Music genre classification and music recommendation by using deep learning

A. Elbir✉ and N. Aydin

Today, music is a very important and perhaps inseparable part of people's daily life. There are many genres of music and these genres are different from each other, resulting in people to have different preferences of music. As a result, it is an important and up-to-date issue to classify music and to recommend people new music in music listening applications and platforms. Classifying music by their genre is one of the most useful techniques used to solve this problem. There are a number of approaches for music classification and recommendation. One approach is based on the acoustic characteristics of music. In this study, a music genre classification system and music recommendation engine, which focuses on extracting representative features that have been obtained by a novel deep neural network model, have been proposed. Acoustic features extracted from these networks have been utilised for music genre classification and music recommendation on a data set.

Introduction: The common usage of the internet paved the way for substantial developments and changes in the music industry. One of the remarkable examples of these developments is the prevalent usage of online music streaming platforms. These platforms are mostly interested in the classification of music genre for mining of music listening and music tagging, music recommendation to increase sales profits and to control the music copyrights for authors. Especially, through music listening and broadcasting platforms such as Spotify, last FM, Fizzy, etc. people can access to millions of songs at any time, from anywhere. Most listening platforms offer music recommendation to their users by means of some methods such as meta-data analysis and collaborative filtering [1]. A music signal has acoustic features such as time, amplitude, phase and frequency. Since these acoustical features are distinctive, they can also be used for music recommendation systems. There are a number of experiments and previous work regarding this topic. Tzanetakis and Cook [1] applied time–frequency (TF) analysis methods such as Mell Frequency Cepstral Coefficient (MFCC), Spectral Centroid and wavelet transformations in feature extraction step on GTZAN data set, which is a benchmark data set used in music information retrieval music processing studies. By using these acoustic features Tzanetakis and co-workers classified the GTZAN with a 61% accuracy. Tzanetakis *et al.* obtained another score as 79.5% using Daubechies wavelet coefficient histogram as a best performing feature and support vector machine (SVM) as a classifier [2]. Using Gaussian mixture model by using three Gaussian components, an accuracy of 63.5% is presented by Li *et al.* [3]. Holzapfel and Stylianou [4] used non-negative matrix factorisation and reached 74% accuracy. Shin *et al.* [5] proposed auditory spice code in their study and they obtained 85% accuracy by using 172 features. In our previous study [6], a convolutional neural network (CNN) was designed and utilised to obtain acoustic features from raw music, spectrogram and mel-spectrogram of each music. By using these input types, an accuracy of 15, 66 and 63%, respectively were obtained. In addition to using CNN, we also used features obtained by TF analysis such as MFCC, Spectral Centroid, Chroma short-time fourier transform (STFT) etc. and wavelet analysis. Since window type, window size and overlap ratio are three important parameters of a TF analysis, we also examined the effects of these parameters over classification performance. According to the results, the maximum accuracy of the music genre classification was obtained by using SVM classifier and TF analysis with the Hanning window, a window size of 96 and overlap ratio of 0.75. Our first target in the previous study was to compare CNN and digital signal processing methods according to their parameters. Furthermore, our second aim of the study was to design a music recommendation engine by acoustic features to be obtained from music. After acquiring the best music genre classification results, the same parameters and features were utilised alongside the nearest neighbour methodology for music recommendation [6].

In this study, we proposed a music genre classifier and recommendation system based on signal processing and a CNN model named as MusicRecNet. The MusicRecNet has been designed to perform better than the previous classifier given in [5].

While other studies solely interested in music genre classification problem, we focused on both music genre classification and music recommendation. Music similarity and music recommendation applications are important for the listener, music sales platforms and authors. The proposed MusicRecNet helps a recommendation engine to recommend to the listener a new music according to his/her music taste. The proposed system can also be used to check whether the music is copied or not, i.e. detection of plagiarism. On the other hand, music broadcasting platforms desire to keep existing customers and to attract new customers, while guaranteeing copyright protection for music owners.

Data set: In order to evaluate MusicRecNet in terms of classification accuracy, the GTZAN data set [1] has been utilised. It contains 1000 music (sampling frequency 22,500 Hz, 16 bits resolution, and 30 s duration). Genres in the GTZAN are blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae and rock. Each genre contains 100 different examples.

Considering the time domain representation, some types of music can easily be distinguished from the others as illustrated in Fig. 1. Here, Jazz displays very distinctive characteristics than Rock and Metal music. However, some types of music can have very similar characteristics that make music classification a challenging task as illustrated again in Fig. 1. Here, comparing to Jazz, Rock and Metal are much more similar, hence more difficult to classify. Considering the frequency domain representation of the signals given in Fig. 1, signal characteristics in the frequency domain appear to be more distinctive as illustrated in Fig. 2. Therefore, more informative features can be utilised in classification process by considering both time and frequency domain representations.

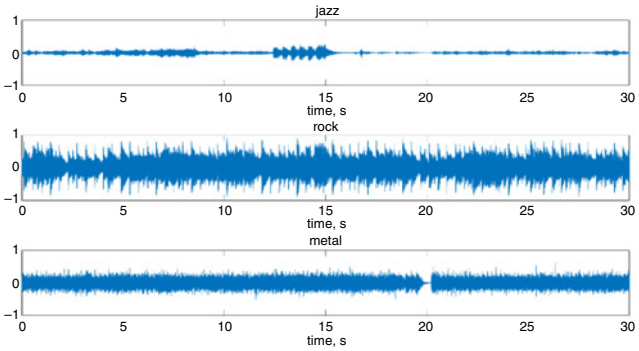

Fig. 1 *Example time domain representation of Jazz, Rock and Metal*

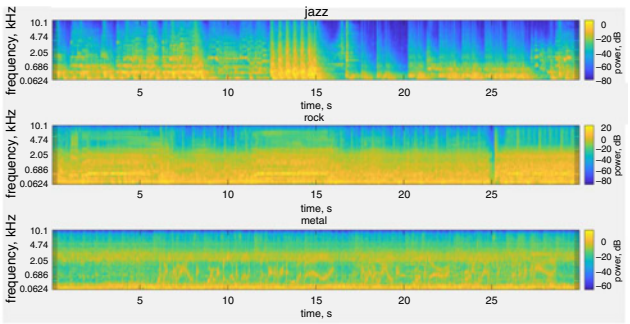

Fig. 2 *Time–frequency representations (spectrograms) of the signals given in Fig. 1*

Methodology: A flow chart of the conducted experiment is illustrated in Fig. 3. Initially, each music in the data set is divided into six parts with a duration of 5 s. Melspectrogram is generated from sampled each 5 s music and saved as an image. Then, this image is applied to the proposed MusicRecNet for training. The MusicRecNet, which is shown as the last block in Fig. 3, is a type of CNN that new layers and artificial dropout features have been added to minimise validation error. Specifically, in our experiments, MusicRecNet is designed to have three layers. Each layer consists of a two-dimensional convolution, an activation function (rectified linear unit), a two-dimensional maximum pooling operation and a dropout operation. The parameters used in the proposed MusicRecNet deep neural network model are given in Table 1.

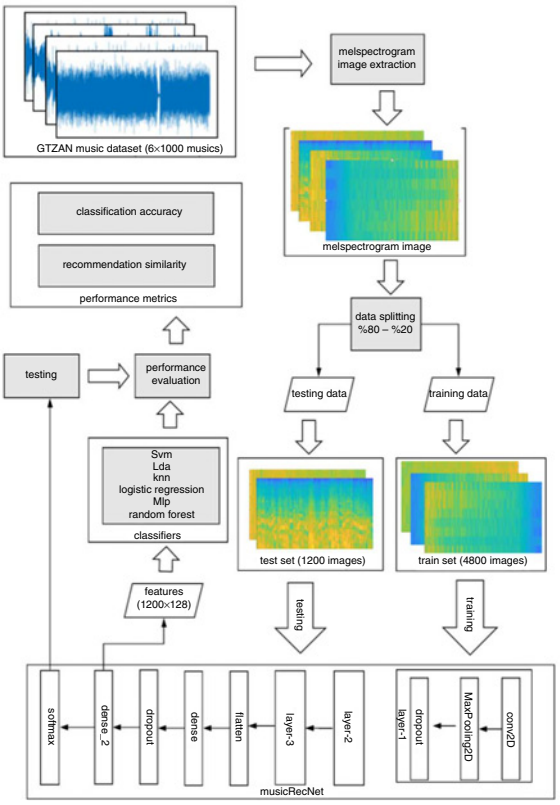

Fig. 3 *Block diagram of the proposed study*

Table 1: MusicRecNet parameters

Parameter type	MusicResNet
input size	128 × 128
batch size	64
kernel size	3 × 3
number of filters	32, 64, 128
loss function	cross entropy
activation function	SoftMax
optimiser	AdaDelta

After the training the MusicRecNet, the model is used for genre classification. Additionally, the last layer of the model named as Dense_2 is used as a feature vector of the test music samples for music genre classification, music similarity and music recommendation. We implemented classification algorithms, some of which are ensemble and model based methods given in Table 2.

Table 2: Classification performance accuracy

Classifier	Avg accuracy, %	
	Five-fold	Ten-fold
MLP	91.2	92.2
logistic regression	97.6	98.4
random forest	87.7	88.7
LDA	96.1	96.3
KNN	88.6	87.6
SVM(Poly-3)	97.6	97.9

Experimental results: Since GTZAN data set consists of ten different genres, accuracy was used as the main performance metric. Although it can be assumed as a subjective measure in view of a listener, the average percentage of music similarity is also used as a metric for quality of music recommendation. Additionally, a confusion matrix, from which precision, recall and *F*-measure scores can be obtained, is used. An example of the confusion matrix is illustrated in Fig. 4.

Mean accuracy of the proposed MusicRecNet is 81.8% when it is used as a standalone classifier. In Table 3, obtained MusicRecNet classification results are compared to the results of other studies. According to these results, it is obvious that the MusicRecNet outperformed the other classifiers.

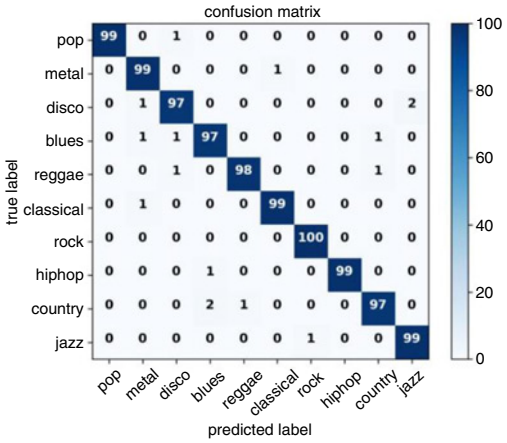

Fig. 4 *Example confusion matrix obtained from SVM*

Table 3: Comparative classification results obtained by MusicRecNet

Studies	Validation accuracy, %
Tzanetakis and Cook [1] and Tzanetakis [2]	61.0 and 79.5
Li and Tzanetakis [3]	74.0
Holzapfel and Stylianou [4]	63.5
Shin *et al.* [5]	84.5
Elbir and Aydin [6]	66.0
MusicRecNet	81.8
MusicRecNet + SVM	97.6

Table 2 summarises some music genre classification results using Dense-2 layer vector. As shown in the results, the classification accuracy increased substantially from 81% to over 90%. This increase in performance to employing classifiers given in Table 2 that are more advanced than the standard CNN SoftMax classifiers.

Fig. 5 shows mean percentages of the same genre recommendation by using the proposed genre classification system when the number of recommended songs are set to 5, 10 and 20, respectively. The best performance is obtained for classical music while the rock is the most difficult genre for the classification and recommendation. The reason is that the rock displays some characteristics similar to other genres such as metal.

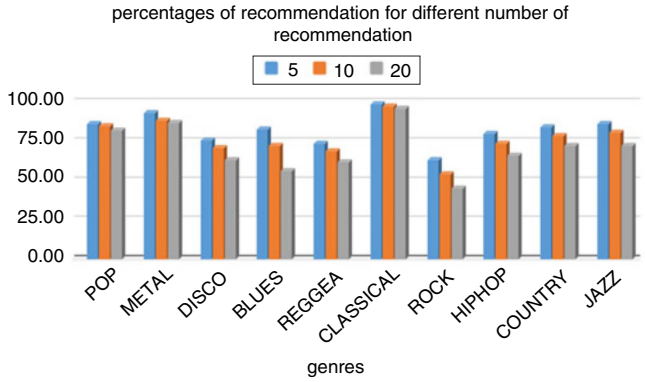

Fig. 5 *Recommendation demo results obtained by MusicRecNet*

Conclusion: In conclusion, the proposed MusicRecNet model has shown improved performance in terms of music genre classification, music similarity and music recommendation, compared to previous studies. In particular, adding some dropout layers into CNN have provided much better results. Owing to the fact that the MusicRecNet has only three layers, it is an appropriate model to employ in the music streaming applications for music similarity and recommendation. When the performance results are examined, some similar music genres can lead to mis-classification and mis-recommendation such as Jazz and Classic. In future studies, in order to improve current results, we plan to design more comprehensive deep neural network models and to add extra data models as an input in addition to using only melspectrogram. Big data processing techniques and tools can also be utilised for feature extraction and model creation in music genre recommendation systems.

Acknowledgment: This research has been supported by the TUBITAK-TEYDEB-1505 Program (Project no: 5180069).

Submitted: *02 January 2020* E-first: *19 March 2020*
doi: 10.1049/el.2019.4202
One or more of the Figures in this Letter are available in colour online.

A. Elbir and N. Aydin (*Department of Computer Engineering, Yildiz Technical University, Istanbul, Turkey*)

✉ E-mail: aelbir@yildiz.edu.tr

References

1 Tzanetakis, G., and Cook, P.: 'Musical genre classification of audio signal', *IEEE Trans. Speech Audio Process.*, 2002, **10**, (3), pp. 293–302
2 Li, T., and Tzanetakis, G.: 'Factors in automatic musical genre classification of audio signals'. Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Platz, NY, USA, 19–22 October 2003
3 Li, T., Ogihara, M., and Li, Q.: 'A comparative study on content-based music genre classification'. Proc. 26th ACM SIGIR Conf., Toronto, ON, Canada, 2003, pp. 282–289
4 Holzapfel, A., and Stylianou, Y.: 'Musical genre classification using non-negative matrix factorization-based features', *IEEE Trans. Audio Speech Lang. Process.*, 2008, **16**, (2), pp. 424–434
5 Shin, S.-H., Yun, H.-W., Jang, W.-J., *et al*.: 'Extraction of acoustic features based on auditory spike code and its application to music genre classification', *IET Signal Process.*, 2019, **13**, (2), pp. 230–234
6 Elbir, A., and Aydin, N.: 'Music genre classification and recommendation by using machine learning and deep learning'. 2018 Innovations in Intelligent Systems and Applications Conf. (ASYU), 2018, Adana, Turkey, 2018, pp. 1–5