



Machine Learning Model for the Classification of Musical Composition Genres

Javor Marosevic, Goran Dajmbic, and Leo Mrsic^(✉) 

Algebra University College, Ilica 242, 10000 Zagreb, Croatia
{javor.marosevic, goran.dajmbic}@racunarstvo.hr,
leo.mrsic@algebra.hr

Abstract. The aim of this paper is to validate the applicability of standard machine learning models for the classification of a music composition genre. Three different machine learning models were used and analyzed for best accuracy: logistic regression, neural network and SVM. For the purpose of validating each model, a prototype of a system for classification of a music composition genre is built. The end-user interface for the prototype is very simple and intuitive; the user uses mobile phone to record 30 s of a music composition; acquired raw bytes are delivered via REST API to each of the machine learning models in the backend; each machine learning model classifies the music data and returns its genre. Besides validating the proposed model, this prototyped system could also be applicable as a part of a music educational system in which musical pieces that students compose and play would be classified.

Keywords: Machine learning · Musical composition genre classification · Logistic regression · SVM · Neural networks

1 Introduction

Music is the generally accepted name for shaping the physical appearance of sound in a purposeful way. Music, although rooted in basic physical phenomena such as sounds, tones, and temporal relationships, is a very abstract phenomenon defined by subjective features and as such presents a challenge when attempting to analyze it using exact computer methods. Recognizing the musical characteristics of a piece of music is often a problem, both for people and especially for computers. Creating models and systems that can recognize these characteristics and classify a musical work in one of the predefined classes is not a simple undertaking and involves expertise in several scientific fields; from physics (to understand the mechanical properties of sound), through digital signal analysis (to overcome the problem of converting sound into its digital form) to knowledge of music theory (to analyze the relationships between the frequencies of sound (tones) so that useful characteristics specific to a particular genre of musical work can be identified). The genre of a musical work is a set of aspects that characterize that musical work so it can be paired with other, similar musical works, thus forming classes into which musical works can be classified. In order to find similarities between musical works,

common terminology is needed that describes the relationships between the physical phenomena of sound and time and allows for their objective comparison. The aim of this paper is to propose a model that can classify the genre of a musical work from a provided soundtrack, with high accuracy. Based on the proposed model, a system was developed that was used for recognition testing. The functionalities of the system are divided into two specific parts: functionalities related to the training of models and functionalities related to the use of trained models. When training models, the system must be able to receive a series of sound recordings, analyze them, generate features based on the analysis in a format suitable for training several types of different models, train several types of prediction models and evaluate trained prediction models. When using trained models, the system must be able to receive the soundtrack in a predefined format, analyze the received soundtrack and return information to the user about the genre of the musical work of the submitted soundtrack and how certain model is that the submitted soundtrack belongs to the recognized class. As the user interface of the system, a mobile application was created that offers the user the functionality of recording audio from the environment, submitting that audio to the backend application and then receiving and displaying feedback on the music class of the submitted recording and prediction reliability. There exist several works that deal with a similar topic as this paper does. In contrast, this paper attempts to identify the genre of a musical work using lower order features and introduces several higher order features that use elements of music theory to better describe the characteristics of individual genres. Bisharad and Laskar use the architecture of repeatable convolutional neural networks and the melspectrogram as input features (Bisharad and Laskar 2019). Nanni et al. have an interesting approach combining acoustic features and visual features of audio spectrograms in genre recognition (Nanni et al. 2016). Cataltepe et al. use MIDI as an input format for recognition which greatly facilitates the extraction of higher order features because it is not necessary to process the input signal to obtain tonal features (Cataltepe et al. 2007). Li and Ogihara apply a hierarchy of taxonomy to identify dependency relationships between different genres (Li and Ogihara 2005). Lidy and Rauber evaluate feature extractors and psycho-acoustic transformations for the purpose of classifying music by genre.

2 Background

Computer-assisted sound analysis is a subset of signal processing because the focus is on characteristics of signal with frequencies less than 20 kHz. Sound can be viewed as the value of air pressure that changes over time (Downey 2014). For computer-assisted sound analysis, it is practical to convert sound into an electrical signal. Such electrical signal can be represented as a series of values over time or graphically. The graphical representation of the signal over time is called the waveform (Fig. 1, top left). It is rare to find a natural phenomenon that creates a sound represented by a sinusoid. Musical instruments usually create sounds of more complex waveforms. The waveform of the sound produced by the violin is shown in Fig. 1, top right. It can be noticed that the waveform is still repeated in time, i.e. that it is periodic, but it is much more complex than the waveform of a sinusoid. The waveform of a sound determines the color or timbre of the sound, which is often described as sound quality, and allows the listener to recognize different types of musical instruments.

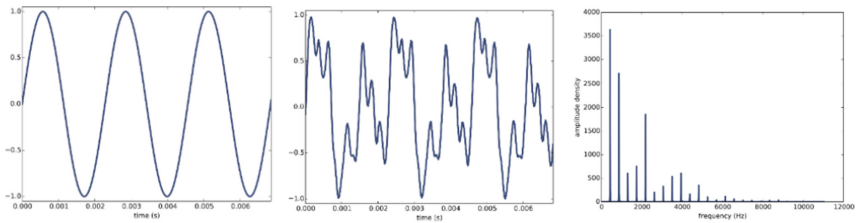


Fig. 1. Left: sound signal of 440 Hz; Middle right: violin sound waveform; Right: violin sound signal spectrum

Spectral Decomposition

One of the most important tools in signal analysis is spectral decomposition, the concept that each signal can be represented as the sum of sinusoids of different frequencies (Downey 2014). Each signal can be converted into its representation of frequency components and their strengths called the spectrum. Each frequency component of the spectrum represents a sinusoidal signal of that frequency and its strength. By applying the spectral decomposition method to the violin signal, the spectrum shown in Fig. 1 bottom will be obtained. From the spectrum it is possible to read that the fundamental frequency (lowest frequency component) of the signal is around 440 Hz or A4 tone. In the example of this signal, the fundamental frequency has the largest amplitude, so it is also the dominant frequency. As a rule, the tone of an individual sound signal is determined by the fundamental frequency, even if the fundamental frequency is not the dominant frequency. When a signal is decomposed into a spectrum, in addition to the fundamental frequency, the signal usually contains other frequency components called harmonics. In the example in Fig. 1, there are frequency components of about 880 Hz, 1320 Hz, 1760 Hz, 2200 Hz, etc. The values of the harmonic frequencies are multiples of the fundamental signal frequency. Consecutive harmonics, if transposed to one octave level from the fundamental tone, form intervals of increasing dissonance. The series of harmonics over the base tone is called the harmonic series or series of overtones and always follows the same pattern. Fourier transforms are a fundamental principle used in the spectral decomposition of signals because they are used to obtain the spectrum of the signal from the signal itself. In this paper, discrete Fourier transforms were used. Discrete Fourier transform takes as input a time series of N equally spaced samples and as a result produces a spectrum with N frequency components. The conversion of a continuous signal into its discrete form is performed by sampling. The level of the continuous signal is measured in equally spaced units of time, and the data on each measurement is recorded as a point in the time series. The appropriate continuous signal sampling frequency is determined by applying the Nyquist-Shannon theorem or sampling theorem: “If the function $x(t)$ does not contain frequencies higher than B hertz, the function is completely determined by setting the ordinate by a series of points spaced $1/(2B)$ seconds.” (Shannon 1949). This means that the sampling frequency of the signal must be at least $2B$, where B is the highest frequency component of the signal. This frequency is called the Nyquist frequency. The problem of sampling a signal with a frequency lower than the Nyquist frequency occurs in the form of frequency aliasing, a phenomenon where the frequency component of a higher frequency signal manifests

itself identically as one of the lower frequency components of that same signal. To avoid this, when analyzing sound signals, sampling is performed at frequencies higher than 40 kHz.

3 Research

The proposed model of the system for classifying the genre of a musical work is designed modularly. The system itself is designed in such a way that each module of the model acts as a separate unit (micro service) and is created as a separated functionality that allows scaling the system according to need and load. The system architecture is shown in Fig. 2. The process begins by analyzing a number of audio files and discovering their characteristic features selected for this paper. Feature detection is aided by the use of the open source LibROSA library, which enables basic and complex signal analysis used in sound analysis. The features used for the analysis are divided into lower order features and higher order features. Lower order features are characterized by general properties related to digital signal analysis and concern basic signal measurements without elements related to music and music theory. Higher order features characterize properties related to elements of human understanding of music and music theory such as tones, the relationship between tones, diatomicity, and chromatic properties. After the feature discovery process, the obtained features serve as an input to the prediction model training process. Model training also involves the process of normalizing and preparing data depending on which type of model is selected. Each model goes through a model evaluation process where the model is tested under controlled conditions and preliminary measures of the predictive accuracy of each model are taken. The trained model is used within an endpoint implemented as an HTTP API service. The service is in charge of receiving, processing, analyzing and returning responses to user queries. The service performs the same steps for analyzing the dedicated soundtrack that are used in the feature discovery process that precedes the prediction model training. The user accesses the service via the HTTP protocol and, respecting the defined contract, can develop her own implementation of the user application or use the mobile application developed for this paper and available on the Android operating system to access the service. The mobile application serves as a wrapper around the HTTP API and takes care of implementation details such as audio retrieval, data transfer optimization, feedback processing, session saving when analyzing longer audio tracks, and displaying service feedback. A GTZAN data set consisting of 1000 30-s audio tracks divided into 10 genres, each containing 100 sound tracks, was used to train the model. Audio recordings are in WAV format characterized as 22050 Hz, mono, 16-bit (Marsyas 2020). The genre classes of this data set are: Blues, Classical, Country, Disco, Hip-hop, Jazz, Metal, Pop, Reggae and Rock, and this also represents the classes used by trained models. Logistic regression, neural network and support vector methods (SVM) were used as machine learning methods. To train the model, it was necessary to select a set of features that describe the differences between the individual classes of elements to be classified, in our case genres. To support the process of classifying the genre of a musical work, several lower order features and several higher order features were selected. Lower order features used are: RMS of an audio signal, spectral centroid, spectral bandwidth, spectral rolloff, and zero crossing

rate. The RMS of an audio signal is the root value of the mean value of the square of the signal amplitude calculated for each sampled frame of the input signal. The RMS of an audio signal is usually described as the perceived volume of sound derived from that same audio signal. The mean value of the RMS of all frames and the standard deviation of the RMS of all frames of the input audio signal were used as classification features. The reason for using these features is that certain genres typically contain high levels of differences in perceived volume within the same piece of music, and these features help to discover these characteristics. The spectral centroid is the weighted mean of the frequencies present in the signal calculated for each sampled input signal frame (Fig. 3). The spectral centroid of an audio signal is described as the “center of mass” of that audio signal and is characteristic of determining the “brightness” of sound. Sound brightness is considered to be one of the most perceptually powerful characteristics of timbre sounds.

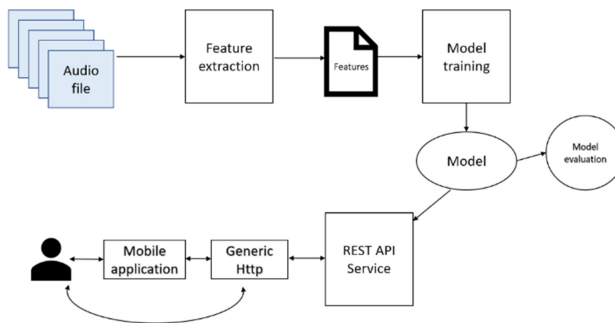


Fig. 2. The architecture of the proposed system

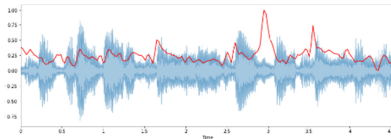


Fig. 3. Spectral centroid of individual audio signal frames

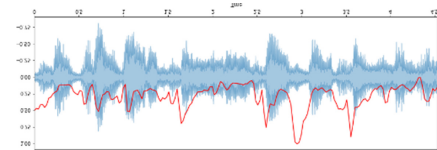


Fig. 4. Spectral rolloff of individual audio signal frames

The mean value of the spectral centroid of all frames and the standard deviation of the spectral centroid of all frames of the input audio signal were used as classification features. The reason for using these features is to reveal the timbre characteristics of individual instruments used in different genres of musical works. Spectral bandwidth is the bandwidth at half the peak of the signal strength (Weik 2000) and helps to detect audio signal frequency scatter. The mean value of the spectral bandwidth of all frames and the standard deviation of the spectral bandwidth of all frames of the input audio signal were used as classification features. Spectral rolloff is the value of the frequency limit below which a certain percentage of the total spectral energy of an audio signal is located (Fig. 4) and helps to detect the tendency of using high or low tones within a piece of music. The mean value of the spectral rolloff of all frames and the standard deviation

of the spectral rolloff of all frames of the input audio signal were used as classification features. The zero crossing rate is the rate of change of the sign of the value of the audio signal (Fig. 5). This measure usually takes on more value in percussion instruments and is a good indicator of their use within musical works. The mean value of the zero crossing rate of all frames and the standard deviation of the zero crossing rate of all frames of the input audio signal were used as classification features. Of the higher order features, the following were used: Mel-frequency cepstrum coefficients, melodic change factor and chromatic change factor. Mel-frequency cepstrum coefficients are a set of peak values of logarithmically reduced strengths of the frequency spectrum of signals mapped to the mel frequency scale. These features are extracted by treating the frequency spectrum of the input signal as an input signal in another spectral analysis. The input signal is converted into a spectrum and such a spectrum is divided into several parts, and the frequency values are taken as the frequency values from the mel scale due to the logarithmic nature of the human ear's audibility. The strength values of each of the parts are scaled logarithmically to eliminate high differences between the peak values of the spectrum strengths. Such scaled values are treated as the values of the new signal which is once again spectrally analyzed. The amplitude values of the spectrum thus obtained represent the Mel-frequency cepstral coefficients. The number of coefficients depends on the selected number of parts into which the result of the first spectral analysis is divided. The number of coefficients selected for this paper is 20.

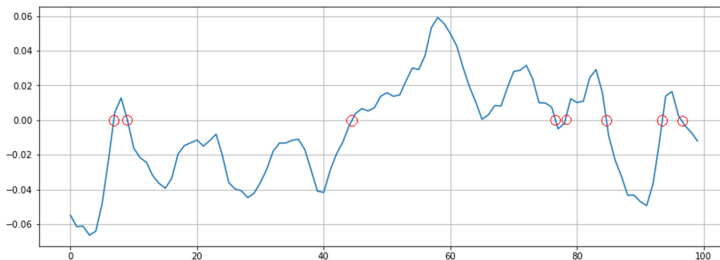


Fig. 5. Changes in the sign of the value of the audio signal

The melodic change factor represents the rate of melodic change of the predominant tones within the input audio signal. The melodic change factor is calculated by first generating a chromagram (Fig. 6) of the input signal. The chromagram represents the distribution of the tone classes of the input audio signal over time. Distribution is calculated for each input signal frame. The melodic change factor helps us to discover the rate of melodic dynamism of a musical work.

The chromatic change factor represents the rate of chromatic change of the predominant tones within the input audio signal. By chromatic change we mean the increase or decrease of two consecutive tones by one half-step. The chromatic change factor is calculated in a similar way as the melodic change factor. A chromagram is generated and the strongest predominant tone class of the input frames is extracted using a sliding window to reduce the error. After extracting the strongest tone class, the cases when two consecutive frames have a change of tone class in the value of one half-degree are

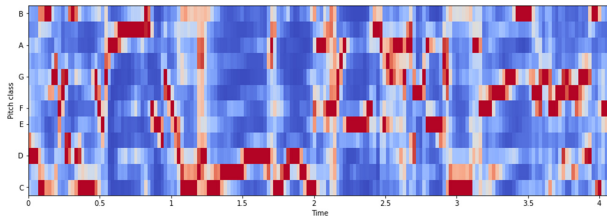


Fig. 6. Chromagram of audio signal

counted. This number is divided by the total number of input frames to obtain a change factor independent of the duration of the audio input signal. The chromatic shift is very characteristic for certain genres of musical works, such as works from the jazz genre or the classical genre, and at the same time rare in parts of the popular genre or rock genre.

4 Results and Discussion

In this paper, the term “controlled conditions” refers to the testing of a model on a test dataset separate from the original dataset on which the model was trained. These data are pseudo-randomly extracted from the original dataset and it is expected that this subset of data shares very similar characteristics as the set on which the models of the genre classification system of the musical work were trained. A rudimentary measure of model accuracy shows that all three models used in this paper have an average prediction accuracy between 70% and 80% (Table 1). The neural network model and the SVM method model show approximately equal accuracy while the logistic regression model is about 5% less accurate. The resulting confusion matrices of prediction models are shown in Fig. 7. It can be observed that all three models have lower accuracy of classification of musical works of the genres Rock, Disco, Country and Reggae. It is interesting to observe that Rock music is often misclassified as Disco music, and the reverse is not valid, i.e., Disco music is not classified as Rock music except in the case of models of SVM model. A similar situation can be noticed with Disco and Pop music. Disco music is more often classified as Pop music, while Pop music is not classified as Disco music.

Table 1. Accuracy of each model

Model	Model accuracy
Logistic regression	0.72
Neural network	0.77
SVM	0.76

A closer look at the confusion matrices of all three models leads to the conclusion that the neural network model has much less scattering of results than the other two

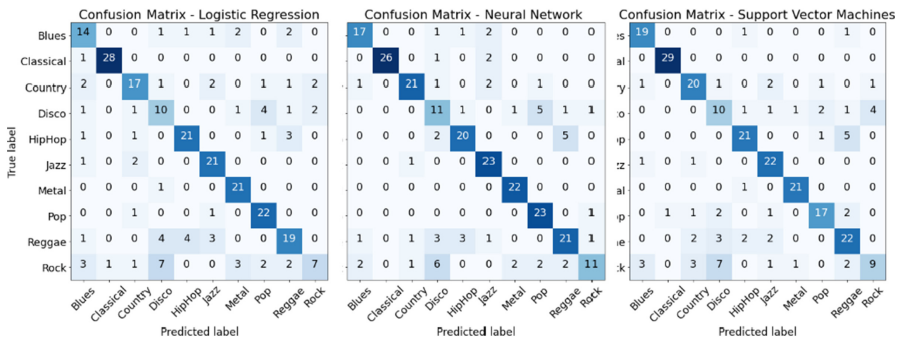


Fig. 7. Confusion matrix of used models

Table 2. Area under the ROC curve of the genres of individual models

Genre	Log. Regres. ROC AUC	Neur. Netw. ROC AUC	SVM ROC AUC
Blues	0.967	0.989	0.981
Classical	1	0.998	1
Country	0.946	0.969	0.971
Disco	0.875	0.917	0.939
Hip-hop	0.941	0.963	0.981
Jazz	0.975	0.988	0.990
Metal	0.999	1	0.997
Pop	0.989	0.992	0.983
Reggae	0.928	0.951	0.934
Rock	0.879	0.914	0.853
Average	0.945	0.968	0.963

models. In the case of most genres, the neural network model classifies the entities of a genre into fewer possible classes than the other two models do, i.e. when classifying works of a genre, there is a good chance that one of the two possible classes will be attributed to the work. This is not the case with the Rock and Reggae genres in which the neural network model also classifies entities into several different classes.

Figure 8 shows the ROC curves for all three models, separately for each prediction class. Areas below the ROC curves were also calculated for each of the prediction models and shown in Table 2 as well as the average area under the curves (AUC) of all genres.

Curves representing the genres Classical, Metal and Jazz have a very steep rise at the beginning which indicates a high certainty that, in case of recognition of these categories, musical works really belong to this category which can be attributed to the characteristics of these genres of musical works. The Blues and Pop genre curves also show good recognition reliability characteristics. In other genres, when classifying a

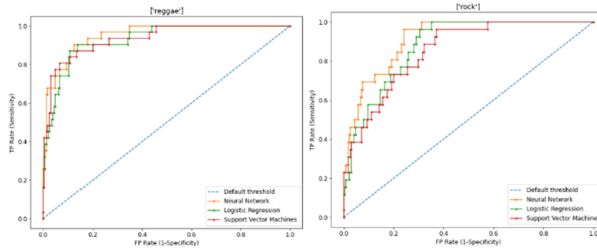


Fig. 8. ROC curves for selected genres

piece of music into one of these categories, we must be careful because the number of false identifications of these categories is slightly higher than the previously mentioned. Such results are expected because these musical genres often borrow elements that could be classified in some of the other genres of musical works, and it is not difficult to imagine a situation where one would misrecognize a genre based on a short clip of one of these musical genres.

In this paper, the term “uncontrolled conditions” is considered to be a set of user selected parts of musical works and their classification by the proposed system. Testing in uncontrolled conditions was performed by selecting three musical works from each genre. A 30-s sample was taken from each selected piece of music and this sample was passed through the system. The 30-s sample was taken so that the beginning of the sample corresponds to the time point of the musical work at 45 s, and the end of the sample corresponds to the point of the musical work at 75 s. This method of sampling was chosen in order to avoid the appearance of silence at the beginning of some musical works and to avoid intros that might not belong to the genre of the musical work. Audio recordings are submitted to processing to the HTTP API where the trained models are located. The result of classification was provided by the neural network model because this model showed the highest recognition accuracy. Selected musical works and their test results are shown in Table 3.

To make it easier to comprehend the test results in uncontrolled conditions, the results are visualized in Fig. 9.

It is interesting to note that certain genres have one hundred percent accuracy of classification, while some genres have not correctly recognized any of the cases. The category of classical music is characterized by the use of a very different spectrum of instruments from other categories so the high recognition accuracy in this category is not surprising. Likewise, the Hip-hop category is characterized by an extremely strong rhythmic and predominant vocal component and the result is expected. The unexpected result is the inability to recognize any randomly selected work in the Country, Metal and Pop genres. A possible explanation for these results is that these genres are relatively similar to other genres, in terms of the features selected and the instruments used. Also, within these genres there are several different subgenres that differ greatly from each other and it is possible that the dataset on which the models are trained is biased in favor of some of these subgenres and that randomly selected musical works that do not have the characteristics needed to classify that musical works into a trained model class. The low recognition accuracy in the Rock and Blues genres might be explained

Table 3. Randomly selected musical works, their actual genres and genres classified by neural network model

Musical work	Actual genre	Classified genre
Muddy Waters - Hoochie Coochie Man	Blues	Jazz
Sam Myers - I Got the Blues	Blues	Blues
The Doors - Roadhouse Blues	Blues	Disco
Bach - Brandenburg Concerto No. 1 in F major	Classical	Classical
Claude Debussy - Clair De Lune	Classical	Classical
Dmitri Shostakovich - Waltz No. 2	Classical	Classical
Hank Williams – Jambalaya	Country	Classical
Johnny Cash and June Carter – Jackson	Country	Jazz
N. Sinatra - These Boots Are Made For Walkin'	Country	Reggae
Bee Gees - Stayin' Alive	Disco	Disco
Boney M. – Rasputin	Disco	Rock
Earth, Wind & Fire - Boogie Wonderland	Disco	Disco
Dr Dre - What's the Difference	Hip-hop	Hip-hop
The Sugar Hill Gang - Rapper's Delight	Hip-hop	Hip-hop
Tupac Ft Elton John - Ghetto Gospel	Hip-hop	Hip-hop
Dave Brubeck - Golden Brown	Jazz	Jazz
John Coltrane - Giant Steps	Jazz	Jazz
Medeski, Martin & Wood – Kota	Jazz	Classical
Disturbed - Down With the Sickness	Metal	Hip-hop
Drowning Pool – Bodies	Metal	Hip-hop
Metallica - Fade to Black	Metal	Rock
Abba – Waterloo	Pop	Classical
Cyndi Lauper - Girls Just Want To Have Fun	Pop	Jazz
Michael Jackson - Billie Jean	Pop	Reggae
Bob Marley - Buffalo soldier	Reggae	Reggae
Peter Tosh - Wanted Dread & Alive	Reggae	Disco
UB40 -Kingston Town	Reggae	Reggae
Dire Straits - Romeo And Juliet	Rock	Reggae
Kansas - Carry on Wayward Son	Rock	Rock
ZZ Top - La Grange	Rock	Disco

in a similar way. Both genres are very general and their subgenres reflect some of the characteristics of other genres. These hypotheses are also supported by controlled test

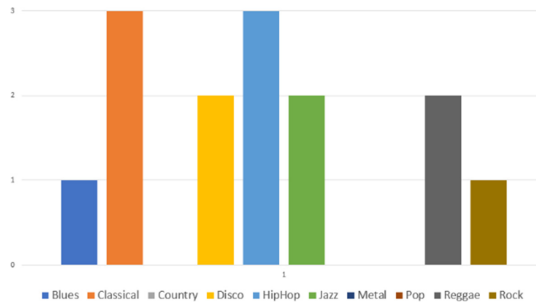


Fig. 9. Number of correct classifications by genre

results that show lower accuracy in recognizing these genres on test data extracted from the original training dataset.

5 Conclusion

The test results of the proposed system for classifying the genre of a musical work show that all considered models have an accuracy between 70% and 80%. The analysis of the results shows that the models show better performance in the classification of certain genres, and for some genres the accuracy is very close to one hundred percent (for example, in the case of classical music genre accuracy is greater than 90%) which shows predispositions to improve classification results in other genres also. Testing in controlled conditions and analysis of results with convolutional matrices and ROC curves showed slightly worse recognition results of the Rock, Disco and Reggae genres, while the results of the Classical, Metal and Jazz genres showed very good recognition results. Testing in uncontrolled conditions in most genres follows the results of controlled testing, except for the genres Metal, Pop and Country, which showed very poor results. Possible explanations for this discrepancy are the diversity of styles within the genres and the similarity of individual subgenres with other music genres and the bias of the input dataset to some of these subgenres, which leads to incorrect classification of individual musical works belonging to these genres. The obtained results suggest that by training the model on a larger dataset the accuracy might be higher and that by generating a larger number of higher orders features it would be possible to increase the accuracy of the classification using the current dataset. It should also be taken into account that the boundaries between genres of musical works are not clearly defined and many musical works belonging to a certain genre also contain elements characteristic for other genres. For example, it is sometimes difficult for a human being to decide whether a work is heavy metal or hard rock. This suggests that the accuracy of classification, from the point of view of classifying a musical work into only one of the available classes, will not be possible to achieve a very high accuracy (95% and more). Classifying works into one or more genres could potentially show better results. In the context of the proposed model and the developed system, the neural network model proved to be the most accurate classification model. The classifier based on the SVM shows the same average accuracy as the classifier based on neural networks, but with a greater scatter of results between

several possible classes. This phenomenon indicates the instability of this classifier for the purpose used in this model and is the main reason for choosing the neural network as the primary method that would be used in the final classification model in case of further development in the intended direction.

References

- Cooley, J.W., Tukey, J.W.: An algorithm for the machine calculation of complex Fourier series. *Math. Comput.* **19**(90), 297–301 (1965)
- Dabhi, K.: Om Recorder Github (2020). <https://github.com/kailash09dabhi/OmRecorder>
- Downey, A.B.: Think DSP. Green Tea Press, Needham (2014)
- National Radio Astronomy Observatory: Fourier Transforms (2016). Preuzeto 2020 iz. <https://www.cv.nrao.edu/course/ast534/FourierTransforms.html>
- Encyclopaedia Britannica: Encyclopaedia Britannica, 16 June 2020. <https://www.britannica.com/art/consonance-music>
- Fawcett, T.: An introduction to ROC analysis. *Science Direct* (2005)
- Hrvatska Enciklopedija: Hrvatska enciklopedija, mrežno izdanje, Leksikografski zavod Miroslav Krleža (2020). Preuzeto 2020 iz. <https://www.enciklopedija.hr/natuknica.aspx?ID=67594>
- ITU: Universally Unique Identifiers (UUIDs) (2020). Preuzeto 3. 8 2020 iz ITU. <https://www.itu.int/en/ITU-T/asn1/Pages/UUID/uuids.aspx>
- Marsyas: GTZAN dataset, 10 July 2020. <http://marsyas.info/downloads/datasets.html>
- Serdar, Y.: What is TensorFlow? The machine learning library explained (2019). Preuzeto 25. 8. 2020 iz. <https://www.infoworld.com/article/3278008/what-is-tensorflow-the-machine-learning-library-explained.html>
- Theano: Theano documentation (2017). <http://deeplearning.net/software/theano/>
- Melo, F.: Area under the ROC curve. In: Dubitzky, W., Wolkenhauer, O., Cho, K.H., Yokota, H. (eds.) *Encyclopedia of Systems Biology*. Springer, New York (2013). https://doi.org/10.1007/978-1-4419-9863-7_209
- Merriam-Webster.Com: Merriam Webster (2020). Dohvaćeno iz. <https://www.merriam-webster.com/dictionary/music>
- Shannon, C.E.: Communication in the presence of noise. *Proc. IRE* **37**(1), 10–21 (1949)
- Square, Inc.: Retrofit (2020). Preuzeto 05. 08 2020 iz. <https://square.github.io/retrofit/>
- Ting, K.M.: Confusion matrix. In: Sammut, C., Webb, G.I. (eds.) *Encyclopedia of Machine Learning*. Springer, Boston (2011). https://doi.org/10.1007/978-0-387-30164-8_157
- Weik, M.H.: *Computer Science and Communications Dictionary*. Springer, Boston (2000). <https://doi.org/10.1007/1-4020-0613-6>
- Bisharad, D., Laskar, R.H.: Music genre recognition using convolutional recurrent neural network architecture. *Expert Syst.* **36**(4), e12429 (2019)
- Nanni, L., Costa, Y.M.G., Lumini, A., Kim, M.Y., Baek, S.R.: Combining visual and acoustic features for music genre classification. *Expert Syst. Appl.* **45**, 108–117 (2016)
- Cataltepe, Z., Yaslan, Y., Sonmez, A.: Music genre classification using MIDI and audio features. *EURASIP J. Adv. Sig. Process.* **2007** (2007). <https://doi.org/10.1155/2007/36409>. Article ID 36409
- Li, T., Ogihara, M.: Music genre classification with taxonomy. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, Philadelphia, PA, vol. 5, pp. v/197–v/200 (2005). <https://doi.org/10.1109/icassp.2005.1416274>
- Lidy, T., Rauber, A.: Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In: *ISMIR*, pp. 34–41 (2005)