Music Genre Classification Using Acoustic Features and Autoencoders

1st Yunus Atahan
Bilgisayar Mühendisliği Bölümü
Yildiz Teknik Universitesi
Istanbul, Turkiye
yunus10atahan@gmail.com

2nd Ahmet Elbir
Bilgisayar Mühendisliği Bölümü
Yildiz Teknik Universitesi
Istanbul, Turkiye
aelbir@yildiz.edu.tr

3rd Abdullah Enes Keskin
Bilgisayar Mühendisliği Bölümü
Yildiz Teknik Universitesi
Istanbul, Turkiye
aekeskin96@gmail.com

4th Osman Kiraz
Bilgi ve Iletisim Teknolojileri
Turkcell
Istanbul, Turkiye
osman.kiraz@turkcell.com

5th Bulent Kirval
Bilgi ve Iletisim Teknolojileri
Turkcell
Istanbul, Turkiye
bulent.kirval@yildiz.edu.tr

6th Nizamettin Aydin
Bilgisayar Mühendisliği Bolumu
Yildiz Teknik Universitesi
Istanbul, Turkiye
naydin@yildiz.edu.tr

2021 Innovations in Intelligent Systems and Applications Conference (ASYU) | 978-1-6654-3405-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ASYU52992.2021.9598979

Özetçe **—Müzik öneri ve sınıflanırma sistemleri sayısal işaret işleme ve sayısal müzik işlemenin ilgilendiği bir alandır. Bu çalışmada dijital sinyal işleme yöntemleri ve otomatik kodlayıcı kullanılarak müziğin özellikleri çıkarılmış ve bu özellikler kullanılarak müzik türü sınıflandırması ve kümelemesi yapılmış, elde edilen sonuçlar ile müzik önerisi yapılmıştır. Elde edilen sonuçlar birbiri ile karşılaştırılmıştır. Çalışmada GTZAN veri seti kullanılmıştır. Amaç otomatik kodlayıcılar ile özellik çıkarımı yapıldığında elde edilen sonuçların dijital sinyal işleme yöntemleri ile karşılaştırılmasıdır. Dijital sinyal işleme için Mel Frekansı Kepstral Katsayıları (MFCC) ve türevi, sıfıra değme noktası (Zero Crossing Rate), spektral bant genişliği (Spectral Bandwidth), spektral etek (Spectal Rolloff), spektral ağırlık merkezi (Spectral Centroid), rms, polinom özelliği, spektral zıtlık (Spectral Contrast), spektral düzlük (Spectral Flatness), sabit q dönüşümü (CQT), kısa zamanlı Fourrier transformu (STFT), tonnetz, dalgacık dönüşümü (Wavelet) yöntemleri kullanılmıştır. Otomatik kodlayıcı için ise yalnızca MFCC özellikleri kullanılmıştır. Sınıflandırma yöntemleri olarak ise MLP, lojistik regresyon, Rassal Orman (Random Forest), Lineer Diskriminant Analizi, K-NN, SVM, Naive Bayes ve Boosting Algoritmaları algoritmaları kullanılmıştır.**

Anahtar Kelimeler—Müzik öneri sistemleri, Müzik türü sınıflandırma, GTZAN veri seti, dijital sinyal işleme, Kümeleme, Otomatik kodlayıcı.

***Abstract*—Music recommendation and classification systems are an area of interest of digital signal processing and digital music processing. In this study by using digital signal processing techniques and autoencoders, music features are extracted and then with these features music classification and clustering has been done, and with the results music recommendation has been made. Obtained results are compared with each other. In the study, GTZAN dataset has been used. Purpose of this study is to compare the result feature extraction with auto encoders and digital signal processing techniques. For digital signal processing, used methods are as following: Mel Frequency Cepstral Coefficients (MFCC) and it's derivative, Zero Crossing Rate, Spectral Bandwidth, Spectral Rolloff, Spectral Centroid, Spectral Contrast, Spectral Flatness, RMS (Root Mean Square Energy), poly features, Chroma CENS, Chroma CQT, Chroma STFT, tonnetz, Wavelet etc. For the classification part MLP Classifier, Logistic Regression, Random Forest Classifier, Linear Discriminant Analysis, K-Neighbors Classifier, SVM, Naive Bayes, Gradient Boosting Classifier, Ada Boost Classifier used for classifying the data.**

Keywords—Music recommendation systems, Music genre classification, GTZAN data set, Digital signal processing, Clustering, Autoencoder.

I. Introduction

Music recommendation systems are an area that has gained importance with the spread of the Internet and the increase in the number of people using music platforms. With these possibilities there are thousands of music published every year. Music platforms want to increase the attention so that people spend more time on the platform. One way for that is recommending music that user might like according to the previous musics the user listened to. In this study, the purpose is to make music recommendation system with music genre classification and clustering using features that are obtained by using auto encoder and digital signal processing techniques. This study is focused on content based classification methods and autoencoders. In this study it is aimed to compare content based methods and autoencoders classification results.

For content based classification each music has different acoustic features like rhythm, timbre, density and pitch. These features is somewhat similar for same genres. We can classify music pieces by extracting acoustic features and analyzing them.

To train the autoencoder Mel Frequency Cepstral Coefficients (MFCC) is used. Latent space vectors that are created using trained autoencoder are used as features for classification and clustering. In autoencoder, Convolutional Neural Network (CNN) architecture used.

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

To see and compare the results GTZAN dataset used. In the second section method used for extracting MFCC's, music genre classification, music clustering and music recommendation has been discussed. Experimental results shown in the third section.

II. METHODS

A. Dataset

In this study to evaluate the results GTZAN dataset [1] has been used. It contains 10 different genres and 100 musics for each genre and is used for obtaining and comparing results in music processing studies. Each music is 30 seconds long. Dataset consists of only raw audio, no metadata is given.

B. Methodology

1) Digital Signal Processing: In this section, the methods of feature extraction used in the study are explained in detail. The GTZAN data set was used for digital signal processing. Wavelet method extracted by PyWavelet while other features extracted by librosa library. In order to obtain feature vector some statistical descriptors such as mean, standart deviation, skewness, kurtosis and median has been used. The number of features extracted from raw music data is 606.

a) Zero Crossing Rate: Zero Crossing Rate is the ratio of the number of moments of signal change from positive to negative or negative to positive to the time. This method is usually one of the first preferred feature extraction techniques in speech recognition systems and MIR (Music Information Retrieval) studies [2].

b) Spectral Centroid: Spectral Centroid indicates which frequency value of the signal is at the center point for its energy [2].

c) Spectral Contrast: For each sub-bands of a spectrogram, the energy contrast calculated by comparing the peak energy and the valley energy [3].

$$Peak_k = log\{\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} x_{k,i}\}$$

$$Valley_k = log\{\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} x_{k,N-i+1}\}$$

$$SC_k = Peak_k - Valley_k$$

d) Spectral Flatness: Spectral Flatness is a measure to determine how much noise-like is the given sound. When the values get close to 1, indicates that the spectrum of sound is similar to white noise [3].

$$\text{Spectral Flatness} = \frac{\sqrt[N]{\prod_k x(k)}}{\frac{1}{N}\sum_k x(k)}$$

e) Spectral Bandwidth: Spectral bandwidth gives the extent of the power transfer function around the center frequency for given signal [3].

$$\text{Spectral Bandwidth} = (\sum_k S(k)(f(k) - f_c)^p)^{\frac{1}{p}}$$

$S(k) \rightarrow$ spectral amplitute of k-th frequency bin
$f(k) \rightarrow$ frequency of k-th bin
$fc \rightarrow$ Spectral Centroid

f) Spectral Rolloff: The rolling frequency is defined as the frequency at which a certain percentage of the total energy of the spectrum is found [3].

$$\sum_{n=1}^{R_i} M_i[N] = 0.85\ x\ \sum_{n=1}^{N} M_i[N]$$

g) Root Mean Square Energy - RMSE: Root Mean Square Energy gives the total magnitude of the signal. For audio signals, it can be said that it roughly calculates the degree of loudness [1].

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_n |x(n)|^2}$$

h) Mel Frequency Cepstral coefficients - MFCC: Basically, MFCC is one of the mostly used feature extraction method for speech and audio signals. Cepstral coefficients are linear. However, the human senses the frequencies below 1KHz in linear scale and the above frequencies in logarithmic scale. The purpose of MFCC is to make cepstral coefficients compatible with human hearing system. [2]

i) Chroma STFT: The Pitch Class Vector (Chroma) is the vector that shows how much energy the music signal has for each of the 12 different pitch classes. With the Chroma STFT, which is the first of the methods used in this study to obtain this vector, the frequency axis obtained from the signal is first divided into parts to ensure that each part presents a pitch class. Once this is done, the frequency axis is linearly labeled to represent MIDI pitch classes with logarithmic separation.

j) Chroma CQT: CQT is an another way of calculating the Chroma vectors. CQT provides transition from time domain to frequency domain similar to fft (fast Fourrier transform). The difference is in CQT frequency axis scaled logarithmically. This is much more closer to human hearing.

k) Tonnetz: Tonnetz used in music to show the harmonic relations. To be able to use tonnetz the Chroma vector has to be calculated. Tonnetz returns centroids of tones.

l) Polynomial Features: A polynomial of the desired degree is generated for the spectrum of the signal and the coefficients of this polynomial are returned. Interpolation methods are used to produce the polynomial. Appropriate interpolation method is used according to the desired degree.

m)Wavelet Transforms: The wavelet transform is a generalized form of Fourier transform. It uses a wavelet function which contains abrupt changes unlike sinusoidal functions. Wavelet coefficients are obtained by stretching and shifting this wavelet function throughout the signal. In this study, discrete wavelet transformation is utilized to extract acoustic features [4].

2) Classification with Machine Learning Algorithms: In this section, the machine learning algorithms used in this study for classification of music genres explained in detail. All algorithms were implemented through Scikit-Learn library.

a) MLP: Multilayer Perceptron is a feed forward neural network method. The input values go through neurons by multiplied by the weights of neurons and the results determine the class. MLP consists of an input layer, hidden layers, and an output layer. Except for input nodes, each node uses a non-linear activation function [3]. In this study, MLP algorithm has been trained with stochastic gradient descent (SGD) algorithm.

b) Logistic Regression: Logistic regression is a classification method based on the possibility of belonging to a class, even though there is regression in the name. This method is a regression method that is used to determine the cause-and-effect relationship with the explanatory variables in cases where the target variable is categorical type. Simple and multiple regression analyzes are used to analyze the mathematical relationship between dependent variables and explanatory variables. [5]

c) Random Forest: This algorithm uses multiple decision trees to classify data. The forest created by the algorithm is often a collection of decision trees trained by the ”bagging“ method [2].

d) Linear Discriminant Analysis - LDA: LDA is a classification method by finding linear combination of features. The obtained model uses for classification or more commonly uses for dimension reduction analysis [6].

e) K Nearest Neighbourhood - k-NN: In k-NN classification algorithm, after the training samples are sorted from small to large according to the distance between the test samples, the overriding class label in the first k sample is determined as the class label of the test sample. The K-NN algorithm is used in two different ways: Majority voting and distance-weighted score. In this study, the distance-weighted score was preferred for the K-NN algorithm [2].

f) SVM: Support Vector Machine is one of the simple but highly effective algorithms used in classification processes. This algorithm works by the logic of determining a boundary between the classes in a plane, obtaining fields that separate the classes from one another. Planes obtained here are called hyper-planes [2]. In this study SVM poly kernel used with 3 degree polynomials.

g) Naive Bayes: When classifying with Naive Bayes based on Bayes's Theorem, a conditional probability is computed for all class labels of the given test sample using a set of training samples. The test sample is classified by this class label if the highest probability is obtained by which class label [3].

h) Gradient Boosting: Gradient boosting is trying to minimize the loss function by using stages like other boosting methods. It is generally used with regression and classification. It gives combination of weak prediction models in form of decision trees [7].

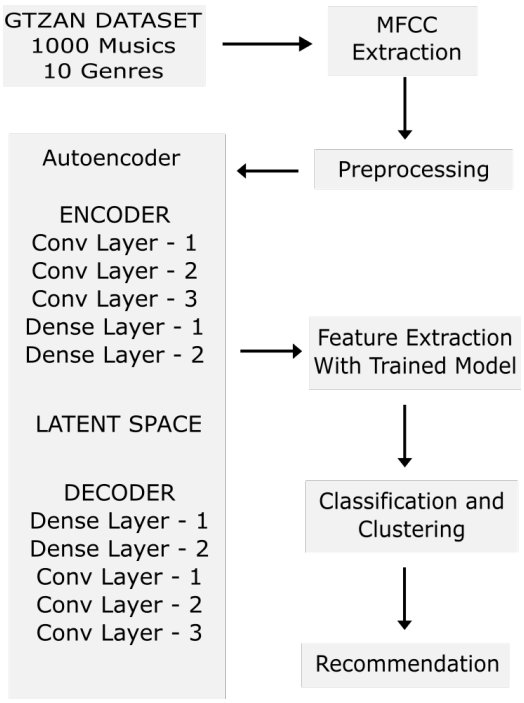

Figure 1 Flowchart of proposed method

3) Autoencoder: As a first step all musics in dataset divided into six segments results in 5 seconds length music. Following this step for each segment Mel spectrogram is generated and saved as an image. For extracting MFCC features, to create mel spectrogram window size is set to 2048 and hop length is set to 512. First 13 MFCC features chosen. Final size of MFCC is 13x259. 259 is the number of MFCC vectors per segment.

Then autoencoder model is created. For encoder part, same features as MusicRecNet [8] is used. In decoder part, number of filters used in reverse order to be able to get the input image back. However since our target with autoencoder is to reconstruct the image with the small loss rate we needed to use mean squared error as a loss function. Additionally activation function needed to be changed from softmax to sigmoid in last layer since we don't do classification on last layer. Also for decoder instead of max pooling layers, up sampling layers are used. A flowchart of this study is shown in Fig. 1.

Latent space size used as dynamic parameter to see the affects of it. In this study we chose to use 64, 128, 256, 512, 1024 latent space size. By changing the latent space size our aim is to see how more compression affects the results. MFCC images are resized and converted into gray scale images before training. Parameters of the proposed autoencoder are given in Table I.

For training purposes 5128 random musics chosen. For testing and validation 570 and 300 random music chosen respectively. After training the model with different latent space sizes, dataset is fed into the model and from the latent space layer feature vectors are extracted. With these features music classification and clustering algorithms implemented

Table I Autoencoder Parameters

Parameter Type	Autoencoder
input size	128 x 128
batch size	64
number of filters encoder	32, 64, 128
number of filters decoder	128, 64, 32
kernel size	3 x 3
loss function	mean squared error
activation function	sigmoid
optimizer	Adam

and obtained results were compared with Elbir and Aydin's [8] study's results.

III. RESULTS

A. Digital Signal Processing

In this section, the different results obtained according to the features used and the classification algorithm will be compared. In the tests conducted with 8 different classification algorithms by using 10-fold cross validation, the highest success was achieved by SVM algorithm. All features were used in all tests. The results obtained with all algorithms are given in the Table II.

Table II Classification Results K=10

Algorithm	Avg Accuracy (%)	Max Accuracy (%)
MLP	74.1	78
Logistic Regression	80.6	84
Random Forest	77.7	84
LDA	81.4	86
KNN	73.2	79
GaussianNB	64.7	73
Gradient Boosting	75.6	79
SVC (Poly)	80.7	86
SVC (Linear)	77.4	86
SVC (RBF)	**81.9**	**88**

Above results were obtained by using every feature extraction method that was mentioned and the highest accuracy obtained from SVM algorithm with rbf kernel with 81.9% accuracy. In Table III, the performance result obtained from using only MFCC feature and it's derivative are shown.

Table III Only MFCC Classification Results K=10

Algorithm	Avg Accuracy (%)	Max Accuracy (%)
MLP	69.7	79
Logistic Regression	72.5	78
Random Forest	70.3	78
LDA	**75.3**	**84**
KNN	61.8	67
GaussianNB	60.9	67
Gradient Boosting	70.3	79
SVC (Poly)	66.6	73
SVC (Linear)	69.4	85
SVC (RBF)	73.4	81

As seen above the highest result obtained by LDA algoritm with 75.3% accuracy. As results indicate using only MFCC doesn't reduce the accuracy substantially. This can be interpreted as that the MFCC itself is enough to classify the musics and other feature extraction methods help classifying algoritms to give better results.

B. Autoencoder

In this study for autoencoder, accuracy is evaluated by using the latent space vectors in classification and clustering algorithms and genres considered as labels. Table IV shows the accuracy of classification algorithms for latent size of 512.

Table IV Classification Results for Latent Space Size 512

Classifier	One Pass	Avg accuracy, %	
		Five-fold	Ten-Fold
MLP	0.58	0.42	0.43
Logistic Regression	0.51	0.43	0.44
Random Forest	0.42	0.37	0.36
LDA	0.48	0.32	0.36
KNN	0.31	0.25	0.26
SVM	0.48	0.33	0.36

When the latent size decreased or increased accuracy has not improved. The best results are obtained by latent size 512. From Table IV, it can be seen that Multi Layer Perceptron MLP gives the best result among the other classification algorithms. Therefore it is seen that using just autoencoder has not made any improvement rather it gives less accuracy. In Table V, clustering with K-Means algorithm results are shown by selecting k as 10.

Table V Clustering Results for Latent Space Size 512
K-Means

Music Genres	Cluster Numbers / Genre Count									
	0	1	2	3	4	5	6	7	8	9
blues	6	31	21	13	111	91	68	14	160	80
classical	-	196	3	30	205	-	105	-	26	
country	37	11	91	15	66	32	230	-	111	4
disco	51	2	232	22	5	68	35	1	166	11
hiphop	151	-	120	64	3	94	27	3	103	30
jazz	27	82	35	21	175	71	116	-	49	17
metal	8	1	35	21	3	36	118	143	78	153
pop	187	4	304	29	14	11	39	1	4	-
reggae	114	4	44	85	18	191	37	-	98	2
rock	25	4	110	22	16	65	159	30	137	24

It can be seen that in some clusters genres spread evenly for every genre. On the other hand some clusters have some genre prior to others. For instance in cluster 7 metal forms most of the cluster. From those observations we can see that most of the genres' most of the musics take place in two or three clusters.Since genres are not clustered homogeneously, with using such clustering getting high performance from music recommendation is not always the case. In order to see the real results it should be tested with real world application to see if recommended musics are drawing the people attention.

Music recommendation tried with the clustering algorithm that is trained using latent space vectors. The

Table VI Recommendation using Clustering

Music Genres	10 different Recommendations
blues	hiphop, hiphop, jazz, jazz, metal, metal, pop, blues, metal, metal
classical	metal, jazz, classical, classical, classical, metal, metal, pop, reggae, jazz
country	pop, hiphop, metal, metal, jazz, reggae, rock, pop, pop, country
disco	reggae, metal, metal, metal, disco, jazz, pop, metal, reggae, disco
hiphop	hiphop, reggae, blues, classical, disco, hiphop, metal, classical, rock, metal
jazz	blues, blues, jazz, metal, country, jazz, blues, country, rock, classical
metal	country, classical, hiphop, classical, blues, pop, metal, hiphop, metal, reggae
pop	hiphop, country, pop, classical, disco, classical, classical, classical, rock, pop
reggae	metal, reggae, blues, metal, metal, reggae, classical, jazz, rock, country
rock	blues, jazz, rock, pop, blues, pop, pop, pop, rock, pop

recommendations however was not of the same genre. Conversely, recommendations were from different genres. In Table VI, sample recommendation can be found.

IV. CONCLUSION

In this study, music genre classification was performed by using the acoustic features of the music and the features that are obtained from autoencoder.

For feature extraction of music; Zero Crossing Rate, Spectral Centroid, Spectral Bandwidth, Spectral Contrast, Spectral Flatness, Spectral Rolloff, RMS, MFCC, MFCC Derivative, Chroma STFT, Chroma CENS, Chroma CQT, Tonnetz, Polynomial and Wavelet transform methods were used. A total of 171 properties were obtained from these methods. For the classification process; K-NN, Naive Bayes, SVM, Random Forest and Decision Tree algorithms were used.

The highest success in classification was obtained by using SVM algorithm with kernel parameter "rbf". MFCC has made the biggest contribution to success.

In a nutshell, our proposed autoencoder method has not shown any improvement on performance in terms of classification accuracy. At the same time, recommendations are not as good as the previous studies. One reason for that failure is considered as the lack of data. For small neural network architecture it is easy to learn from a small dataset. However in this study since the model is big more data may be required. In future studies, with a larger dataset can be tested again. Also variational autoencoder could give better results. Another thing to be experimented is using not just MFCC features but also features from raw audio that are generated by autoencoder like in this study. Then two features can combined and classification and clustering can be done on this combined feature.

REFERENCES

[1] G. Tzanetakis, G. Essl, and P. Cook, "Automatic musical genre classification of audio signals," 2001. [Online]. Available: http://ismir2001.ismir.net/pdf/tzanetakis.pdf

[2] O. Yildiz and A. Karatana, "Music genre classification with machine learning techniques," 05 2017.

[3] A. Elbir, H. Bilal Çam, M. Emre Iyican, B. Öztürk, and N. Aydin, "Music genre classification and recommendation by using machine learning techniques," in *2018 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2018, pp. 1–5.

[4] N. Kehtarnavaz, "Chapter 7 - frequency domain processing," in *Digital Signal Processing System Design (Second Edition)*, second edition ed., N. Kehtarnavaz, Ed. Burlington: Academic Press, 2008, pp. 175 – 196. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780123744906000076

[5] H. Bircan, "Lojistik regresyon analizi: Tıp verileri Üzerine bir uygulama," *Kocaeli Üniversitesi Sosyal Bilimler Dergisi*, pp. 185 – 208, 2004.

[6] M. I. Dogan, A. Orman, M. Örkcü, and H. H. Örkcü, "A new approach based on regression analysis and mathematical programming to multi-group classification problems," 2019.

[7] J. H. Friedman, "Stochastic gradient boosting," *Computational statistics & data analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[8] A. Elbir and N. Aydin, "Music genre classification and music recommendation by using deep learning," *Electronics Letters*, vol. 56, no. 12, pp. 627–629, 2020.