, and visual features in music genre classification. As a result of the studies, they observed that the features attained by classical methods and the features used by CNN contain different information. By using the properties obtained from these two situations together, they acquired more successful results than previous studies.

4. Results

In this section, we share the results of the study. We obtained the results with the k-fold cross-validation method, in which test and training data were changed to ensure a more accurate evaluation. The testing accuracy rate was attained as the average of five-fold. In our study, we took a 94.5% accuracy rate into consideration as state of the art, as in Choi et al. (2017). Even if this rate is exceeded in different studies, the highest meaningful performance rate that can be achieved based on noise and errors in the GTZAN dataset can be 94.5%, as explained in Sturm (2013). In Table 1, we give the result of our model on the dataset presented in Tzanetakis and Cook (2002) and the results of other models applied to the same dataset.

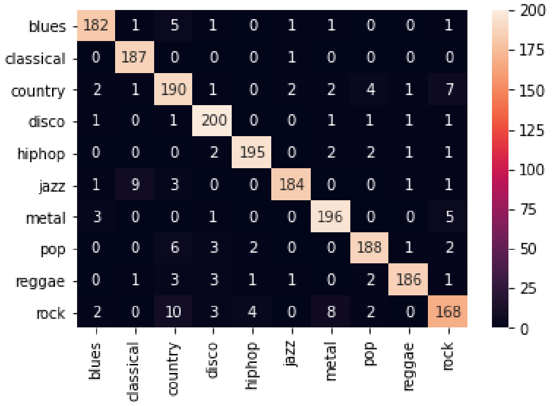
Table 1. Results on GTZAN Dataset

	Testing Accuracy (%)
Proposed Model	93.4
Lee et al. (2018)	82.1
Choi et al. (2017)	89.8
Lee et al. (2009)	90.6
Nanni et al. (2017)	90.6
Zhang et al. (2016)	87.4
Feng et al. (2017)	92.0

The proposed model achieved 93.4% testing and 99.3% training accuracy. To evaluate the genres separately, we give the precision, recall, and f1-score values acquired for each class in Table 2. In Figure 13, for one of the five-folds, we give a confusion matrix created according to the model's predictions.

Table 2. Precision, Recall, and f1-scores by Genres

	Precision (%)	Recall (%)	F1-Score (%)
Blues	94.9	94.4	94.6
Classical	93.6	99.1	96.2
Country	86.8	90.1	88.4
Disco	93.1	96.7	94.8
Hip-Hop	96.1	95.7	95.9
Jazz	97.0	92.1	94.4
Metal	92.9	95.2	94.1
Pop	94.1	92.7	93.4
Reggae	97.0	93.6	95.2
Rock	89.4	85.0	87.1



	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock
blues	182	1	5	1	0	1	1	0	0	1
classical	0	187	0	0	0	1	0	0	0	0
country	2	1	190	1	0	2	2	4	1	7
disco	1	0	1	200	0	0	1	1	1	1
hiphop	0	0	0	2	195	0	2	2	1	1
jazz	1	9	3	0	0	184	0	0	1	1
metal	3	0	0	1	0	0	196	0	0	5
pop	0	0	6	3	2	0	0	188	1	2
reggae	0	1	3	3	1	1	0	2	186	1
rock	2	0	10	3	4	0	8	2	0	168

Figure 12. Confusion Matrix of 93.8% Accurate Fold

We obtained the highest performance rates in the classical and hip-hop genres. We obtained the lowest rates in the rock and country genres. Although not certain, the reason for this situation may be that the 3-second pieces do not contain enough information for these two genres or there are similar sounds in these two types. The reason why a definitive judgment cannot be made is that there are still situations in deep learning that are not fully explained.

5. Conclusion

Music has a multi-element and complex structure. Therefore, examining music by dividing it into subclasses provides better understanding and learning. One of the most valid methods we can use when making this distinction is to classify music genres. Because in the process of learning and teaching music, music genre classification is of great importance in terms of facilitating perception and interpretation and creating more distinct lines in the learning process. The idea that deep learning can perform this classification process faster and with more successful results than humans has been the main starting point of this study. Therefore, considering the results of the models used in previous studies and the experiments on the subject, we carried out the classification process using CNN and 26x65 MFCCs. As a result of the study, we achieved a level of success that can be compared with studies in which many features or more complex and combined deep learning models were used. Also, the use of smaller data resulted in less training time and lower memory usage.

In future studies, by creating and using datasets containing more genres and tracks, models that classify music genres more accurately and provide more benefits to the learning and teaching process of music can be revealed. Besides, if the developed models are integrated into music recommendation systems, the efficiency and accuracy of these systems can be increased.

Acknowledgments

We thank Hilal Tuğçe Ceylan for helping us in translation and editing for this article and comments on the manuscript.

References

Choi, K., Fazekas, G., Sandler, M. B., & Cho, K. (2017). *Transfer learning for music classification and regression tasks*. Proceedings of the 18th International Society for Music Information Retrieval Conference, (ISMIR), Suzhou, China, 141-149.

Ersoy, İ. (2017). Müzikte Tür Kavramı: Müzik Türleri Sınıflandırmasında Yeni Bir Model Önerisi. *EÜ Devlet Türk Musikisi Konservatuvarı Dergisi, 11*, 1-16. http://doi.org/10.31722/konservatuvardergisi.517061

Feng, L., Liu, S., & Yao, J. (2017). Music Genre Classification with Paralleling Recurrent Convolutional Neural Network. *arXiv preprint, arXiv:1712.08370.*

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, Nevada, USA, 770-778. http://doi.org/10.1109/CVPR.2016.90

Humphrey, E., & Bello, J. (2012). *Rethinking automatic chord recognition with convolutional neural networks*. Proceedings of International Conference on Machine Learning and Application (ICMLA), Boca Raton, Florida, USA, 11(2), 357-362. http://doi.org/10.1109/ICMLA.2012.220

Karatana, A., & Yıldız, O. (2017). *Music genre classification with machine learning techniques*. 25th IEEE Signal Processing and Communications Applications Conference (SIU), Antalya, Turkey, 1-4. http://doi.org/10.1109/SIU.2017.7960694

Khvatova, K. (2019, April 1) Is Convolutional Neural Network a black box? Not anymore. *Medium*. Retrieved from https://medium.com/@kristinakhvatova/is-convolutional-neural-network-a-black-box-not-anymore-caca1429952d

Kızrak, M. A., & Bolat, B. (2015). *Classification of Classic Turkish Music Makams by using Deep Belief Networks*. 23nd Signal Processing and Communications Applications Conference (SIU), Malatya, Turkey, 527-530. http://doi.org/10.1109/SIU.2015.7129877

Lee, C. H., Shih, J. L., Yu, K. M., & Lin, H. S. (2009). Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features. *IEEE Transactions on Multimedia*, *11*(4)*,* 670-682. http://doi.org/10.1109/TMM.2009.2017635

Lee, J., Park, J., Kim, K., & Nam, J. (2018). SampleCNN: End-to-end deep convolutional neural networks using very small filters for music classification. *Applied Sciences, 8*(1), 150. http://doi.org/10.3390/app8010150

Molau, S., Pitz, M., Schluter, R., & Ney, H. (2001). *Computing Mel-frequency Cepstral Coefficients on the Power Spectrum.* IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, Utah, USA, 73-76. http://doi.org/10.1109/ICASSP.2001.940770

Nanni, L., Costa, Y. M. G., Aguiar, R. L., Silla, C. N., & Brahnam, S. (2018). Ensemble of deep learning visual and acoustic features for music genre classification. *Journal of New Music Research, 47*(4), 383-397. http://doi.org/10.1080/09298215.2018.1438476

Nanni, L., Costa, Y. M. G., Lucio, D. R., Silla, C. N., & Brahnam, S. (2017). Combining visual and acoustic features for audio classification tasks**.** *Pattern Recognition Letters*, *88*, 49-56. http://doi.org/10.1016/j.patrec.2017.01.013

Soykan, Ö. N. (2012). Müzik Nedir? Felsefi Bir Araştırma*. Doğu Batı, 62,* 29-42.

Stowell, D., & Dixon, S. (2011). *MIR in school? Lessons from ethnographic observation of secondary school music classes.* Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR), Miami, Florida, USA, 347-352.

Sturm, B. L. (2013). The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use. *arXiv preprint, arXiv:1306.1461.*

Tzanetakis, G., & Cook, P. (2002). Music genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing, 10*(5), 293-302. http://doi.org/10.1109/TSA.2002.800560

Zhang, W., Lei, W., Xu, X., & Xing, X. (2016). Improved music genre classification with convolutional neural networks. *Proc. Interspeech 2016*, 3304-3308. http://doi.org/10.21437/Interspeech.2016-1236

Copyrights