International Journal of Music Science, Technology and Art
(IJMSTA) 2 (1): 10-18, Music Academy "Studio Musica", 2020 ISSN: 2612-2146 (Online)

Evaluation of Deep Learning and Data Mining Techniques for Audio Data Stream Classification

Hayder K. Fatlawi [1,2]✉, Attila Kiss [1]

[1] Eötvös Loránd University, Budapest, Hungary, Faculty of Informatics, Department of Information Systems
[2] University of Kufa, Najaf, Iraq, Information Technology Research and Development Center
✉ Corresponding author: hayder.fatlawi@uokufa.edu.iq

ARTICLE INFO

Received: March 16, 2019
Accepted: April 10, 2020
Published: April 16, 2020

Keywords:
Audio Stream
Convolution Neural Network
Random Forest
Decision Tree
Bayesian model
SVM
Music Genres

ABSTRACT

Classification of audio streams into a meaningful category like music genres has increasing research interest due to the need of the multimedia websites categorization and user profiling. Audio signals require many of preprocessing and features extraction operations. In this paper, two types of data mining classifiers (batch and stream) has been used with 40 extracted features while CNN deep learning classifier utilized to classify images that obtained from waveform and spectrogram representations of music data. The results showed that Random Forest classifier had the best performance in both batch and stream classification with accuracy 71% and 74.6% respectively.

1 Introduction

The increasing growth of the internet produces spreading millions of music tracks which make the organizing and categorizing those files an important need for more fast and efficient access. Music genres represent the conventional method for music categorization although some genres can be overlapped. This method is so popular in multimedia websites, digital media stores, and online radio stations. The need for classifying a new music track in a suitable genre and suggest the most related tracks for the users (profiling) increased the interest of the researchers for developing a music classification model based on machine learning techniques.

Classification is one of the major tasks of machine learning that aims to label data into previously defined categories, it generally has three main phases; the data is prepared to training process then machine learning techniques are applying for knowledge extraction and finally the results will evaluated and processed in understandable form for decision making [1].

1.1 Audio Data Features

Basically, digital audio can be represented by a sequence of quantized pulses in time. The type of required information they are attempting to extract from the audio signal specifies the grouping audio features. In this work, three types of features are extracted; (1) Spectral Features (2) Mel-Frequency Cepstral Coefficients (3) Pitch/Harmony Features [2].

- **Spectral Features.** It related to magnitude spectrum and represents summarized information about energy distribution across frequency. An important feature from this type is the spectral centroid that considered as the gravity center the spectrum [2].
- **Mel-Frequency Cepstral Coefficients (MFCCs).** are used widely in speech recognition and music analysis. Those Coefficients require three steps; the first is calculating Mel-scale filter bank using Fourier transformation, second step is to find

the Log energy by calculating the logarithm of the magnitude of filter bank outputs. Finally, the dimensionality of the filter bank outputs is reduced using discrete cosine transform [2].

- **Pitch/Harmony Features.** it is related with underlying harmony of music piece and aim to find occurrence of specific musical pitches in a segment of music. It can be calculated by mapping and folding all the magnitude bins of a fast Fourier transform, and this process is called chroma or chromagram [3].

1.2 Audio Features Extraction

It aims to capture the higher-level information of underlying musical content by filtering the large amounts of raw data into more compact representations [music mining]. The representations of sound signals could have a separate notion of time and frequency. Many of common audio representation could be used for features extraction such as short-time Fourier transform and wavelets.

Short-time Fourier transform (STFT) produces elementary signals as a linear combination from an original signal to be in more understandable and manipulated form. It also expresses the energy distribution of the signal in the time domain and frequency domain. The STFT is an adoption from a discrete Fourier transform (DFT) for providing localization in time [2].

Wavelet representation is formed by using variable time-frequency resolution. It includes high precisely detection for low frequencies which are not placed in time very accurately. In the other hand, high frequencies have lower precisely detection and more accurate placing in time. The discrete wavelet transform has same number of coefficients of the original discrete signal [2].

1.3 Classification Techniques

The input of a classification task is a collection of data records. Each record is represented by a tuple (x, y), where x is the features set and y is the target of classification. The type of target attribute y detects the required technique; if values of y are binary, binary class classification techniques can be used, multiclass classification is used if y has more than two categories. If values of y are continuous, regression techniques should be used for predictive task [1]. In this work, music genres contain more than two categories therefore it requires a multiclass classification task. Batch and online classification techniques are used in this work including Decision tree, Bayesian Model, Rule based, SVM, and Random forest in addition to CNN classifier.

Batch Data Mining Classifiers

In this type of classifiers, the data is treated as a batch in which all data sampled are available as a file before building the classifiers. Four batch classifiers are used in this works as follow:

- **Decision Tree DT** represents by a flowchart-like tree structure, where every non terminal node denotes a condition on an attribute, to split data records. Each branch represents the result of that condition, and each leaf node (i.e. terminal node) holds a class label [4].
- **Random Forest** is one of bagging ensemble classifiers that combine a set of single models for obtaining a better integrated model. Bagging combines the decisions of multiple base classifiers using voting concept. Random Forest uses randomness in two steps as follows: (i) chooses data instances of for each single model (ii) chooses a subset of features for each node [4].
- **Bayesian Classifier** is a probabilistic model for performing classification tasks. In this classifier, the features (F1,F2,…,Fn) of each data record including the class label are considered as a random variables. The aim is to find the value of class Y that maximizes the conditional probability P(C| F1, F2,…,Fn) [1].
- **Support Vector Machine.** In this classifies, the margin among closest points of the classes is maximized for obtaining optimal separating hyper plane between them. SVM is designed to deal with binary classification tasks, so for multiclass classification, it utilizes one-versus-one method [5].

Stream Data Mining Classifiers

Mining of data stream led to developing many classification techniques. Hoeffding Tree or Very Fast Decision Tree (VFDT) is a classification technique depends on replacing the terminal nodes with decision nodes based on a statistic evaluation [6]. Oza Bagging method is an online bagging strategy which simulates bagging task by training every arriving data sample K times [7]. Adaptive Random Forest is another bagging techniques that utilizes Hoeffding tree for adapting with distribution changes and Adaptive Window ADWIN for detecting the change is a data stream [8].

Deep Learning Convolutional Neural Networks CNN.

It represents a neural network-based classification technique in which the input data is passed into a group of pairs contain convolutional and pooling layer. Output of the last layer is considered as input into another group of fully connected layers that pass its output to a SoftMax layer. Choosing the suitable architecture of CNN has dependency on the type of classification problem, so the network may be built in different ways [9].

1.4 Related Works

In [10] a comparison was made among deep neural networks and set of typical learning models like SVM and logistic regression, music data was preprocessed using Spectrogram and MFCC and the best accuracy was 38.8%. Utilization of cepstral modulation spectrum is presented by [11] for building a deep neural network (DNN). GTZAN music dataset is used in this research and its result showed that the temporal features can obtain classification accuracy comparable or better than the spectral features.

A comparison between Convolutional Neural Networks CNN and Supported Vector Machine SVM is presented by [12] for Music Classification. The evaluation is performed using three different music datasets, and the results showed a significant performance of CNN with two datasets.

In [13] the performance of CNN learning technique based on the images of the spectrogram of audio signal is compared with four typical classifiers which were trained base on hand-crafted features. The result of this work showed that CNN had better accuracy (64%) compared with Extreme Gradient Boosting XGB (59%) and other three typical classifiers, also the combination between CNN and XGB as an Ensemble Classifiers reached to (65%) accuracy.

2 Methodology

Classification of music genres in this work divides into two major independent parts; the first part includes extracting the important features from audio signals after applying Fast Fourier Transformation FFT, then four classification techniques (Decision Tree, Bayes Model, SVM, and Random Forest) are trained. The evaluation of classifiers performance is performed by five popular metrics (accuracy, precision, Recall, F-measure, and ROC).

The second part starts with represent the audio signal as an image using wave form representation and spectrogram representation. Images that produced by wave form will be converted to binary images while the RGB image that was produced from spectrogram representation is split into three different images. CNN classification model is applied on the resulted images, and finally the performance is evaluated using validation accuracy.

2.1 Shorter Form Features Analysis

The first step after gathering music data is to represent it in a shorter form. This will ensure getting the highest valuable information and reduce the required resources for the classification process. This reduction can be performed by applying FFT on the audio signals; thereby the important features can be extracted. For stream data mining, data records are streamed for the classifiers according to a specific frequency

For each feature (Chroma and spectral), mean and standard deviation are calculated to summarize the values of a music track in addition to 10 features from MFCCs. In the end of this step, each music track is converted to an instance with 40 features and class label value. These values represent the required dataset for applying classification techniques.

After preparing batch and stream data from extracted features, eight classifiers are applied, four for each type. Decision tree is generated by using multi branch J48 algorithm. It produces a pruned decision tree based on information gain metric, (2) Bayesian classifier is built using a hill climbing as a search algorithm (3) Support Vector classifier is built using SMO implementation in which all features are normalized (4) Random forest is generated with no limitation used for random tree depth. For performance evaluation, 10 cross-validation method is used to distribute data between training and testing process and five metrics are used (accuracy, precision, Recall, F-measure, and ROC).

2.2 Shorter Form Features Analysis

Handling the sound signal as an image can facilitate the classification task by utilizing all image classification tools and techniques. Artificial Neural Networks and specially the Convolution Neural Networks have an interested performance for image classification, therefore for each music track, two images are created; the first by wave form and the other using spectrogram.
Wave Form Representation. The wave form resulted image has one color for representing waves in addition to white background color as can be seen in figure (2, a). Because the color of waves isn't meaningful, it can be converted to black color, then, the whole image can be represented as a binary image from 1 and 0 values.

- **Spectrogram Representation.** Color image with three channels (RGB) is resulted from Spectrogram and for more precise classification, each channel is separated as a single gray level image which contains s color values in range 0-255. Threshold is needed to covert gray level image to binary image and 128 is chosen in which every color value less than 128 will considered as 0 and every color value more than or equal 128 will considered as 1.

The resulted images from both representations feed into two separated CNNs with same architecture for both networks. It starts with four 2D convolution layers; each layer is represented by 2*2 array with Rectified Linear Unit as an activation

function. Then, 2D max pooling layer with 2*2 pool size is added to the network for reducing the number of samples of the previous layer therefore the next operations will be minimized. Two Dense layers have been added to the architecture of CNN with a Flatten layer between them. The first Dense layer contains 8 neurons and uses Rectified Linear Unit as an activation function. The second Dense layer uses softmax Unit as an activation function and represents the last layer, therefore it contains 10 neurons for 10 classes of music genres.

3 Implementation and Results

Music classification methodology is implemented using many of programming and data analysis tools. A standard music dataset is used also for this implementation. Figure 1 illustrates all the steps of methodology and implementation.

3.1 Description of Dataset

The evaluation of methodology requires testing its performance on real dataset; therefore, a popular dataset is chosen. GITZAN music dataset [14] contains 1000 data rows which distributed in equal among 10 music genres (100 data rows for each genre). It is used in many of music classification related works.

3.2 Frameworks and Tools

The implementation of music classification methodology included utilizing a wide and various range of tools. Two different environments are used to this task; the first one was on a PC with windows 10 operating system, Core i5 1.8 GHz processor, and 4 Gigabytes of RAM). The second environment was an AWS Amazon server with Windows Server 2016 operating system, 8 vCPU Xeon 2.5 GHz, 32 Gigabytes RAM).

Python [15] programming language with many of its libraries alongside with Weka [16] are used. Anaconda [17] with Spider editor is chosen as the programming framework in both environments. For reading the music data from files, Librosa [18] library is used. It is also used for features extraction in the first environment and for representing the music as wave form and spectrogram in the second environment.

Applying FFT in the first environment is performed using SciPy [19] python library. OpenCV library package is used for generating the binary images from wave form images and for splitting RGB images of spectrogram representation. Tensorflow [20] and Keras [21] are used together for applying CNN classification technique on the images that resulted from both wave form and spectrogram representations. Figure 1 illustrates all the steps of the implementation of music classification methodology.

MOA Platform [22] is an improvement for Weka platform for mining data stream. It provides many of popular mining techniques, stream generator, and concept drift detection techniques, in our comparison, it performed the data streaming and implementation of stream classification techniques.

3.3 Results

Many experiments are performed to validate the performance of the methodology based on GITZAN dataset. In shorter form features analysis, after applying FFT on the music data, 10 chroma and spectral features are extracted. Mean and standard deviation are used for each feature to construct 20 features, in addition to 20 MFCCs features, we get 40 features and the class label as shown in Table 1.

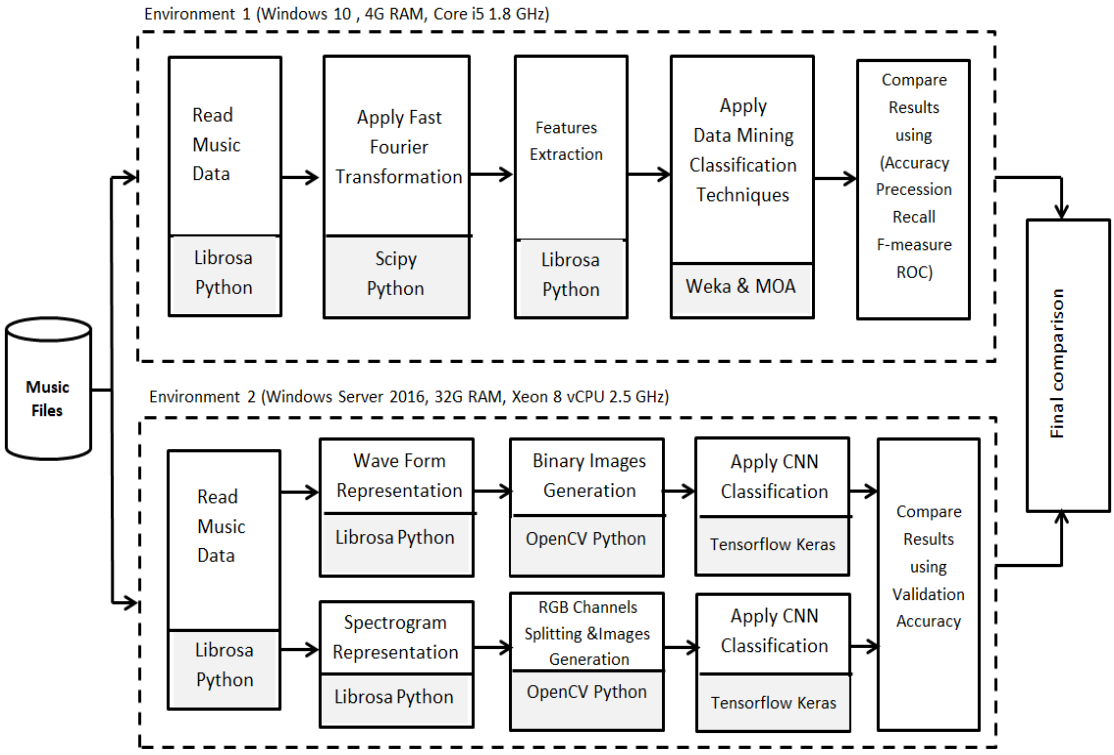

Figure 1. Block Diagram of Music Classification Methodology and Implementation.

In the first experiment, we use only mean to obtain 30 features, and we made a comparison between the performance of the four classifiers with and without using FFT before features extraction. Five metrics are used in this comparison, and the results in Table 2 showed that the performance of all classifiers is better without applying FFT. The best classification is achieved by Random forest with 0.69 accuracy.

Table 1. Description of dataset features extracted.

Feature Name	Feature Description	Feature Name	Feature Description
F1-F2	Chroma shift	F13-F14	Spectral contrast
F3-F4	Spectral centroid	F15-F16	Spectral flatness
F5-F6	Spectral bandwidth	F17-F18	Chroma cqt
F7-F8	Rolloff mean	F19-F20	Chroma Energy
F9-F10	Zero crossing rate	F21-F40	Mfcc1 - Mfcc20
F11-F12	Roof mean square error	Music Genre	Class Label

In the second experiment, the full set of 40 features is used, and the performance of all classifiers is improved. Accuracy of Random forest reached to 0.71 after 220 iterations as shown in Table 3 and Figure 2.

From Confusion Matrix of Random Forest Classifier in Table 4 we can recognize different classification accuracy for music genres. The best result reached to 89 instances classified correctly in classical genre while with rock genre only 41 instances classified correctly.

In both experiments, the four stream classifiers had mostly same results. The best results obtained by adaptive random tree was 74.66% in the average of accuracy followed by random hoeffding tree with 74.41% as shown in Figure 4 and Figure 5.

Another two experiments are performed for two representations of music signals as an image, one using waveform and the other using spectrogram. CNN classifier which explained in section 2.2 is applied on waveform images (Figure 6.a illustrates waveform of Blues Music track), 700 images are used for training and 300 for validation through 16 iterations. Training accuracy reached to the best value 1 (in 0-1 range) after 10 iterations while the best validation accuracy has been obtained 0.45 in iteration 11 as shown in Figure 3. The last experiment was with 3000 Spectrogram images that resulted from splitting the three channels of RGB images, Figure 6.b, Figure 6.c, Figure 6.d and Figure 6.e illustrate Spectrogram images of music signals

before and after channels splitting. 2100 images are used to training CNN classifier; the remaining images are used for validation. Five iterations are used in this experiment; the best training accuracy was 0.974 while best validation accuracy was 0.505 as shown in Figure 7. The comparison among all the five classifiers that used in this work showed that Random Forest had best performance followed by SVM as shown in Figure 8.

Table 2. Evaluation of four classifiers with and without using FFT for audio music files using 30 extracted features.

Classifier	Accuracy		Precision		Recall		F-Measure		ROC Area	
	With FFT	Without FFT	With FFT	Without FFT	With FFT	Without FFT	With FFT	Without FFT	With FFT	Without FFT
DT	0.190	0.491	-	0.492	0.190	0.491	-	0.491	0.641	0.741
Bayesian	0.387	0.452	0.372	0.432	0.387	0.452	0.369	0.425	0.787	0.865
SVM	0.416	0.641	0.423	0.637	0.416	0.641	0.414	0.636	0.804	0.902
Random forest	0.474	0.692	0.458	0.689	0.474	0.692	0.464	0.689	0.850	0.936

Table 3. Evaluation of four classifiers with and without using FFT for audio music files using 40 extracted features.

Classifier	Accuracy		Precision		Recall		F-Measure		ROC Area	
	With FFT	Without FFT	With FFT	Without FFT	With FFT	Without FFT	With FFT	Without FFT	With FFT	Without FFT
DT	0.429	0.522	0.424	0.522	0.429	0.522	0.426	0.520	0.702	0.754
Bayesian	0.440	0.506	0.421	0.492	0.440	0.506	0.419	0.489	0.832	0.888
SVM	0.514	0.700	0.504	0.702	0.514	0.700	0.498	0.697	0.851	0.926
Random forest	0.558	0.713	0.544	0.711	0.558	0.713	0.546	0.709	0.896	0.949

Table 4. Confusion Matrix of Random Forest Classifier.

Genre	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock
blues	83	0	2	1	1	3	5	0	2	3
classical	0	89	5	1	0	3	0	0	1	1
country	2	0	68	4	0	11	1	3	4	7
disco	6	0	3	68	9	1	3	1	3	6
hiphop	2	0	0	7	63	0	6	4	17	1
jazz	1	6	5	3	0	79	1	2	3	0
metal	3	0	0	2	6	1	84	0	0	4
pop	0	0	5	3	3	3	0	80	5	1
reggae	2	1	6	6	11	2	0	6	61	5
rock	7	0	12	15	1	3	6	5	10	41

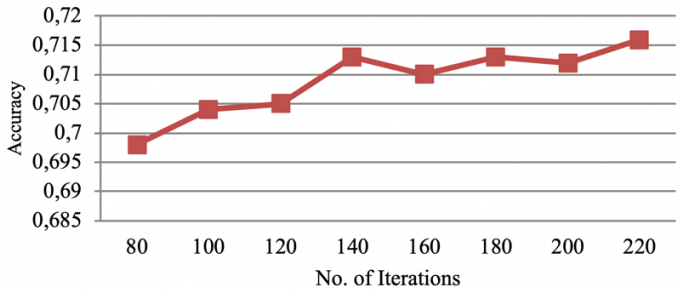

Figure 2. Accuracy of Random Forest Classifier with variable number of iterations.

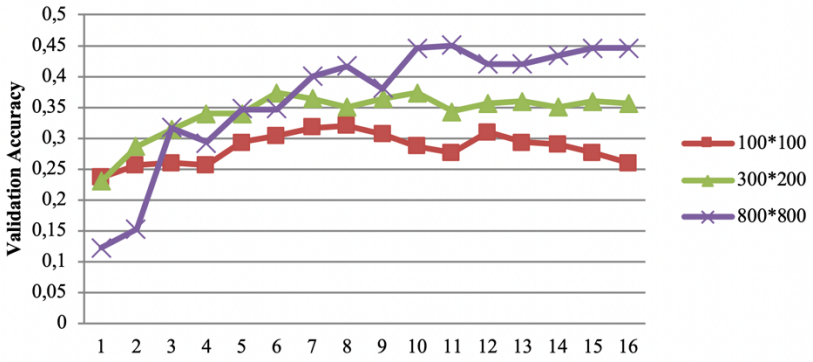

Figure 3 CNN Validation Accuracy for three different sizes of waveform images.

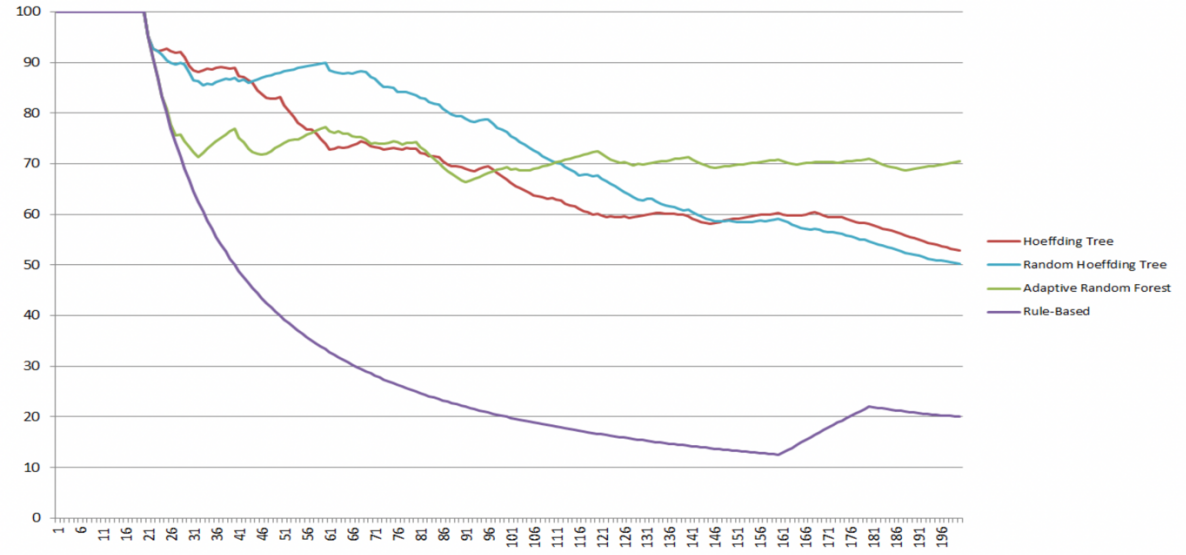

Figure 5. Comparisons of stream classifiers performance with audio music streams without FFT.

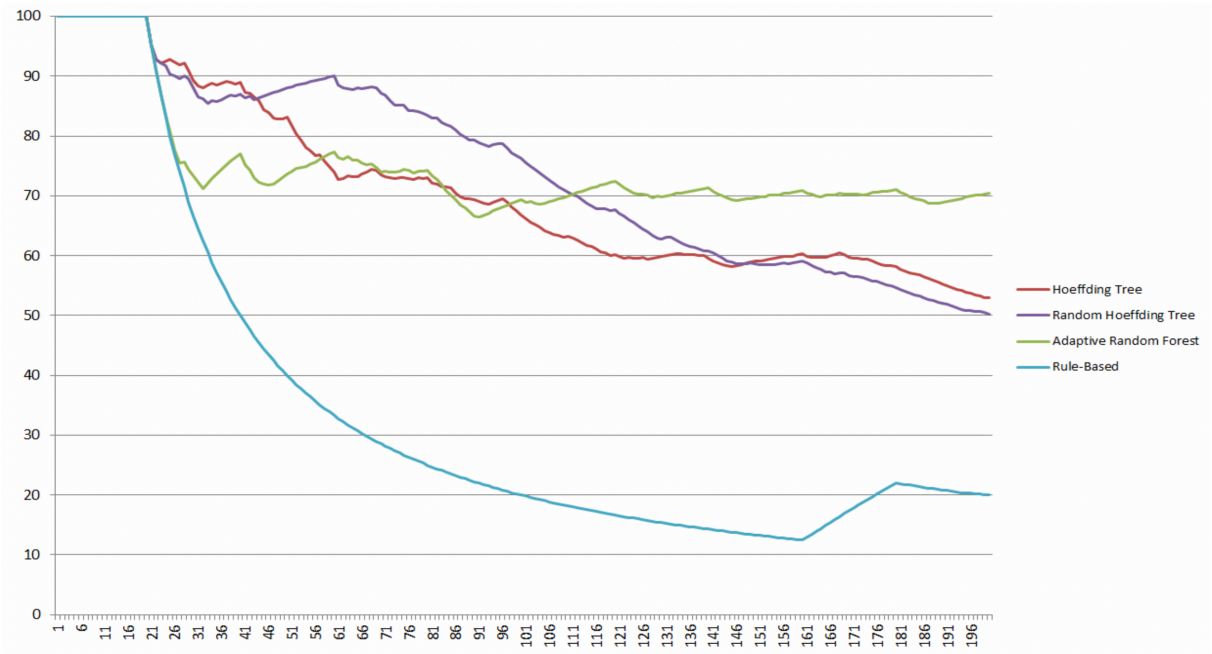

Figure 6. Comparisons of stream classifiers performance with audio music streams with FFT.

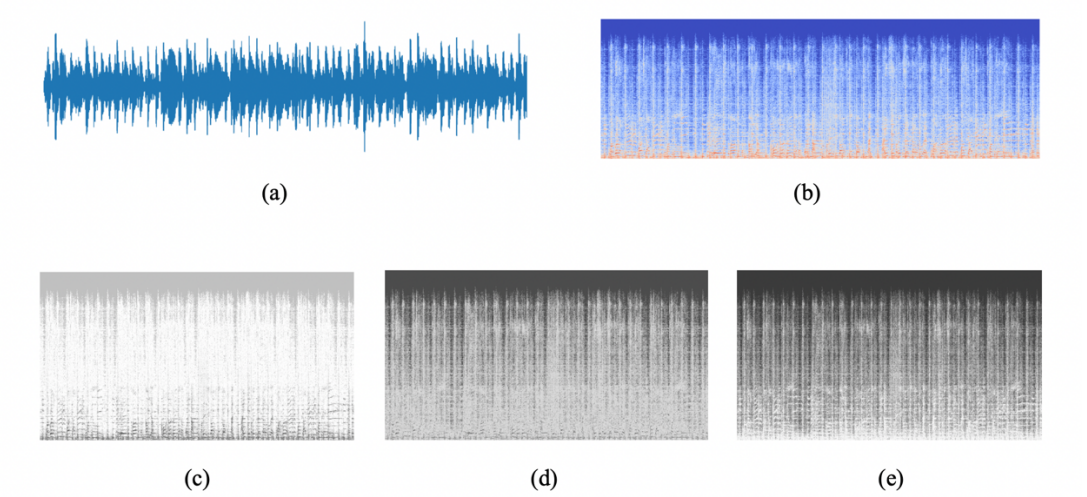

Figure 6. Blues Music track representation in (a) wave form (b) Spectrogram as three channels image (c,d,e) Spectrogram as three images for three channels.

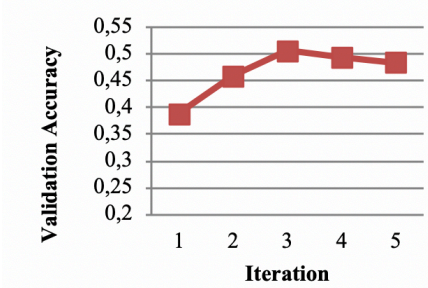

Figure 7. CNN validation accuracy for Spectrogram images

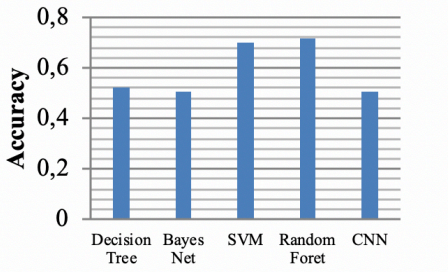

Figure 8. Comparison of validation accuracy among five classifiers

4 Conclusion

Music data as a part from the digital content need to be classified and genre of music can represent a target of this classification. This work aims to compare the performance of typical classifiers that trained on 40 audio extracted features with CNN classifier which trained on images that obtained from waveform and Spectrogram representations of music. From the results we can conclude that the performance of all four classifiers is improved by increasing number of extracted features. The performance of CNN is enhanced by increasing number of trained data rows. The limitation of the available resources led to simplify CNN architecture in which the validation accuracy didn't exceed 50% although the training accuracy reached to 100%. Also, increasing the resolution of images had a preferred impact on CNN results. Finally, Random Forest had best accuracy in both batch and stream classification with accuracy 71% and 74.6% respectively, it followed by SVM classifier in batch classification with 70% and random hoeffding tree in stream classification with 74.4 accuracy in average.

5 Acknowledgment

The project has been supported by the European Union, co-financed by the European Social Fund (EFOP-3.6.3-VEKOP-16-2017-00002).

References

[1] Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. ISBN-13: 978-0321321367, Addison-Wesley (2005).
[2] Tao, L., Ogihara, M.: Music data mining. CRC Press (2011).
[3] Mark, A., Wakefield, G.H.: To catch a chorus: Using chroma-based representations for audio thumbnailing. Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics. IEEE (2001).

[4] Al-Fatlawi, Ali H., Hayder K. Fatlawi, and Sai Ho Ling. : Recognition physical activities with optimal number of wearable sensors using data mining algorithms and deep belief network. 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE (2017).
[5] Meyer, David, and FH Technikum Wien.: Support vector machines The Interface to libsvm in package (2015).
[6] Gama, J.: Knowledge discovery from data streams. Chapman and Hall/CRC, 2010.
[7] Nikunj C Oza: “Online bagging and boosting". In: 2005 IEEE international conference on systems, man and cybernetics. Vol. 3. Ieee. 2005, pp. 2340-2345.
[8] Babenko, B., Yang, M.H., Belongie, S.: “A family of online boosting algorithms". In: 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops. IEEE. 2009, pp. 1346-1353.
[9] Venkatesan, Ragav, and Baoxin Li.: Convolutional Neural Networks in Visual Computing: A Concise Guide. CRC Press. (2017).
[10] Rajanna, Arjun Raj, et al.: Deep neural networks: A case study for music genre classification. 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA). IEEE. (2015).
[11] Jeong, Il-Young, and Kyogu Lee.: Learning Temporal Features Using a Deep Neural Network and its Application to Music Genre Classification. Ismir. (2016).
[12] Costa, Yandre MG, Luiz S. Oliveira, and Carlos N. Silla Jr. : An evaluation of convolutional neural networks for music classification using spectrograms. Applied soft computing 52. 28-38 (2017).
[13] Bahuleyan, Hareesh.: Music genre classification using machine learning techniques. arXiv preprint arXiv:1804.01149. (2018).
[14] http://opihi.cs.uvic.ca/sound/genres.tar.gz, access date: 02-04-2019.
[15] https://www.python.org/, access date: 01-04-2019.
[16] https://www.cs.waikato.ac.nz/~ml/weka/, access date: 02-04-2019.
[17] https://www.anaconda.com/, access date: 02-04-2019.
[18] https://librosa.github.io/librosa/, access date: 03-04-2019.
[19] https://www.scipy.org/, access date: 03-04-2019.
[20] https://www.tensorflow.org/, access date: 03-04-2019.
[21] http://keras.io/, access date: 03-04-2019.
[22] https://moa.cms.waikato.ac.nz.