**Music Genre Classification and Comparison between Feature Based and CNN model**

**Venkatesh Nalluri**

UH\_ID: 2048174Department of Computer Science  
University of HoustonHouston, Texas, USA  
Vnallur3@cougarnet.uh.edu  
  
**Nithin Nerella**

UH\_ID: 2094680  
Department of Computer Science  
University of HoustonHouston, Texas, USA  
nnerella@cougarnet.uh.edu  
  
**Sai Asish Samineni**

UH\_ID: 2088595  
Department of Computer Science  
University of HoustonHouston, Texas, USA  
Ssaminen@cougarnet.uh.edu

**CONTENTS**

1. **Abstract 2**
2. **Survey 2**
3. **Development 6**

**3.1 Methodology 6**

**3.2 Dataset 6**

**3.3 Feature based classification 7**

**3.4 Neural Network based classification 9**

**4. Conclusion 12**

**5. References 12**

# 

# **Abstract**

# Software techniques have been playing a big role in music since the audio signals were made digital. The signal processing domain has found a lot of research and hence the music industry has started making use of signal processing techniques to shape music today. Artificial Intelligence (AI) and Machine Learning (ML) have also played a big role in the innovation in the way the music is composed, generated and consumed.AI led the music domain to allow different music learning approaches, music marketing and music generation as well as music synthesis. Music can be thought of as an audio signal with a certain pitch and tone. AI can make use of the fact that music is a semi-structured audio signal, and it has recognizable patterns that may be of use in the interpretation of music. The applications of AI even in the domain of Music is large, and in this project, we study the use of AI in the interpretation of music, compare the performance of 2 Machine Learning Approaches Feature Based Methodology – XGBoost Model and Neural Networks (CNN).

# **Survey**

We surveyed the domain of Music and Songs, in the domain, we found many relations of Music with AI such as Music generation with AI, AI Music generation accompanist, Music Classification & Interpretation and Musical expression understanding with AI.

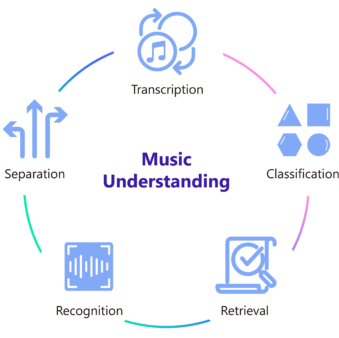


Figure 1: Music Understanding (Microsoft)

**Music Generation**

Music generation dates to 1234 A.D when a person tried to generate music using a piano roll. In the roll he specified the timing, the velocity, and frequency of the piano notes.

This field is of great interest for composers who are interested in generating a single music with different themes and pitches. Most approaches are learning approaches where the machine learning algorithm tries to learn from the audios, music scripts, keywords etc. to try to generate similar music. It learns the pattern that is present in that music and creates an imitation similar to what it has heard. This approach may be good, but it disables the creation of novel music. Hence research is today focused on autoencoders for generation which can generate music out of nothing.

In the [MuseGAN paper](#_cyto8v9h94pm) based on generative networks, under the framework of GAN’s 3 models are proposed for symbolic multi-track music generation. The 3 differ in network architectures and in the assumptions, are called jamming, the composer and the hybrid model. They showed generation of coherent music of four bars without any human input.

## **AI Accompanist**

AI can act as an accompanist as well. Work has been done for creation of AIs focused on a single task such as piano accompanist or violin etc. AI accompaniment involves the AI listening to the partner and synchronizing the music with the music of the partner. This is a much simpler task, and a lot has been accomplished in this field of study. The core of accompaniment is the harmony of music. To solve the coordination problem of automatic accompaniment of partners using AI, hidden Markov models, Probabilistic approaches, Chord construction with Neural Networks and now Autoencoders are being used.

A great project on this is the [AI Duet](#_cyto8v9h94pm) project by Google which takes some inputs from the user in the form of piano playing and then responds and plays a similar accompanist track.

## 

## 

Figure 2: AI Duet Application by Google (Input is blue keystrokes, Yellow keystrokes are generated in response by the AI)

## **Expressive Music Generation**

Machine Learning can be used to learn from original human sung songs or composed music to try to imitate the expressiveness and the structure that is usually the core part of what makes a human sung song different from a Machine generated one. Expressiveness of a music or song is hard to understand by a Machine and thus remains a difficult problem even today.

The [Jazz Transformer](#_cyto8v9h94pm) paper tries to understand and compare the differences between real Jazz music and AI generated Jazz music. Generative model Jazz Transformer utilizes Transformer-XL, a neural sequence model for modelling lead sheets of Jazz music. This paper analyses and answers the question as to why an AI generated music still cannot compare to the actual piece composed by a human and lacks expressiveness. The paper also defines some future goals on music composition to improve the expressiveness of AI models.

## 

## **AI Marketing with Music**

Music is now also being used for marketing purposes to understand the preference of a customer and use the insights for targeted advertisements and marketing.

An outstanding paper in journal [Artificial intelligence-driven music biometrics](#_cyto8v9h94pm) Surveys 386 customers in a five-day experiment with different types of music (enhanced by music-recognition biometrics). They found that the music impacts the cognition, and the user may have a change in preference due to the music.

## **Music Classification and Interpretation**

Music classification is an important problem space today with the rise of online music platforms such as Spotify, YouTube Music, Gaana etc. Classifying music can be difficult due to the music produced today electronically and through new and old instruments combined with multiple singers and sometimes even AI accompanists.

In this paper about [Music Genre Classification](#_cyto8v9h94pm), the authors studied the GTZAN data. They made the study where one group of classifier models is trained with randomly stratified data and the second group of classifiers is trained with data where the train and the test data do not have songs from the same artists. They analysed the differences between responses using Item response theory.

**Inspiration**

Our project is inspired by the domain of Music Classification and Interpretation. The music classification problem may look simple but the problem becomes more difficult when there are hundreds of thousands pieces of music being composed and songs being released through the internet. This becomes a hard problem for music platforms to manage all the music. To solve the issue, the music may be classified in some pre-defined domains such as jazz, pop, rock, metal. But there are always going to be arguments regarding the genre of a music or song as different users and even music experts may classify the music in a different class. In this research we find that this topic and problem is very interesting and with new Machine learning techniques, we can try to compare them with older more ‘feature’ dependent approaches.

Experts in the field have tried to understand sound and what differentiates one music from another music. First thoughts on this question always relate to the tone, harmonies and the beat of the sound. Some experts argue that the Timbre or “tone colour” of the instrument played is the most important factor to differentiate between musical instruments. Here lies yet another question, what makes a song different from the other?

A song is a musical composition of music and human voice. The human voice may have distinct pitches or fixed harmonics, while the music may be produced through an instrument or a combination of instruments. Hence the identification of the genre of the song is a challenging task and even experts may have different thoughts on the genre of the music. In the music domain, it is well known that music can be divided into genres, such as rock, metal, pop, hip hop etc. These genres are defined by the musical characteristics as well as the harmonics and distinct pitches, but a human can classify them just by listening to these songs. A computer can only understand numbers, not images, not text, not audio, not any other format except numbers. Hence, we need to find a way to create a way that a computer may be able to interpret audio.

# **Development**

* 1. **Methodology**

Most of the research for music genre classification is based on using the Neural networks, Recurrent Neural Networks such as LSTMs, GRUs, Autoencoders and Variational Autoencoders and now even Transformers. While the results may be great, it is difficult to use such a model and retrain it again and again due to its size and the number of songs being uploaded every day. The large amount of data needs to be trained and large models increase the resources that are required to perform training and inference in real-time. This may be easier with high end Accelerated Computing servers but is it even necessary to use such models?

Hence, in this project we focus on comparing the performance of 2 Machine Learning Approaches, the first approach is to extract features of the music and use those numeric features to try to classify the music. The second approach is to use raw audio features and use Convolutional Neural Network to analyse the music sequentially and classify it. As noted here, the CNN is computationally expensive and requires a Graphics Processor (GPU) to train. While for features we may use any Machine Learning model able to work on Tabular data but for this task to make it fair we have used XGBoost Machine Learning model.

* 1. **Dataset**

The GTZAN dataset can be used for achieving this task. We try to apply Machine Learning techniques to classify music into genres using the dataset. The GTZAN dataset is the most-used public dataset for evaluation

in machine learning research for music genre recognition tasks. Music files were collected in 2000-01 from different sources, mainly CDs, radio, mic recordings, recording conditions differed for all these audios. The

GTZAN audio dataset contains 1000 tracks of 30 second length. There are 10 genres, each containing 100 tracks which are all 22050 Hz Mono 16-bit audio files in .wav format. The genres are:

| **Genres** | |
| --- | --- |
| Blues | Jazz |
| Classical | Metal |
| Country | Pop |
| Disco | Reggae |
| Hip-Hop | Rock |

Table 1: Genres in the GTZAN dataset

## 

## **Feature Based Methodology**

Our idea about music classification was that the music usually has a specific pitch, timbre, spectral behaviour to it that may define its unique characteristics and after some visualization we were able to confirm this through plotting spectrums of different classes.

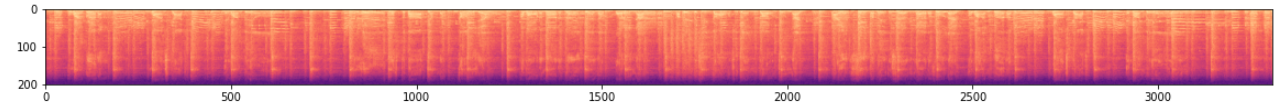


Figure 3: Spectrogram for Music with class ‘blues’

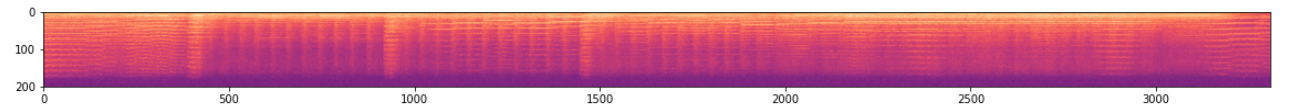


Figure 4: Spectrogram for Music with class ‘classical’

There is a significant difference between their respective spectrums and thus we can use spectrogram features, and many other features to build a numeric table for features.

Our feature extraction-based methodology is as follows:

### **Audio chopping**

All the audios (1000 tracks) are chopped into 10 sec segments each. This results in an increased amount of data (3000 tracks each of 10 seconds) and makes our model more robust. This also enables us to use more data for training and testing.

### **Creating Data loaders**

We create data loaders to process audio in sequence, this process is for code encapsulation.

**Applying feature extraction**

We apply feature extraction to all the audios in the data. Feature extraction is done using the **Librosa audio processing library** in Python. The features are extracted and their mean and variance are calculated to create a total of 54 features:

| **Mean and Variance of the Audio features** | |
| --- | --- |
| Chromagram | Root-mean-square (RMS) |
| Spectral centroid | Spectral bandwidth |
| Tempogram: local autocorrelation | Roll-off frequency |
| Mel-scale spectrogram | Zero-crossing rate of an audio |

Table 2: Features Extracted from Audio segments

Saving features into a comma separated file, we save the data into a csv file, then split it into 80% train and 20% test.

* **Applying Minimum-Maximum Scaling**

The data is scaled between the values 0-1 for better optimization and faster convergence.

* **Label Encoding**

We encode the label strings (rock, pop, metal etc) into numerics (0,1,2,...).

* **XGBoost**

We use the XGBoost classifier with estimators equal to 1000 and learning rate set to 0.05 and train it on the train set.

* **Result**

We tested the XGBoost model on the test set we put aside. The results were great, and we got an **accuracy** of **85.64%**.

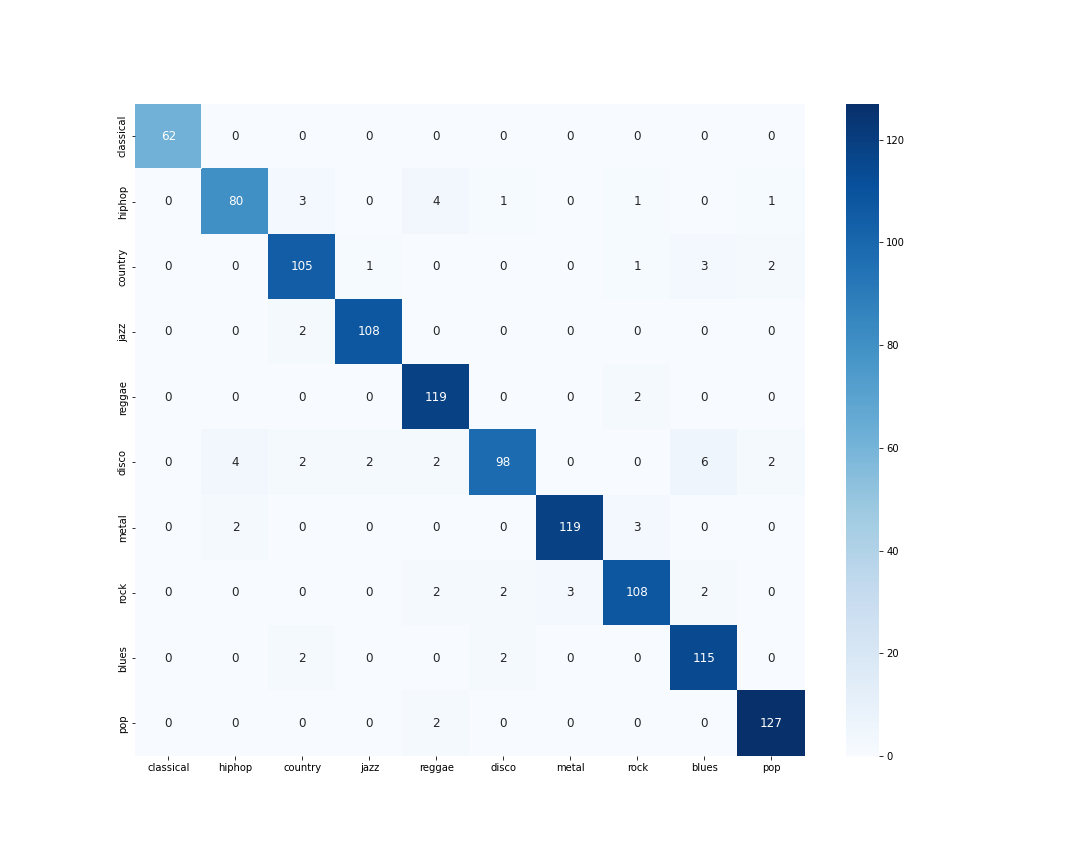
* **The confusion matrix**
* ****

Figure 5: Confusion Matrix of XGBoost Classification on Test set

* 1. **Neural Network based classification**

The neural network for music classification needs to take into account the sequence of the input, hence we use neural network architecture presented by [[7](#_mdmss7gwrrmd)]. The network consists of three Convolution 1-D blocks for

processing the raw audio and extracting features, four Convolutional 2-D blocks to find patterns in the extracted features. The network is deep and illustrates further on the fact that music classification is a complex problem and requires very deep neural networks with multiple convolution or recurrent layers to give good classification results.

* **Model Inputs**

The audios were chopped into 10 seconds segments each and split into 80% Train and 20% Validation. The data loaders were created with a batch size of 32. The labels were encoded using the same method as used in feature-based method.

* **Model Architecture**

The model can be summarized as using initial sequential 1D convolutions to extract features with 11, 51, and 111 filter size, from the audio separately. These features are then stacked on top of each other to create a 2D matrix. The 2D CNN blocks are used to process these features and next are passed to ResNet blocks. Finally the ResNet features are passed to the Classifier with three Linear Layers and 10 output units.

* **Model Parameters**
  + Batch size: 32
  + Epochs: 20
  + Optimizer: Adam
  + Learning rate: 5e-2
  + Loss: Categorical Cross Entropy
  + Metrics: Accuracy
* **Results**

Testing was done on the validation set that we used in training. The model was only able to achieve an **accuracy** of **80.12%**. One of the reasons can be that we have used less number of epochs to train the model. But that is beside the point of this comparison. In use cases where application building is as simple as extracting some defined set of features which define the data, it is much simpler to use feature extraction and train the model on these features as we did here.

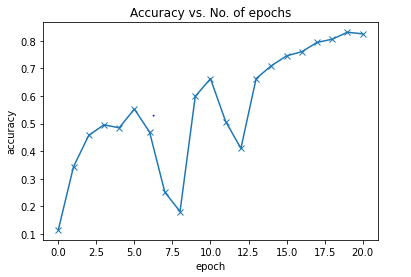


Figure 6: Model Training Accuracy vs Epochs

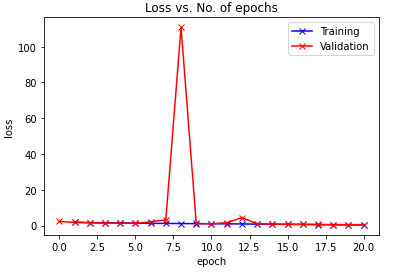


Figure 7: Model Training / Validation Loss vs Epochs

* + - **Model architecture is added in the Appendix.txt file.**

1. **Conclusion**

We compared the performance of **Feature Based classification approach** (XGBoost Model) and **Neural Network** **based feature extraction and classification** approach on GTZAN dataset. The Feature Based Classification Model - XGBoost Model resulted 85.64% Accuracy and Neural Network Model resulted 80.12% Accuracy. The sole purpose was to demonstrate how easily the model can be trained to perform well on classification of music just based on the features of the audio. While neural networks which are difficult to train on small datasets take a lot of training time and even after 20 epochs do not fit the data properly.

However, it can be said that our model only performed well because the data set is about 20 years old, and the music industry has changed a lot during this time. The music today contains all kinds of new electronic musical instruments, electronic compositions, software tuning, modified vocals, noise filters and many more pre-processing and postprocessing techniques that make music today far different from what it was 20 years ago. Due to these reasons our model may or may not perform well on newer datasets and may require complex neural networks to solve this problem.

# **References**

1. [Dong, Hao-Wen, et al. "Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment." Thirty-Second AAAI Conference on Artificial Intelligence. 2018.](https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewPaper/17286)
2. [“AI DUET.” *Experiments with Google*,](https://www.frontiersin.org/research-topics/10088/music-and-ai.) https://[experiments.withgoogle.com](https://experiments.withgoogle.com/ai/ai-duet/view/)
3. [Wu, Shih-Lun, and Yi-Hsuan Yang. "The jazz transformer on the front line: Exploring the shortcomings of ai-composed music through quantitative measures." arXiv preprint arXiv:2008.01307 (2020).](https://arxiv.org/abs/2008.01307)
4. [Rodgers, Waymond, et al. "Artificial intelligence-driven music biometrics influencing customers’ retail buying behavior." Journal of Business Research 126 (2021): 401-414.](https://www.researchgate.net/publication/348591607_Artificial_Intelligence-Driven_Music_Biometrics_Influencing_Customers'_Retail_Buying_Behavior)
5. [Mulongo, Benedith. "Analyzing music genre classification using item response theory: A case study of the GTZAN data." (2020).](http://kth.diva-portal.org/smash/record.jsf?pid=diva2:1472081)
6. [Caramiaux, Baptiste, and Marco Donnarumma. "Artificial Intelligence in Music and Performance: A Subjective Art-Research Inquiry." Handbook of Artificial Intelligence for Music. Springer, Cham, 2021. 75-95.](https://hal.archives-ouvertes.fr/hal-03046229/document)
7. [“Music and Ai.” *Frontiers*, https://www.frontiersin.org/research-topics/10088/music-and-ai.](https://www.frontiersin.org/research-topics/10088/music-and-ai.)
8. [Sturm, Bob L. "The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use." arXiv preprint arXiv:1306.1461 (2013).](https://arxiv.org/abs/1306.1461)
9. [Shi, Shuo, Shuting Xi, and Sang-Bing Tsai. "Research on Autoarrangement System of Accompaniment Chords Based on Hidden Markov Model with Machine Learning." Mathematical Problems in Engineering 2021 (2021).](https://www.hindawi.com/journals/mpe/2021/6551493/)
10. [Dhariwal, Prafulla, et al. "Jukebox: A generative model for music." arXiv preprint arXiv:2005.00341 (2020).](https://openai.com/blog/jukebox/)
11. [Microsoft. (2021, November 21). AI Music. Microsoft Research. https://www.microsoft.com/en-us/research/project/ai-music/](https://www.microsoft.com/en-us/research/project/ai-music/)