Music Genre Classification

Manuel Cortez-Muñoz

- **University:** Universidad Autónoma de San Luis Potosí
- **Professor:** Dr. Juan Carlos Cuevas Tello
- **Generation:** 2021
- **UASLP Code:** 340664
- **Course:** Machine Learning
- **Date:** May 28, 2024

Abstract

In the digital era, the expansive growth of music libraries on streaming platforms necessitates effective organization and retrieval systems. This project explores the development and implementation of automatic music genre classification models using machine learning techniques. Leveraging the GTZAN dataset, we employed both a Multilayer Perceptron (MLP) and a Convolutional Neural Network (CNN) to classify music into distinct genres based on audio features such as Mel-Frequency Cepstral Coefficients (MFCCs) and Fast Fourier Transform (FFT) coefficients. The MLP model achieved an accuracy of 55.58%, while the CNN model outperformed with an accuracy of 73.35%. These results highlight the potential of deep learning models in automating music genre classification, offering a promising avenue for enhancing music information retrieval systems.

1 Introduction

In today's digital age, music has become an integral part of our lives, thanks to the accessibility offered by streaming platforms like Spotify and Apple Music. With just a few clicks, anyone can immerse themselves in a vast ocean of musical content.

According to data from the Spotify Community, the platform hosts an impressive collection of over 40 million songs. And this number continues to grow steadily as new releases are added to Spotify's ever-expanding library.

However, with such a vast array of music available, many people find it challenging to organize and manage their listening experiences. One effective

method of sorting through this abundance of music is by categorizing songs into genres, based on their rhythmic structure, harmonic content, and instrumentation. Musical genres serve as convenient labels crafted by humans to help navigate and understand the diverse world of music. Yet, they are fluid and dynamic, shaped by a complex interplay of factors such as public perception, marketing trends, historical influences, and cultural contexts.

In this digital landscape, hierarchical genre structures play a crucial role in organizing the immense collection of music available online. Currently, the task of annotating musical genres often relies on manual efforts. However, the emergence of automatic musical genre classification presents an exciting opportunity to streamline and enhance this process. By harnessing the power of machine learning algorithms, automatic classification systems have the potential to assist or even replace human users in categorizing music, thereby enriching music information retrieval systems.

2 Problem Definition and Motivation

The origin of my passion for working with audio can be traced back to my love for music and my fascination with the intricate complexities found within each musical piece. Beyond the emotional and artistic weight carried by songs, I am captivated by the scientific and engineering elements that underpin their creation.

My interest was sparked by the functionality of the Shazam application, which, in 2002, it allowed users to identify songs simply by dialing "2580" on their phones and holding them up to the music. This innovative approach, driven by Wang's algorithm, intrigued me and prompted me to delve deeper into the field of music recognition.

Figure 1: Shazam 2002

Upon reading Avery Li-Chun Wang's paper, "An Industrial-Strength Audio Search Algorithm," I was further compelled to explore the realm of music recognition. However, I recognized the need to start with a more manageable project within my reach before tackling such an extensive endeavor. This realization led me to embark on the creation of a music classifier.

Automatic musical genre classification not only offers a practical solution for organizing vast music collections but also serves as a foundation for developing and evaluating features for various types of content-based analysis of musical signals. By automating the classification process, researchers and music enthusiasts can gain valuable insights into the underlying characteristics of musical compositions, revealing new perspectives and patterns within the diverse landscape of musical expression.

3 Objectives

This music classifier project is has three primary objectives. Firstly, it aims to achieve precise genre classification. Secondly, it seeks to benchmark its performance against existing methodologies, as illustrated in Table 1, showcasing methods utilized by other researchers. Moreover, the project endeavors to explore the potential advantages of genre-based music classification. Recognizing the capacity of automated genre classification to streamline processes, it aims to play a pivotal role in a comprehensive music information retrieval system.

Furthermore, the project aims to establish a robust framework for developing and evaluating features that characterize musical content. These features can

be harnessed for various tasks such as similarity retrieval, classification, segmentation, and audio thumbnailing, thereby laying the groundwork for numerous proposed audio analysis techniques within the music domain.

Table 1: Various studies that have shown capability to perform genre classification

Author(s)	**Dataset**	**Model**	**Accuracy**
Sturm (2013) [12]	GTZAN	Sparse Representation Classification	83.00%
Bergstra et al. (2006) [3]	GTZAN	ADABOOST	82.50%
Li et al. (2003) [7]	GTZAN	Support Vector Machines	78.50%
Lidy et al. (2007) [8]	GTZAN	Sequential Minimal Optimization	76.80%
Benetos and Kotropoulos (2008) [2]	GTZAN	Non-negative Tensor Factorization	75.00%
Bahuleyan (2018) [1]	Audio Set	CNN	65.00%
Tzanetakis and Cook (2002) [13]	GTZAN	Gaussian Mixture Model	61.00%

4 Algorithms

Sound is all around us, but what exactly is it? Imagine strumming a guitar string. The string doesn't just move back and forth, it creates tiny wiggles in the air around it. These vibrations cause air molecules to bump into each other, creating areas of high and low pressure. This creates a wave that travels through the medium, which can be air, water, or even solid objects.

We can represent these vibrations with a waveform, a graph that shows how the pressure changes over time. The amplitude of the wave corresponds to the height of the peaks and valleys, which tells us how strong the vibration is. A higher amplitude generally translates to a louder sound.

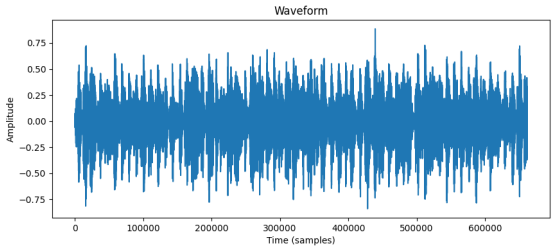

Figure 2: Waveform representation of sound

The frequency of the wave, on the other hand, refers to how often these peaks and valleys repeat themselves in a given time frame. It's measured in Hertz (Hz) and determines the pitch we perceive. Understanding complex sounds is crucial for many machine learning applications. But how do we break down that rich symphony of instruments and voices into a language machines can comprehend?

Imagine a complex sound as a painting composed of countless brushstrokes. The Fast Fourier Transform (FFT) acts like a special magnifying glass, revealing

the individual frequencies that contribute to the overall sound. By performing an FFT, we obtain a power spectrum, a map displaying the intensity of each frequency. This is essentially the "fingerprint" of the sound, capturing its frequency content.

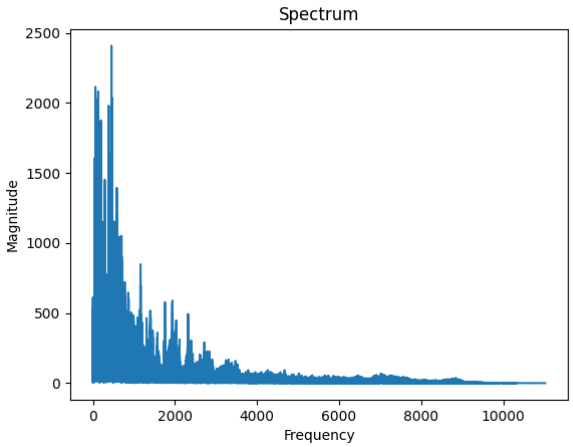

Figure 3: Power spectrum obtained from FFT

However, the power spectrum has a limitation – it lacks time information. In music, for example, the way a sound evolves over time is just as important as its overall frequency makeup. This is where the Short-Time Fourier Transform (STFT) comes to the rescue. Imagine slicing the sound into small windows and performing an FFT on each slice. The STFT does just that, creating a spectrogram – a visual representation of how the frequency content of the sound changes over time.

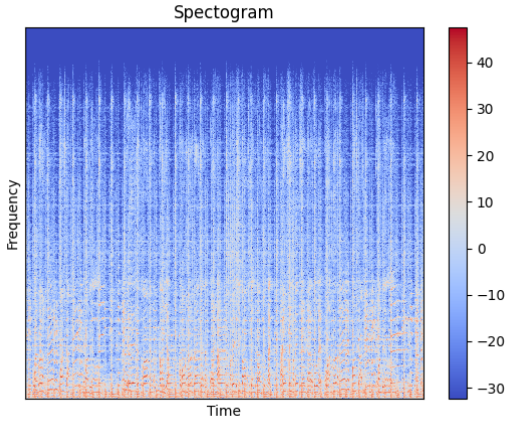

Figure 4: Spectrogram showing frequency content over time

From a waveform, we can extract various features to train a deep learning model. Among these features, Mel-Frequency Cepstral Coefficients (MFCCs) hold a special place. Inspired by the human auditory system, MFCCs capture the "timbre" or texture of a sound – its unique sonic fingerprint. By focusing on frequencies that humans perceive more distinctly, MFCCs offer a more efficient representation for tasks like music genre classification.

Unlike spectrograms, MFCCs are a more compact representation of audio data, making them computationally efficient for machine learning models. Additionally, by mimicking human hearing, MFCCs allow machines to "hear" the world in a way that's more relevant to how we perceive sound. This makes them invaluable for tasks like speech recognition and music information retrieval.

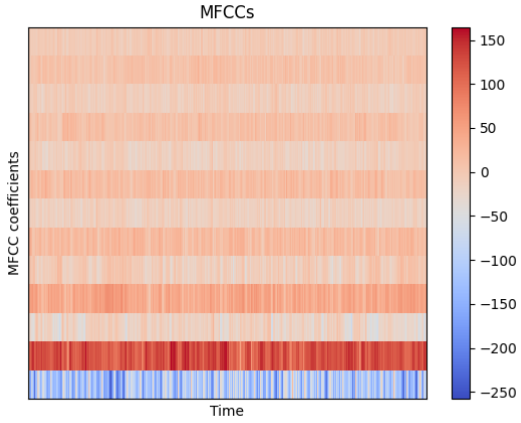

Figure 5: Mel-Frequency Cepstral Coefficients (MFCCs)

4.0.1 Fast Fourier Transform (FFT)

The Fast Fourier Transform (FFT) is a fundamental algorithm in scientific computation, allowing for the rapid computation of the discrete Fourier transform and its inverse. Widely employed across various domains including signal processing, image analysis, and scientific simulations, its efficiency in analyzing and transforming data between time and frequency domains is unparalleled.

4.0.2 Short-Time Fourier Transform (STFT)

The Short-Time Fourier Transform (STFT) is a signal processing technique that examines how a signal's frequency content evolves over time. By segmenting the signal into short segments and computing the Fourier transform for each segment, it generates a time-frequency representation. Noteworthy for its real-time feasibility and simplicity, the STFT's efficacy is exemplified in applications such as speech time-scale modification, where it produces high-quality output surpassing existing methods.

5 Methodology

My project entails coding a Multilayer Perceptron (MLP) and a Convolutional Neural Network for music genre classification using Python and the Librosa library, leveraging features extracted from a musical dataset. This endeavor integrates techniques from Artificial Intelligence (AI), focusing on constructing intelligent agents capable of perceiving their environment and influencing it. Central to this effort is the utilization of Machine Learning (ML), a paradigm through which computer systems learn from data to improve performance on

specific tasks over time. Moreover, Deep Learning (DL), a specialized branch of ML, will be employed, emphasizing the use of artificial neural networks characterized by multiple layers for enhanced computational capability.

5.1 Dataset

The dataset chosen for this project is the most popular among the authors, it the GTZAN Dataset - Music Genre Classification. A collection of 10 genres with 100 audio files each, all having a length of 30 seconds (the famous GTZAN dataset, the MNIST of sounds).

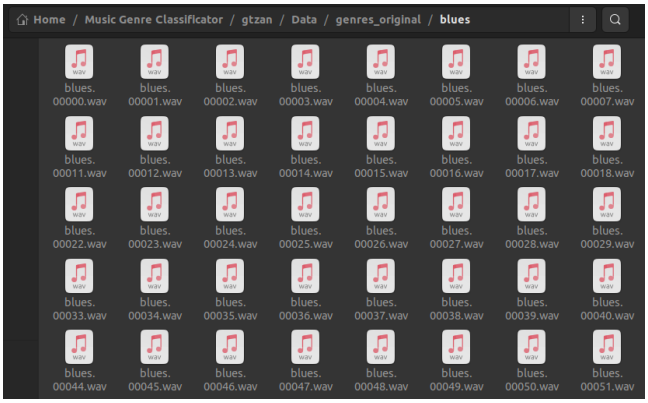

Figure 6: GTZAN dataset

5.2 Feature Extraction

5.2.1 Fast Fourier Transform Coefficients

Fast Fourier Transform (FFT) coefficients refer to the numerical values obtained from applying the FFT algorithm to a discrete set of data. These coefficients represent the amplitudes and phases of the individual sinusoidal components that make up the original signal in the frequency domain. FFT coefficients are crucial for various applications such as signal processing, spectral analysis, and filtering, as they provide insights into the frequency content of the input data and enable efficient manipulation and transformation between the time and frequency domains.

5.2.2 Mel Frequency Cepstrum Coefficients (MFCCs)

Mel Frequency Cepstrum Coefficients (MFCCs) are pivotal in audio recognition, capturing the subjective pitch and frequency content of audio signals effectively. Derived through a series of steps including Fourier transform, mel scale mapping, logarithmic conversion, and discrete cosine transform, MFCCs provide a

compact representation of audio data. Their increasing adoption in music information retrieval tasks such as genre classification and audio similarity measures underscores their significance in the field.

Table 2: Comparison of Audio Features

Author(s)	**Features**
Sturm (2013) [12]	MFCCs, OSC, Spectral centroid, Spectral flux, Spectral rolloff, ZCR
Bergstra et al. (2006) [3]	FFTCs, LPC, MFCCs, RCEPS, Spectral centroid, Spectral rolloff, Spectral spread, ZCR
Li et al. (2003) [7]	Low Energy, MFCCs, Spectral centroid, Spectral flux, Spectral rolloff, ZCR
Lidy et al. (2007) [8]	RH, RP, SSD
Benetos and Kotropoulos (2008) [2]	AC, AFF, AP, ASF, ASS, ASC, LAT, MFCCs, SONE, SRF, TL, ZCR
Bahuleyan (2018) [1]	Chroma features, CM, MFCCs, RMSE, Spectral bandwidth, Spectral centroid, Spectral contrast, Spectral rolloff, Tempo, ZCR
Tzanetakis and Cook (2002) [13]	DS, EAC, FWR, LEF, LPF, MFCCs, MR, Spectral centroid, Spectral flux, Spectral rolloff, ZCR

6 Model Implementation

6.1 Multilayer Perceptron (MLP)

This project involves developing a Multilayer Perceptron (MLP) for music genre classification using Python and the Librosa library. The model leverages features extracted from the GTZAN dataset.

The MLP is constructed as a sequential Keras model. The input layer is configured to match the shape of the dataset's features. The architecture includes three hidden layers: the first with 512 neurons, the second with 256 neurons, and the third with 64 neurons. To mitigate overfitting, both L2 regularization and dropout techniques were employed. The ReLU activation function is used for all hidden layers. The output layer consists of 10 neurons, corresponding to the ten music genres in the dataset, and uses the softmax activation function for multi-class classification.

6.1.1 Training Process

During the training phase, various data splits were tested to determine the optimal configuration, with a 70% training and 30% testing split proving most effective. Several learning rates were experimented with, and a learning rate of 0.0001 was found to provide the best balance between training duration and model accuracy. The model was trained using the Adam optimizer over 70 epochs. The loss function utilized was sparse categorical crossentropy, and a batch size of 32 was employed.

6.1.2 Evaluation Metrics

The performance of the model was monitored throughout the training process using metrics for both loss and accuracy on the validation set. The final accuracy on the test set was reported upon completion of the training process, providing a quantitative measure of the model's classification performance.

6.2 Convolutional Neural Network (CNN)

In addition to the Multilayer Perceptron (MLP), a Convolutional Neural Network (CNN) was incorporated to enhance classification accuracy. CNNs offer a potent methodology, particularly advantageous for audio classification tasks, as they can effectively process Mel-Frequency Cepstral Coefficients (MFCCs) and spectrograms akin to images.

The CNN architecture comprises three convolutional layers, each wielding a (3,3) kernel and 32 neurons, all activated by Rectified Linear Units (ReLU). MaxPooling layers, employing a (3,3) kernel and (2,2) stride with 'same' padding, were strategically inserted.

Subsequently, a flattening layer precedes a dense layer populated by 64 neurons, activated via ReLU, while a dropout layer is thoughtfully introduced to mitigate overfitting.

The network's output layer hosting 10 neurons, mirroring the ten music genres, governed by a softmax activation function.

6.2.1 Training Process

The model's training regimen was augmented with a validation set, optimizing its efficacy. With a training set constituting 75% of the data and a validation set comprising 20%, the Adam optimizer, with a learning rate of 0.0001, steered the learning process. A batch size of 32, previously proven effective in the MLP model, was adopted, and training persisted over 40 epochs, a duration deemed sufficient for achieving a nice performance.

6.2.2 Evaluation Metrics

Throughout the training phase, the model's proficiency was meticulously tracked, gauging loss and accuracy metrics on the validation set. Upon training completion, the model's accuracy on the test set was reported, furnishing a tangible measure of its classification prowess.

7 Results and Discussion

7.1 Multilayer Perceptron Model

The Multilayer Perceptron (MLP) model achieved an accuracy of 55.58%. Given the simplicity of the model architecture, this accuracy is quite satisfactory. The

model consists of only three hidden layers, which makes it computationally efficient while still providing a reasonable performance on the music genre classification task.

The accuracy plot (Figure 7) illustrates the learning behavior of the model over 70 epochs. Both training and testing accuracy show a steady improvement, indicating that the model is learning effectively from the data without significant overfitting. The training accuracy gradually increases, converging towards the test accuracy, which suggests that the model generalizes well to unseen data.

The loss plot (Figure 8) further supports the model's performance. The training and validation loss decrease consistently, demonstrating that the model is effectively minimizing the loss function over time. The convergence of the training and test loss curves indicates that the model is not overfitting and maintains a good balance between bias and variance.

When compared to another author's models with more complex architectures and additional features, the MLP's performance is commendable. Although state-of-the-art models might achieve higher accuracy, they often require more computational resources and sophisticated feature engineering. This MLP model, with its straightforward design and moderate accuracy, provides a good baseline for music genre classification tasks.

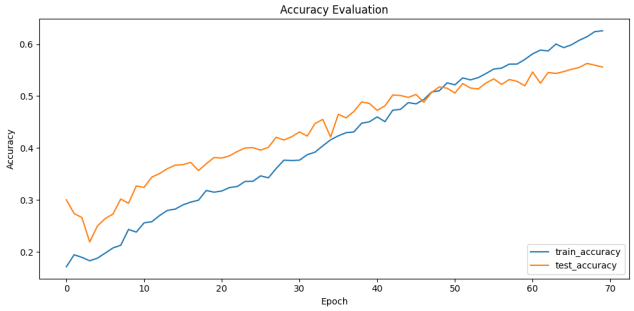

Figure 7: Accuracy Evaluation of MLP Model

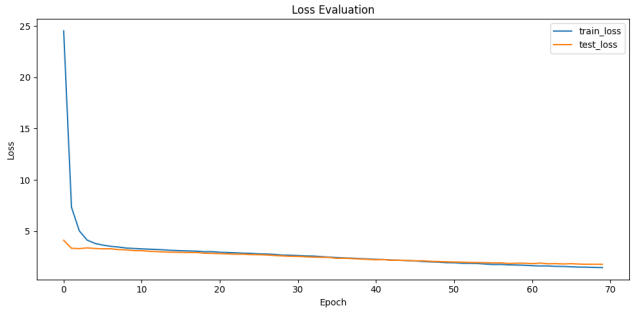

Figure 8: Loss Evaluation of MLP Model

Overall, the results indicate that the implemented MLP model is a viable option for music genre classification, especially in scenarios where computational efficiency is prioritized. Future work could explore more complex architectures, additional features, and data augmentation techniques to further improve the classification performance.

7.2 Convolutional Neural Network Model

The Convolutional Neural Network (CNN) model achieved an accuracy of 73.35%. This result indicates that the CNN architecture is a viable option for implementing a music genre classifier, primarily due to the effectiveness of convolutional kernels in extracting relevant features from the input data.

When comparing this performance to other models in the literature, such as the one implemented by Bahuleyan [1], who achieved an accuracy of 65% using an ensemble model that combines a Convolutional Neural Network (CNN) based on VGG-16 with XGBoost (XGB) using only spectrograms, our CNN model demonstrates competitive results.

The accuracy plot (Figure 9) shows the training and validation accuracy over 40 epochs. The steady increase in both training and validation accuracy demonstrates that the model is learning effectively without significant overfitting. The model's validation accuracy closely follows the training accuracy, indicating good generalization to unseen data.

The loss plot (Figure 10) further supports this conclusion. Both training and validation loss decrease consistently, indicating effective optimization. The convergence of training and validation loss curves suggests that the model maintains a balance between underfitting and overfitting.

Compared to the performance of other models listed in Table 1, our CNN model shows a commendable performance. While models with more sophisticated architectures or additional features might achieve higher accuracy, the simplicity and effectiveness of our CNN model make it an attractive option for music genre classification tasks.

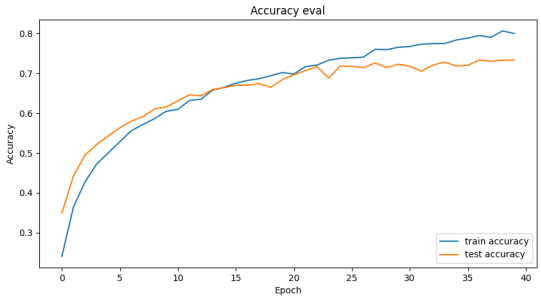

Figure 9: Accuracy Evaluation of CNN Model

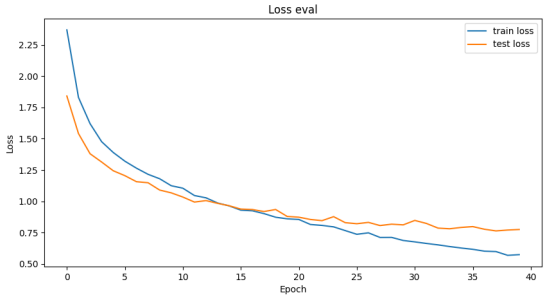

Figure 10: Loss Evaluation of CNN Model

Overall, the results indicate that the implemented CNN model is highly effective for music genre classification. Future work could explore the integration of additional features, more complex architectures, and data augmentation techniques to further enhance the classification performance.

8 Conclusion

This project demonstrates the development and evaluation of two machine learning models, a Multilayer Perceptron (MLP) and a Convolutional Neural Network (CNN), for the task of music genre classification. The MLP model achieved an accuracy of 55.58%, indicating its capability to handle basic genre classification tasks with moderate computational requirements. In contrast, the CNN model achieved a higher accuracy of 73.35%, highlighting its effectiveness in extracting intricate features from audio data. These results underscore the efficiency and potential of CNNs in music genre classification, especially for complex audio signals.

Future work could focus on incorporating more sophisticated architectures, such as Long Short-Term Memory (LSTM) networks, which may be better suited for capturing temporal dependencies in audio data. Additionally, utilizing a more comprehensive dataset, integrating an API to access extensive music metadata, and expanding the number of genres could further enhance classification performance. Data augmentation techniques and advanced feature extraction methods should also be explored to develop more robust and accurate music information retrieval systems. These advancements have the potential to significantly improve the precision and reliability of genre classification models.

References

[1] Hareesh Bahuleyan. *Music genre classification using machine learning techniques*. Tech. rep. 2018.

[2] Emmanouil Benetos and Constantine Kotropoulos. "A tensor-based approach for automatic music genre classification". In: *16th European Signal Processing Conference*. 2008, pp. 1–4.

[3] James Bergstra et al. "Aggregate features and Adaboost for music classification". In: *Machine Learning* 65.2-3 (2006), pp. 473–484.

[4] Keunwoo Choi, György Fazekas, and Mark Sandler. *Explaining deep convolutional neural networks on music classification*. Tech. rep. 2016.

[5] D. Griffin and J. Lim. "Signal estimation from modified short-time Fourier transform". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32.2 (1984), pp. 236–243.

[6] J. R. Johnson and R. W. Johnson. "Challenges of computing the fast Fourier transform". In: *DARPA Conference*. Citeseer, 1997.

[7] Tao Li, Mitsunori Ogihara, and Qi Li. "A comparative study on content-based music genre classification". In: *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '03. 2003, pp. 282–289. URL: https://doi.org/10.1145/860435.860487.

[8] Thomas Lidy et al. *Combining audio and symbolic descriptors for music classification from audio*. Tech. rep. 2007.

[9] Tom M. Mitchell. *Machine learning*. Vol. 1. McGraw-Hill, 1997.

[10] Nthabiseng Ndou, Ritesh Ajoodha, and Adnan Jadhav. "Music genre classification: A review of deep-learning and traditional machine-learning approaches". In: *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. 2021, pp. 1–6.

[11] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 1995.

[12] Bob L. Sturm. "On music genre classification via compressive sampling". In: *2013 IEEE International Conference on Multimedia and Expo (ICME)*. 2013, pp. 1–6.

[13] George Tzanetakis and Perry Cook. "Musical genre classification of audio signals". In: *IEEE Transactions on Speech and Audio Processing* 10.5 (2002), pp. 293–302.

[14] Unknown Author. *Introduction to TensorFlow*. Accessed: 2024-05-17. 2024. URL: https://www.tensorflow.org/learn.

[15] Unknown Author. *Shazam*. Accessed: 2024-05-17. 2024. URL: https://www.shazam.com/company.

[16] Avery Wang et al. "An industrial strength audio search algorithm". In: *ISMIR 2003*. Washington, DC, 2003, pp. 7–13.