Music genre classification based on spectrograms of the sound recording using an ensemble of CNNs

Tadej Lahovnik
tadej.lahovnik@student.um.si
Faculty of Electrical
Engineering and Computer Science,
University of Maribor
Koroška cesta 46
SI-2000 Maribor, Slovenia

Vili Podgorelec
vili.podgorelec@um.si
Faculty of Electrical
Engineering and Computer Science,
University of Maribor
Koroška cesta 46
SI-2000 Maribor, Slovenia

ABSTRACT

Several papers have attempted to classify music genres based on the features extracted from sound recordings. However, none have implemented an ensemble classifier of different CNNs for various types of spectrograms.

One thousand sound recordings from the GTZAN database were used for classification by the authors. Each sound recording was converted into three different spectrogram types, resulting in 3000 spectrograms. 85% of the spectrograms were used to train three CNN models, and the remaining 15% were used for testing. The individual CNN models formed a classifier ensemble, which combined the predictions of respective models into a single prediction based on the sum of the scores of respective genres.

Since the accuracy of the classifier ensemble (54.67%) is higher than the accuracy of the individual classification models (44.00%, 53.33%, 26.67%), it was beneficial to combine the CNN models into one. The confusion matrix revealed some common errors in genre prediction. The somewhat low accuracy is likely a consequence of the truncated sound recordings. Although the classifier ensemble did not achieve high accuracy, it predicted the genre based on the spectrograms of the sound recording more accurately than a human. Weighting the individual CNN models could significantly improve the results.

KEYWORDS

classification, spectrogram, machine learning, convolutional neural network, music genre, sound recording

ACM Reference Format:
Tadej Lahovnik and Vili Podgorelec. 2022. Music genre classification based on spectrograms of the sound recording using an ensemble of CNNs. In *Proceedings of Student Computing Research Symposium (SCORES'22).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

SCORES'22, October 6, 2022, Ljubljana, Slovenia

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

During the last two decades, we have seen an increase in AI-driven recommendation systems on audio streaming platforms. There are two general approaches to music recommendation - collaborative and content-based recommendation. While the former recommends objects that the user group of similar preference has liked, the latter analyses the content of objects that a user has previously preferred and recommends the ones with relevant content [8]. For content-based music recommendation, automatic music genre identification plays an important role.

Various machine learning (ML) methods have been developed for accurate music genre classification [7]. The algorithms behind these systems are often based on metadata and features extracted from sound recordings using audio processing techniques. Audio signal processing algorithms generally involve analysis of a signal, extracting its properties, predicting its behaviour, recognising if any pattern is present in the signal, and how a particular signal is correlated to other similar signals [16]. The extracted audio signal features represent training data for a selected ML algorithm.

While such ML approaches have shown some excellent results, there are other ways of representing and processing audio signals. An audio signal can be visually represented with spectrograms, potentially revealing patterns characteristic of different types of music. Existing works [3, 12] have already approached the task using spectrograms or features extracted from sound recordings. Different spectrogram types can be used to represent an audio signal. For example, Mel spectrograms[1,2] can be used instead of typical spectrograms [3, 9, 17]. As spectrograms are visual representations of music, the most advanced deep neural network-based image classification techniques can be applied to music genre classification.

In this paper, we propose a new ensemble classification method that predicts the music genre of a sound recording based on several individual convolutional neural network (CNN) models, each trained on its type of spectrogram. To the best of our knowledge, this is the first report on music genre classification using an ensemble classifier combining different types of spectrograms. An essential advantage of the proposed method is the direct use of spectrogram images for training the CNN models, which does not require the often demanding process of extracting and selecting features from sound recordings.

[1] a method of representing audio visually [17]
[2] substitutes the frequency on the y-axis with the mel scale and indicates colours using the decibel scale instead of the amplitude [6]

2 BACKGROUND

2.1 Music genre

A music genre is a category of music characterised by a particular style. A genre can also be influenced by social conventions, marketing, association with a particular artist, and other external influences [2].

Repetition is the foundation of genres. A genre codifies past repetitions and encourages new repetitions [14].

In this paper, genres represent classes for classification. A class is a set of things that can be grouped meaningfully. We often think of a class in terms of the common properties of its members, especially those that distinguish them from other things that are similar in many ways [13].

2.2 Spectrograms

A spectrogram represents a signal's intensity or 'loudness' over time at the different frequencies present in a particular waveform. Changes in energy levels over time are displayed in a spectrogram. [1].

Spectrograms are two-dimensional graphs, where colours represent the third dimension [1]. Time is represented on the horizontal axis. The vertical axis represents frequency, which can also be interpreted as pitch or tone. The lowest frequencies are located at the bottom and the highest at the top.

2.3 Spectrogram generation process

The Librosa[3] library allows us to generate a simple spectrogram from a sound recording. The sound recording is converted into a floating point time series during the upload process. The resulting time series must be converted from the square of the amplitude to decibels before the spectrogram can be displayed.

Figure 1 shows an example of the first type of spectrograms we used for classification. The spectrogram shows the presence of specific frequencies over time. Orange represents a high presence of a particular frequency, and blue represents a low presence.

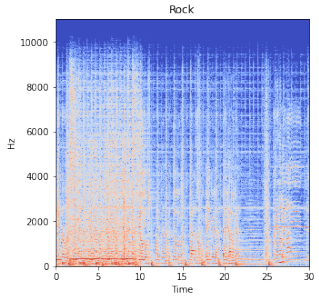

Figure 1: Spectrogram (frequency/time)

Figure 2 shows an example of the second type of spectrograms, which shows the presence of certain tones over time. Orange colour represents a high presence of a particular tone, and blue represents a low presence. The markings on the y-axis indicate the individual octaves (C1-C9).

[3]https://librosa.org/doc/latest/index.html

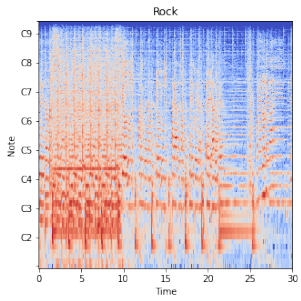

Figure 2: Spectrogram (tone/time)

Figure 3 shows an example of the third type of spectrograms, which shows the presence of tones across all octaves over time. Orange represents a high presence of a particular tone, and blue represents a low presence. Individual marking on the y-axis includes all matching tones from different octaves.

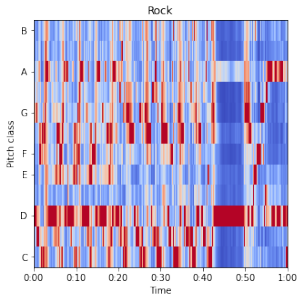

Figure 3: Spectrogram (chroma/time)

2.4 Classification

Classification occurs in many human activities. When used in its broadest sense, the term can refer to any situation in which a prediction or decision is made based on currently available information [10].

Numerous classification algorithms exist, including decision trees, rule-based learning, support vectors, Bayesian networks, and convolutional neural networks (CNNs). If necessary, classification algorithms may also be combined into ensembles (e.g., boosting, bagging, stacking, tree forests, and more) [12].

2.4.1 Neural network. A neural network is a set of algorithms that seek to identify the underlying connections in a set of data through a process that mimics the human brain [5].

Neural networks have three main components: an input layer, a processing or hidden layer, and an output layer [5]. Inputs can be weighted based on different criteria.

In neural networks, learning is accomplished by altering the weights across connections in response to new input data or learning patterns [15].

A convolutional neural network is adapted to analyse and recognise visual data such as digital images or photographs [5].

3 THE PROPOSED METHOD

The GTZAN database [11], which contains 1000 sound recordings, was used for the development. Each sound recording is 30 seconds long and belongs to one of the ten genres[4] provided by the database. The sound recordings were later converted into spectrograms using the Librosa library.

Three different types of spectrograms were created for each sound recording in the dataset. 85% of the spectrograms were used to train the CNN models, and the remaining 15% were used for testing. The training set was partitioned into training and validation sets using the Keras[5] library. The validation set consisted of 30% of the training data.

The generated images were classified using CNNs provided by the Sklearn[6] library. Three separate CNN models were created, trained, and combined into a classifier ensemble. The predictions of the individual models were combined into a single prediction based on the sum of the predictions.

The classifier ensemble was evaluated with several metrics[7]. Additionally, we displayed the results with a confusion matrix, revealing common errors of the implemented classifier ensemble.

3.1 The ensemble of CNN models

3.1.1 The CNN model. For classification, we used the sequential model from the Keras[8] library. Figure 4 shows the visualisation of the model.

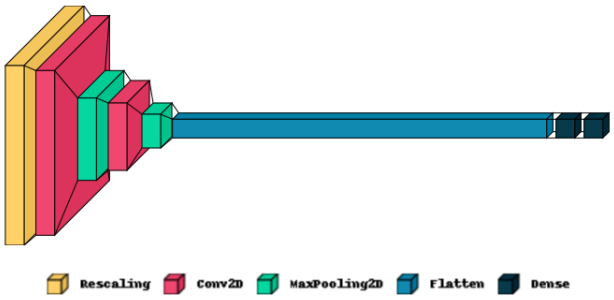

Figure 4: CNN model

The *Rescaling* layer standardises the input data. *Conv2D* creates a 5x5 convolution kernel that is convolved with the layer input to produce a tensor of outputs. *MaxPooling2D* downsamples the input along its spatial dimensions (height and width) using a 2x2 pooling window. The *Flatten* layer flattens the input. The *Dense* layer implements an operation that returns a vector with a length equal to the number of classes, providing the classification scores for each respective class (music genre in our case).

3.1.2 Training a CNN model. Each CNN classification model has been trained separately on its type of spectrogram images (each CNN model is trained from scratch with random initialisation of weights). Fifty epochs were used to train the model. Each batch contained 128 samples. Stochastic gradient descent was used to minimise the loss function of the CNN model. The CNN model outputs a vector with the same length as the number of classes (music genres). This vector represents the scores of individual classes - the higher the score for a particular class, the higher the expectation that the recording belongs to the corresponding music genre. Since these values can be arbitrary, we used a function called Softmax, which ensures that the sum of all the values is 1, thus constraining the individual scores between 0 and 1 [4].

3.1.3 Combining CNN modes into an ensemble. After each CNN model is trained on its type of spectrogram from sound recordings, the individual CNN classifications are combined into an ensemble, comprising the classification results of all three specific CNN models. The instances of three specific CNN models have been combined in a single ensemble model by averaging the outputs of the corresponding Softmax layers. Figure 5 showcases the proposed ensemble method.

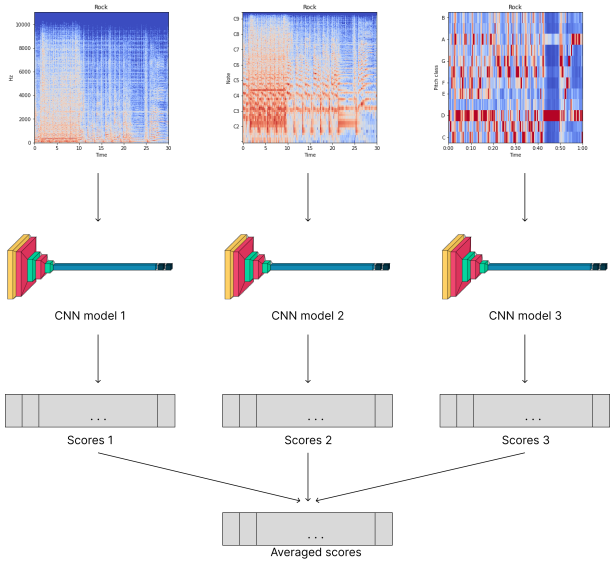

Figure 5: The proposed ensemble classifier

4 RESULTS

Three different CNN classification models were used to perform the classification. We used 150 test instances or 15% of the total dataset for prediction.

The first classification model (which predicted the genre based on classical spectrograms) correctly classified 66 test instances. The accuracy of the first classification model was 44.00%.

The second classification model (which predicted the genre based on spectrograms showing the presence of certain tones over time) correctly classified 80 test samples. The accuracy of the second classification model was 53.33%.

The third classification model (which predicted genre based on spectrograms showing the presence of individual tones across all octaves over time) correctly classified merely 40 test samples. The accuracy of the third classification model was 26.67%.

[4] blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, rock
[5] https://keras.io
[6] https://scikit-learn.org/stable
[7] accuracy, recall, precision, F-score
[8] https://keras.io/guides/sequential_model

After combining the three individual CNN classifiers into an ensemble, the ensemble classifier correctly predicted 82 test instances. The accuracy of the ensemble classifier is 54.67%, which is higher than the accuracy of each CNN classification model. In this manner, merging the individual CNN classification models into an ensemble classifier was worthwhile.

Figure 6 shows the final confusion matrix of the proposed ensemble classifier. The labels of the individual rows show the actual genres of the test instances, while column labels show the genre predicted by the classification ensemble. Values in the cells show the number of test instances belonging to the genre shown in the row, classified as the genre shown in the column. In addition to the numerical labels, we can use the colour scale found next to the confusion matrix.

The confusion matrix shows that there are some common errors in genre prediction. The ensemble often predicted blues as rock, country as jazz and rock, and disco as hip hop, pop, and rock. The most incorrect predictions occurred for the disco genre. On the other hand, classical music seems to be the easiest to distinguish from other genres, as it was only misclassified on a few occasions with jazz.

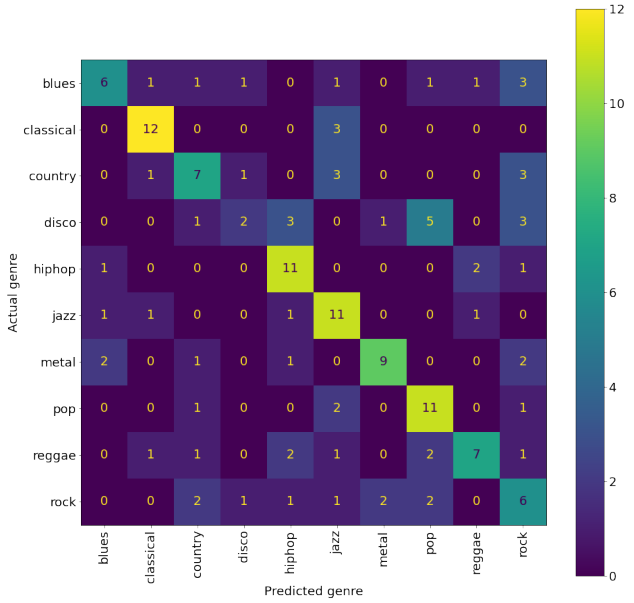

Figure 6: Confusion matrix

5 CONCLUSIONS

Although the achieved classification accuracy (54.67%) does not seem to be very high, we have to consider that it is difficult to distinguish between 10 different music genres, even for a human. Thus, we are satisfied with the result. However, there is still plenty of room for improvement.

Librosa produces wide white margins around the spectrogram, which are useless for classification. The spectrograms could be cropped to ensure that only the spectrogram is present in each image.

Classification of a music genre based on spectrograms is not the most accurate. The low accuracy is most likely also due to the truncated music files. Each sound recording is only 30 seconds long. If the dataset contained full tracks, we would have a more representative sample, which would most likely improve the results.

A single track may belong to several different genres. Therefore, performing a multi-label classification and comparing the results would be reasonable. Alternatively, the multi-label classification model could be used within the classifier ensemble.

Features could be extracted from the sound recordings, and another classification model could be added to the classifier ensemble. The classification would likely be more accurate as these features contain additional information.

It might also be worth considering weighing the individual classifiers in the ensemble. As a result, each classifier would not necessarily contribute the same amount to the final ensemble prediction.

ACKNOWLEDGMENTS

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P2-0057).

REFERENCES

[1] [n. d.]. Spectrogram. https://en.wikipedia.org/wiki/Spectrogram Accessed: 2022-07-28.
[2] Shlomo Argamon, Kevin Burns, Shlomo Dubnov, and Roger B Dannenberg. 2010. Preprint from The Structure of Style: Algorithmic Approaches to Understanding Manner and Meaning Style in Music. , 45-58 pages. http://www.cs.cmu.edu/~rbd/
[3] Hareesh Bahuleyan. 2018. Music Genre Classification using Machine Learning Techniques. (4 2018). http://arxiv.org/abs/1804.01149
[4] Hmrishav Bandyopadhyay. 2022. Image Classification Explained. https://www.v7labs.com/blog/image-classification-guide
[5] James Chen. 2021. Neural Network. https://www.investopedia.com/terms/n/neuralnetwork.asp
[6] Ketan Doshi. 2021. Audio Deep Learning Made Simple - Why Mel Spectrograms perform better. https://ketanhdoshi.github.io/Audio-Mel/
[7] Ahmet Elbir, Hilmi Bilal Çam, Mehmet Emre Iyican, Berkay Öztürk, and Nizamettin Aydin. 2018. Music genre classification and recommendation by using machine learning techniques. In *2018 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 1–5.
[8] Dongmoon Kim, Kun-su Kim, Kyo-Hyun Park, Jee-Hyong Lee, and Keon Myung Lee. 2007. A music recommendation system with a dynamic k-means clustering algorithm. In *Sixth international conference on machine learning and applications (ICMLA 2007)*. IEEE, 399–403.
[9] Jash Mehta, Deep Gandhi, Govind Thakur, and Pratik Kanani. 2021. Music Genre Classification using Transfer Learning on log-based MEL Spectrogram. *Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021*, 1101–1107. https://doi.org/10.1109/ICCMC51019.2021.9418035
[10] Donald Michie, David J Spiegelhalter, and Charles C Taylor. 1994. Machine learning, neural and statistical classification. (1994).
[11] Andrada Olteanu. 2020. GTZAN Dataset - Music Genre Classification. https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification
[12] Nikki Pelchat and Craig M. Gelowitz. 2020. Neural Network Music Genre Classification. *Canadian Journal of Electrical and Computer Engineering* 43 (6 2020), 170–173. Issue 3. https://doi.org/10.1109/CJECE.2020.2970144
[13] Claude Sammut and Geoffrey I Webb. 2011. *Encyclopedia of machine learning*. Springer Science & Business Media. https://books.google.si/books?id=i8hQhp1a62UC
[14] Jim Samson. 2001. *Genre*. Vol. 1. Oxford University Press. https://doi.org/10.1093/gmo/9781561592630.article.40599
[15] Yi Shang and Benjamin W Wah. 1996. Global optimization for neural network training. *Computer* 29 (1996), 45–54. Issue 3.
[16] Garima Sharma, Kartikeyan Umapathy, and Sridhar Krishnan. 2020. Trends in audio signal feature extraction methods. *Applied Acoustics* 158 (2020), 107020.
[17] Sugianto Sugianto and Suyanto Suyanto. 2019. Voting-based music genre classification using melspectogram and convolutional neural network. *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 330–333.