# Music Genre Classification by Machine Learning Algorithms

Xuze Zhang*

Department of Mathematics, University College London, London, United Kingdom

*Corresponding author: xuze.zhang.21@ucl.ac.uk

**Abstract.** Working with audio data has been a machine learning issue that has received relatively little attention. With the fast emerging of machine learning and deep learning algorithms, benchmarks for the most recent groundbreaking deep learning research are frequently determined by how well it performs on text and image data. Additionally, models that use text and graphics are where deep learning has made the biggest strides. Speech and audio, equally significant data types, are frequently ignored in this. Audio data has many uses in daily life, ranging from converting spoken words to text, live translation, and employing audio and music for analysis or filtering. A crucial component of music information retrieval systems, content-based music genre classification has featured prominently as a consequence of the emergence of digital music on the Internet. The field of automatically classifying music genres is still incredibly young, and the claimed classification performance levels are currently relatively low.

**Keywords:** Music genre classification; Data analysis; Machine learning.

## 1. Introduction

The majority of music genre classification methods classify feature vectors—extracted from brief re-cording segments—into genres using pattern recognition algorithms. Timbral texture characteristics, rhythmic features, and pitch content features are the general divisions of the features used for music genre classification. Support vector machines (SVMs), nearest-neighbor (NN) classifiers, or classifiers that use Gaussian Mixture Models, Linear Discriminant Analysis (LDA), etc. are examples of frequently used classifiers.

This work proposes DWCHs1, a revolutionary feature extraction strategy for musical genre characterization. The precision of the classification phase is substantially enhanced by DWCHs, which imitate mu-sic sounds by building histograms on Daubechies wavelet coefficients in various frequency bands. The classification accuracy of the learning algorithms was evaluated on Dataset but used a multitude of feature combinations. The accuracy values are computed using ten-fold cross-validation. Numbers in parentheses are used to indicate standard deviations. SVM1 and SVM2 stand for the paired SVM and the one-versus-the-rest SVM, respectively. The graphic illustrates that, regardless of the method used, the accuracy rate of DWCHs is significantly increased after characteristics from DWCHs are extracted.[1]

Throughout another publication, a multilinear framework is applied to examine the issue of autonomous music genre classification. Three multilinear subspace analysis techniques were employed to feature extraction from the cortical representation of sound. NTF HOSVD and MPCA models are proposed and evaluated using datasets from GTZAN and ISMIR2004Genre. As per the statistics, the HOSVD model's ISMIR2004Genre dataset precision value is 80.95%. [2]

**Table 1.** Results of HOSVD in GTZAN and ISMIR2004 Genre dataset.

| | GTZAN | ISMIR2004Genre |
|---|---|---|
| **NTF** | 78.20%(3.82) | 80.47%(2.26) |
| **HOSVD** | 77.90% (4.62) | 80.95% (3.26) |
| **MPCA** | 75.01% (4.33) | 78.53% (2.76) |

# 2. Method

## 2.1. Decision Tree

A "test" on an attribute is defined by each internal node, the result of the test is represented by each branch, and the class label is displayed by each leaf node (decision taken after computing all attributes). The routes from root to leaf stand in for classification standards.

To find out features and distinctions, the author will examine a selection of songs from six major genres: pop, rap, rock, Latin, and electronic dance these six various musical genres in various properties. Actually, rap is perceived by most as having a strong sense of rhythm and quick phrases while the sound of Latin music makes people show their strong desire for dance. What's more, taking rock as an example, the author discover that the amplified electric guitar has historically been at the heart of rock music, which means, unlike Latin music, rock is much louder and much more enthusiastic. Regarding Electronic Dance Music. the author finds that although people would like to dance when Latin music and Electronic Dance Music are on playlists, people exhibit various emotions and feelings. They are comparable in certain respects because of this while others are not. In consequence, EDM tunes are most likely to have strong intensity and low valence and are least likely to be acoustic. Latin music is highly balenced and danceable. Rock songs are more likely to be live recordings and have low danceability, whereas rap compositions score higher for talkativeness and danceability. Pop, Latin, and EDM songs are more likely to be shorter in length than R&B, rap, rock, and EDM tracks. It's conceivable that instrumentality and key are not particularly effective at defining genres. The most essential trait in the decision tree model is talkativeness, which sets rap apart from the other classes on the initial decision. Songs that are challenging to dance to are then classified as Rock. Likewise, EDM music is reported to experience a high pace. Longer songs are then designated as R&B. Consequently, songs with a bunch of danceability are classified as Latin. Finally, the remaining songs are defined as pop. According to the decision tree model, 43.2% of songs were correctly classified into their respective genres overall, which is far less than what people actually require undoubtedly.

## 2.2. Random Forest

The bagging method is extended by the random forest algorithm, which uses feature randomness in addition to bagging to produce an uncorrelated forest of decision trees. The random subspace method, also known as feature bagging, creates a random subset of features that guarantees a low correlation between decision trees. The main distinction between decision trees and random forests is this. Random forests merely choose a portion of those feature splits, whereas decision trees take into account all possible feature splits. If the author return to the "should I surf?" example, the questions I would pose to arrive at the forecast might not be as thorough as those posed by someone else. By taking into consideration every possible variation in the data, the author can lessen the chance of overfitting, bias, and total variance, leading to predictions that are more accurate.

This process lessens the likelihood of overfitting and raises the precision of predictions. By starting a random forest model with 100 trees, total accuracy of 65.7% percent is displayed in the random forest model.

## 2.3. Deep Neural Network

Convolutional neural network (CNN)[3-6], a subclass of artificial neural networks that have gained dominance in a number of computer vision tasks, is gaining popularity in a number of fields, including radiology. Using a variety of building pieces, including convolution layers, pooling layers, and fully connected layers, CNN is intended to automatically and adaptively learn spatial hierarchies of features through backpropagation.

In a recurrent neural network (RNN)[7-10], a category of artificial neural networks where linkages between nodes can constitute a loop, the output from some nodes potentially affects subsequent input to the same nodes. It can operate in a contextually compelling manner owing to this. RNNs, which

are derived from feedforward neural networks, can assess variable-length input sequences by using their internal state (memory). They are capable of performing tasks like speech recognition or interlinked, unsegmented handwriting recognition due to this feature. A CNN contrasts with an RNN in that it possesses the capacity to evaluate temporal information—data that occurs in sequences, like syllables. Recurrent neural networks were developed explicitly to comprehend temporal information, whereas convolutional neural networks are unable to perform this function. As a result, CNNs and RNNs are employed for very different objectives, necessitating variations in the neural network architectures themselves.

For CNN: As shown in Figure 1, the author will use Librosa to produce mel spectrograms for the audio files once the author has divided the audio files into empty directories for each genre. The author must divide the data into a training set and a validation set now that the author has all of the data.

The final training accuracy was 82.37% after 50 epochs, while validation accuracy was 78.73%. There-fore, our validation results were extremely accurate.

For RNN: As shown in Figure 2, the audio file needs to be processed before the author can use the data from the GTZAN dataset as input for our RNN-LSTM neural network. Additionally, preprocessing entails taking the audio signal and extracting useful features. Mel-frequency cepstral coefficient (MFCC), which determines how bright a sound is, is thus one of the techniques to extract usable information from the signal. It can also be used to determine the sound's timbre (quality).
The model's accuracy on the training dataset was 76.3%, whereas, after 50 iterations on the test data, it averaged 68.93% accuracy.

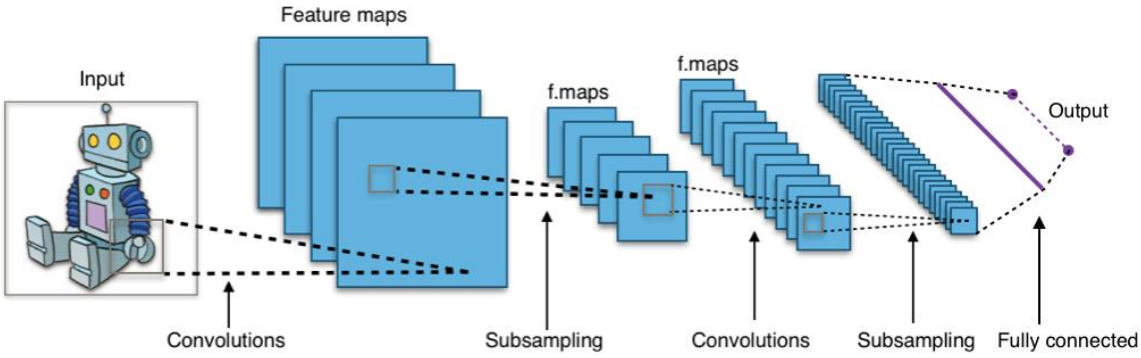


**Fig. 1.** The Architecture of CNN.

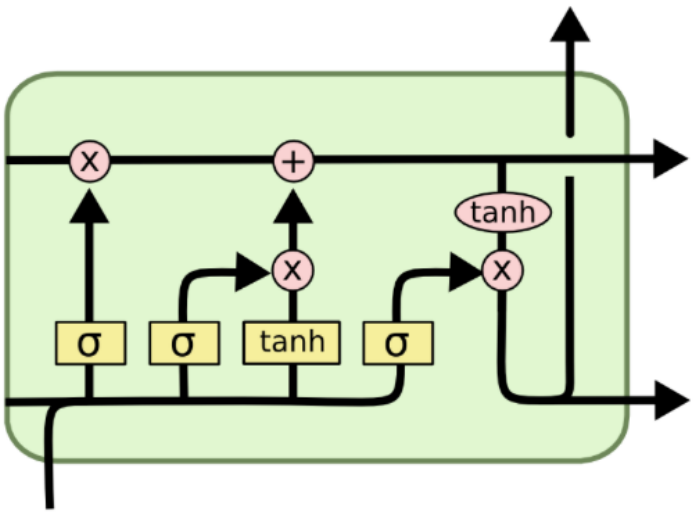


**Fig. 2.** The framework of LSTM network.

## 3. Experiments

Over-fitting your training data is a significant issue with decision trees. While over-fitting may produce 100% accuracy on your training data, it may produce disastrous outcomes on untested data. Limiting the depth of the nodes or pruning (removing arbitrary leaves after overfitting) are two strategies for reducing overfitting. For instance, the author can't tell if a piece of music is RAP since some characteristics (such as talkativeness) is above a particular value. Likewise, the author cannot guarantee that any song with a number lower than this is not a rap song.

Decision trees' poor accuracy is primarily caused by the inability to accurately calculate the fork value. The best strategy to increase accuracy is to include additional adjectives or qualities that describe mu-sic.

The study of CNNs is progressing, but knowledge of them is still limited. They may not perform well with new data and are frequently over-specified for really precise data. This is because it might be challenging to foresee or determine which kind of layer or activation function will be most effective in a certain application. Using a lot of data sets in the CNN model is a good technique to increase accuracy. In general, the accuracy of a CNN model increases with the number of datasets learned.

Another approach is to combine the use of RNNs and CNNs. The visual representation of audio in both frequency and time dimensions is called a spectrogram. CNN makes sense since spectrograms of songs are similar to images in that they each have unique patterns. By making the hidden state at time t de-pendent on the hidden state at time t-1, RNNs excel in comprehending sequential data. RNNs are far more effective at distinguishing the song's short and long-term temporal aspects since the spectrograms have a time component.

As shown in Table 2, there is no doubt that the accuracy will be far higher with CNN+RNN than with CNN or RNN alone as shown in the graph extracted from the following essay. [3]

**Table 2**. Accuracy in training and test dataset for different models.

| Method | Training Accuracy | Test Accuracy |
|---|---|---|
| Decision Tree | -- | 43.2% |
| Random Forest | -- | 65.7% |
| CNN | 82.37% | 78.73% |
| RNN | 76.3% | 68.93% |

## 4. Conclusion

In this paper, the author aims to conduct music genre classification task using machine learning methods. Three types of methods are included, such as decision tree, random forest and deep neural network. As for deep neural network, Convolution Neural Network(CNN) and Recurrent Neural Network(RNN). The study indicates that all CNNs with a parallel RNN block perform better than CNNs alone. Nevertheless, this does not automatically entail that the effectiveness will advance with the depth of RNN layers. Future research on automatic music genre classification will emphasize on supervised multilinear sub-space analysis utilizing a variety of and coupled models (e.g. CNN+RNN or CNN+RNN+ another model). The author aims to apply the machine learning and deep learning algorithms into arts, which can also bring a satisfying performance.

## References

[1] Cheng Qiyun, Sun Caixin, Zhang Xiaoxing, et al. Short-Term load forecasting model and method for power system based on complementation of neural network and fuzzy logic. Transactions of China Electrotechnical Society, 2004, 19(10): 53-58.

[2] Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology, 2011.

[3] Amjady N. Short-term hourly load forecasting using time series modeling with peak load estimation capability. IEEE Transactions on Power Systems, 2001, 16(4): 798-805.

[4] Ma Kunlong. Short term distributed load forecasting method based on big data. Changsha: Hunan University, 2014.

[5] SHI Biao, LI Yu Xia, YU Xhua, YAN Wang. Short-term load forecasting based on modified particle swarm optimizer and fuzzy neural network model. Systems Engineering-Theory and Practice, 2010, 30(1): 158-160.

[6] Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology, 2011.

[7] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

[8] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

[9] Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., ... & Iyengar, S. S. (2018). A survey on deep learning: Algorithms, techniques, and applications. ACM Computing Surveys (CSUR), 51(5), 1-36.

[10] Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. IEEE Access, 7, 53040-53065.