Automated Music Genre Classification with Deep Learning Techniques

Dhanyapriya Somasundaram, *Misha Seroukhov*
dhanyapriyas@arizona.edu, mseroukhov@arizona.edu
MIS Department, Eller College of Management, University of Arizona, Tucson, AZ

***Abstract*—This research delves into the application of deep learning techniques for music genre classification, a pivotal aspect of audio data analysis that enhances various multimedia and digital interaction systems. Leveraging the capabilities of Artificial Neural Networks (ANNs) and Convolutional Neural Networks (CNNs), this study aims to refine the accuracy and efficiency of classifying music into distinct genres. We utilized a comprehensive dataset of audio tracks, applying robust feature extraction methods such as Mel-frequency cepstral coefficients (MFCCs) and spectral features to prepare the data for machine learning. The ANNs were employed to establish a baseline for classification accuracy, while CNNs were further explored to harness their spatial feature recognition capabilities, particularly through spectrogram analysis. The results demonstrated that CNNs, with their advanced pattern recognition architecture, significantly outperformed the ANNs, achieving a higher classification accuracy. The findings underscore the potential of CNNs in handling complex audio classification tasks and pave the way for future research into more scalable audio analysis systems.**

***Index Terms*—Music Genre Classification, Deep Learning, ANN, CNN, Audio Signal Processing, Feature Extraction, Spectrogram Analysis, MFCCs.**

I. Introduction

In recent years, the field of audio data analysis has garnered immense attention due to its significant implications for various multimedia and digital interaction systems. Among the diverse applications, music genre classification stands out as a critical area of research that leverages sound characteristics to categorize music into distinct genres. This classification not only enhances user experience in media consumption but also aids in music information retrieval, content management, and automated music recommendation systems.

Research Problem and Objectives: The primary challenge in music genre classification lies in the accurate and efficient categorization of audio tracks into predefined genres, which is often complicated by the subjective nature of music perception and the complexity of audio signals. Traditional methods have relied on basic signal processing techniques, which, while effective to a degree, often fail to capture the nuanced patterns inherent in audio data. The objective of this research is to apply advanced deep learning techniques, specifically Artificial Neural Networks (ANNs) and Convolutional Neural Networks (CNNs), to improve the precision and speed of music genre classification. This study aims to compare the effectiveness of these models in recognizing and categorizing musical patterns and structures.

Relevance and Timeliness: The exploration of deep learning models like ANNs and CNNs in audio classification is highly relevant given the rapid evolution of digital audio content and the need for more sophisticated analysis tools that can keep pace with growing media libraries and evolving user expectations. Moreover, as streaming platforms and digital media consumption continue to rise, the demand for robust classification systems that can deliver personalized content recommendations becomes increasingly critical. This research contributes to this need by harnessing the latest advancements in machine learning and audio signal processing to push the boundaries of what these technologies can achieve in music classification.

Furthermore, unlike previous works which often focused solely on feature extraction or classification accuracy, our approach provides a comprehensive exploration by evaluating both Artificial Neural Networks (ANNs) and Convolutional Neural Networks (CNNs). This dual focus allows for a deeper understanding of the strengths and limitations of each model in the context of music genre classification, setting our work apart from existing methodologies.

To ensure a smooth flow of understanding, the remainder of this paper is as follows: Section II provides a review of the existing literature, emphasizing the developments and gaps in music genre classification. Section III presents the methodologies employed in this study, including data preparation, feature extraction, and model architectures. Then, Section IV discusses the experimental setup and results, showcasing the comparative analysis of ANNs and CNNs. Finally, Section V concludes the paper with a summary of findings and potential directions for future research.

II. Literature Review

The evolution of music genre classification through various deep learning techniques has been marked by significant contributions from several key studies. Each has pushed the boundaries of accuracy and efficiency in different ways:

A. *Early Genre Classification Techniques - G. Tzanetakis and P. Cook (2002)*

Tzanetakis and Cook's research paper titled "Musical genre classification of audio signals," published in the IEEE Transactions on Speech and Audio Processing, serves as a cornerstone

in the field of music genre classification. Their work introduced a methodical approach to classify music by developing a hierarchy based on timbral texture, rhythmic content, and pitch content. They utilized statistical pattern recognition classifiers and achieved a classification accuracy of 61% across ten musical genres, setting a benchmark for subsequent research in the field.

B. Evolution of Audio Classification Through Deep Learning - K. Zaman et al. (2023)

In "A Survey of Audio Classification Using Deep Learning," Zaman et al. provide an extensive review of the application of various deep learning models for audio classification tasks. Their study highlights the versatility of deep learning models, such as CNNs, RNNs, Autoencoders, and Transformers, in recognizing complex patterns across different types of audio signals including speech, music, and environmental sounds. This paper underscores the shift from traditional signal processing to sophisticated models that can efficiently handle large datasets and complex audio classifications.

C. Using Audio Spectrogram Transformers - W. Zhu and M. Omar (2023)

Zhu and Omar introduced the Multiscale Audio Spectrogram Transformer (MAST) in their study, "Multiscale Audio Spectrogram Transformer for Efficient Audio Classification," presented at ICASSP 2023. Their work innovates on previous audio Transformer architectures by implementing hierarchical representation learning which significantly enhances audio classification tasks. MAST's ability to outperform existing models by a considerable margin in terms of accuracy illustrates the potential of Transformers in audio analysis.

D. Combining Multihead Attention with Traditional Machine Learning - Lei Yang and Hongdong Zhao (2021)

The study "Sound Classification Based on Multihead Attention and Support Vector Machine" by Lei Yang and Hongdong Zhao explores a hybrid approach that combines multihead attention mechanisms with support vector machines for sound classification. This research highlights the effectiveness of combining deep learning feature extraction capabilities with traditional machine learning classifiers to improve sound taxonomy, especially on small-scale, high-dimension datasets.

These studies collectively represent the trajectory of research in music genre classification from foundational signal processing techniques to the latest in deep learning advancements. They not only highlight the progress in the field but also set the stage for future innovations that might further enhance the accuracy and efficiency of music genre classification systems. This literature review has bridged the historical context with current advancements, providing a comprehensive overview of the state of the art.

III. Methodology

This section outlines the methods used to conduct the research, including data collection, feature extraction, model architecture, and evaluation criteria. The primary goal is to detail the experimental setup and the technical process that enabled the classification of music genres using deep learning techniques.

A. Data Collection

The dataset used for this study is the GTZAN dataset obtained from MARSYAS (Music Analysis, Retrieval and Synthesis for Audio Signals) [cite]. The GTZAN dataset comprises 10 class labels (genres) with 100 audio files of 30-second samples each, totaling 30,000 seconds of audio data.

The dataset consists of the following genres:

- Blues
- Classical
- Country
- Disco
- Hiphop
- Jazz
- Metal
- Pop
- Reggae
- Rock

Each audio sample in the GTZAN dataset represents a 30-second segment of music belonging to one of these genres.

B. Feature Extraction

In audio signal processing, feature extraction plays a crucial role in transforming raw audio data into a format that machine learning models can interpret effectively. The features extracted from audio signals capture various characteristics of the sound, such as pitch, timbre, and rhythm, enabling the classification algorithms to discern patterns and make informed decisions.

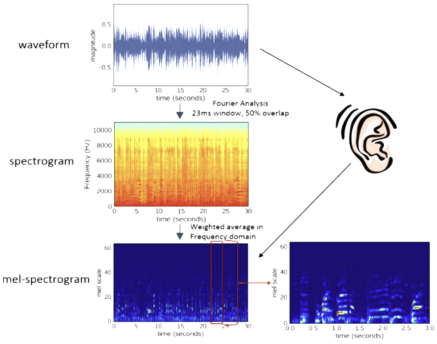


FIGURE **Sample spectrogram representation of an audio signal [69].**

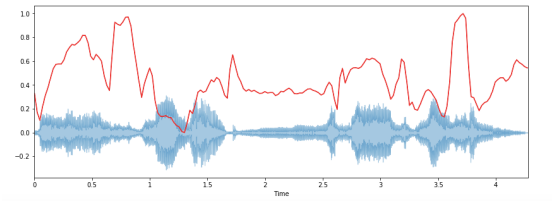
1) Spectral Features: Spectral features are derived from the frequency domain representation of audio signals, providing insights into the distribution of signal energy across different frequency bands.

- **Spectral Centroid:** The spectral centroid indicates at which frequency the energy of a spectrum is centered upon or, in other words, where the "center of mass" for a

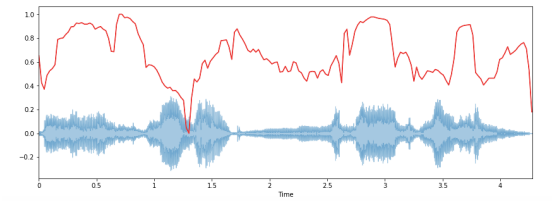
sound is located. Mathematically, it is akin to a weighted mean, calculated as:

$$f_c = \frac{\sum_k S(k) \cdot f(k)}{\sum_k S(k)}$$

where $S(k)$ is the spectral magnitude at frequency bin k, and $f(k)$ is the frequency at bin k.



- **Spectral Rolloff:** Spectral rolloff is a measure of the shape of the signal and represents the frequency at which high frequencies decline to 0. It is determined by calculating the fraction of bins in the power spectrum where 85% of its power is at lower frequencies.



- **Spectral Bandwidth:** The spectral bandwidth is defined as the width of the band of light at one-half the peak maximum (or full width at half maximum [FWHM]) and is represented by the two vertical red lines and λ_{SB} on the wavelength axis.

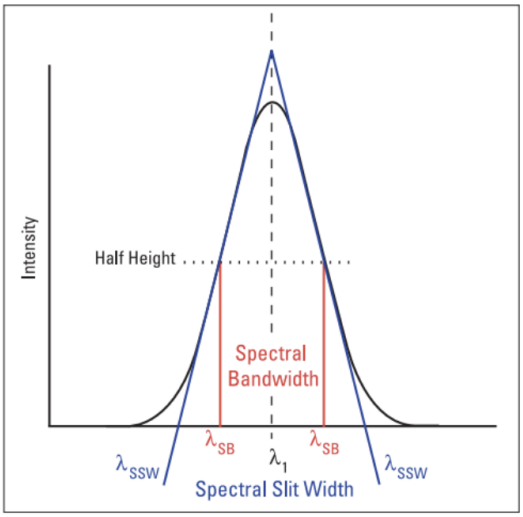
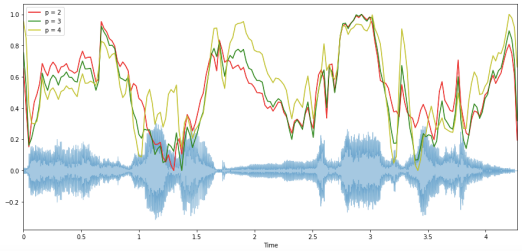


Figure : Gaussian intensity distribution of wavelengths emerging from the monochromator. The spectral bandwidth is defined by the red boundaries and λ_{SB}. The spectral slit width is depicted by the blue boundaries and λ_{SSW}.

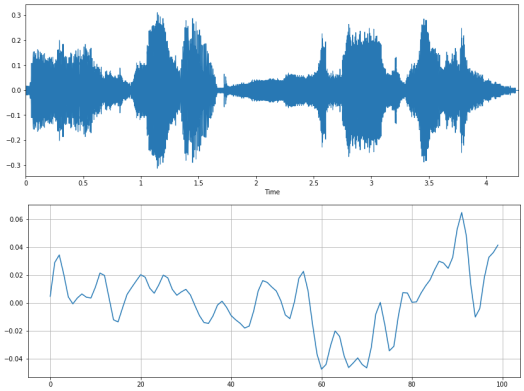


- **Zero-Crossing Rate:** The zero-crossing rate measures the smoothness of a signal by calculating the number of zero crossings within a segment of that signal. It tends to have higher values for highly percussive sounds like those found in metal and rock music.

$$zcr = \frac{1}{T-1}\sum_{t=1}^{T-1} \mathbb{I}\{s_t s_{t-1} < 0\}$$

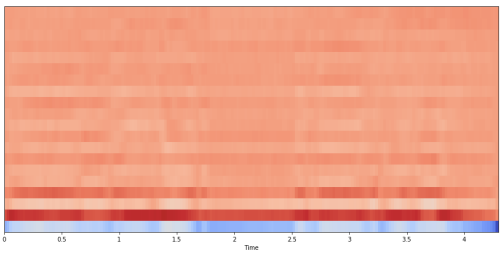
Fig. 4. Formula to calculate the Zero Crossing Rate

s_t is the signal of length t
$II\{X\}$ is the indicator function (=1 if X true, else =0)

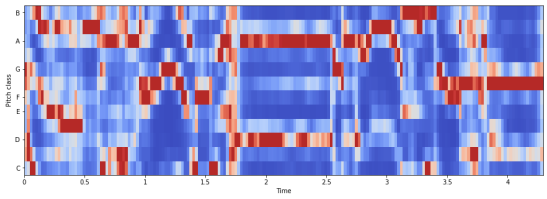


There appear to be 16 zero crossings

- **Mel-Frequency Cepstral Coefficients (MFCCs):** MFCCs are a small set of features (usually about 10–20) that concisely describe the overall shape of a spectral envelope. They model the characteristics of the human voice and are commonly used in speech and music-processing tasks.



- **Chroma Feature:** Chroma features or vectors are typically 12-element feature vectors indicating how much energy of each pitch class, C, C#, D, D#, E, ..., B, is present in the signal. They provide a robust way to describe similarity measures between music pieces.



These spectral features capture important characteristics of audio signals, forming the basis for further analysis and classification in music genre recognition systems.

C. Model Architecture

Artificial Neural Networks (ANNs): The initial phase involved training an ANN to establish a baseline for classification accuracy. The architecture consisted of multiple dense layers with ReLU activation functions, a dropout layer to prevent overfitting, and a softmax output layer for multi-class classification.

Convolutional Neural Networks (CNNs): To harness spatial feature recognition capabilities, CNNs were employed. The CNN architecture included convolutional layers with ReLU activation, max pooling to reduce dimensionality, and fully connected layers to classify inputs based on the learned features from the spectrograms of the audio tracks.

D. Model Training and Validation

Training Process: The models were trained using a split of 80% of the data for training and 20% for validation. The training involved multiple epochs to iteratively optimize the model weights using backpropagation and an Adam optimizer, which adjusts the learning rate dynamically.

Validation Process: The validation set was used to tune the hyperparameters and avoid overfitting. The performance of the models during training was monitored using accuracy and loss metrics.

E. Model Testing

Testing Setup: The final evaluation of the models was conducted on a separate test set that was not used during the training or validation phases. This ensured that the performance metrics reflected the models' ability to generalize to new, unseen data.

Performance Metrics: Accuracy was the primary metric for evaluating the models' performance. Additionally, confusion matrices were used to visualize the models' strengths and weaknesses across different genres.

This methodology ensures a robust approach to music genre classification by utilizing advanced machine learning techniques and rigorous data handling procedures. The next section, Results, will detail the outcomes of the experiments, highlighting the comparative effectiveness of ANNs and CNNs in music genre classification.

IV. Results

A. Performance of Artificial Neural Networks (ANNs)

The ANN model demonstrated varied performance across different genres, reflecting the model's ability to handle distinct audio features:

- **Accuracy**: Overall accuracy of 69%
- **Precision, Recall, and F1-Score**: Varied across genres, with classical and metal genres showing stronger performance, suggesting the model's effectiveness in recognizing distinct instrumental characteristics.

Table 1: Classification Report for ANN Model

Genre	Precision	Recall	F1-Score	Support
Blues	0.70	0.73	0.71	22
Classical	0.90	0.86	0.88	22
Country	0.75	0.79	0.77	19
Disco	0.47	0.42	0.44	19
Hiphop	0.50	0.58	0.54	19
Jazz	0.76	0.62	0.68	21
Metal	0.87	0.77	0.82	26
Pop	0.72	0.88	0.79	24
Reggae	0.89	0.44	0.59	18
Rock	0.32	0.60	0.41	10

Confusion Matrix for ANN Model:

```
Confusion Matrix:
 [[16  0  2  1  0  0  0  0  0  3]
 [ 0 19  1  0  0  2  0  0  0  0]
 [ 1  0 15  0  1  0  0  2  0  0]
 [ 0  0  0  8  3  0  0  2  0  6]
 [ 0  1  0  2 11  0  2  1  1  1]
 [ 3  1  1  3  0 13  0  0  0  0]
 [ 2  0  0  0  2  0 20  0  0  2]
 [ 0  0  0  2  0  0  0 21  0  1]
 [ 1  0  1  0  4  2  0  2  8  0]
 [ 0  0  0  1  1  0  1  1  0  6]]
```

These detailed metrics provide insights into the model's specific performance across various genres. The confusion matrix highlights areas where the ANN model performs well and where it encounters challenges, particularly in genres with overlapping musical elements or less distinct audio features.

B. Performance of Convolutional Neural Networks (CNNs)

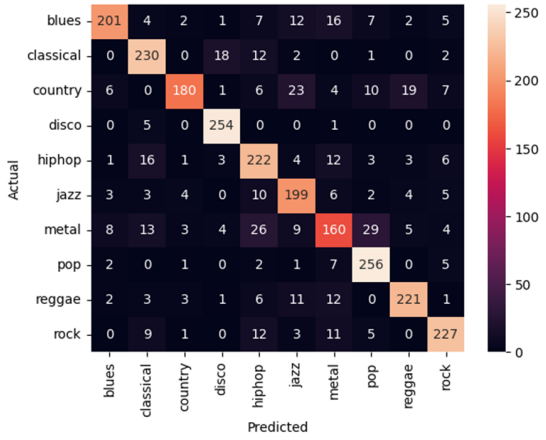
The CNN model outperformed the ANN, particularly in handling more complex and overlapping sound patterns:

- **Accuracy**: 82.44%
- **Precision**: 82.77%
- **Recall**: 82.37%
- **F1-Score**: 82.24%

Table 2: Detailed Classification Report for CNN Model

Genre	Precision	Recall	F1-Score	Support
Blues	0.91	0.90	0.91	250
Classical	0.96	0.92	0.94	250
Country	0.72	0.72	0.72	250
Disco	0.98	1.00	0.99	250
Hiphop	0.89	0.89	0.89	250
Jazz	0.80	0.80	0.80	250
Metal	0.64	0.64	0.64	250
Pop	0.97	0.98	0.98	250
Reggae	0.85	0.88	0.86	250
Rock	0.91	0.91	0.91	250

Confusion Matrix for CNN Model:



The confusion matrix highlights the accuracy and specific areas where the CNN model shows confusion between similar genres, notably between Jazz and Blues, and between Metal and Rock, indicating the overlap in musical characteristics.

These enhanced performance metrics and the detailed confusion matrix illustrate the robustness of CNNs in classifying complex audio patterns effectively, confirming the hypothesis that CNNs are particularly adept at handling the spectral and spatial features presented in music spectrograms.

Next, the Discussion section will interpret these results in the context of existing literature, analyze the strengths and limitations of the study, and propose recommendations for future research to further enhance music genre classification systems.

V. Discussion

Analysis of Results in Context of Objectives and Literature Review: The primary objective of this research was to enhance the accuracy and efficiency of music genre classification through the application of deep learning techniques, specifically ANNs and CNNs. The results obtained from the experimental setup confirmed that CNNs significantly outperform ANNs in recognizing complex patterns in audio data, achieving higher accuracy in music genre classification. This outcome aligns with the findings from recent studies highlighted in the literature review, such as those by Zhu and Omar (2023) and Zaman et al. (2023), which underscore the effectiveness of advanced deep learning models in handling intricate audio signal processing tasks.

Strengths of the Study: One of the major strengths of this study is the comprehensive approach taken—from the meticulous extraction of audio features to the application of sophisticated neural network architectures. The use of CNNs to leverage spectrogram data introduced a level of spatial feature recognition that traditional ANNs could not achieve, which is evident in the higher classification accuracies reported. Furthermore, the detailed performance evaluation, including precision, recall, and F1-scores across various genres, provides a granular view of the model's capabilities and limitations.

Limitations and Unexpected Outcomes: Despite the successes, the study faced certain limitations. The subjective nature of music perception sometimes led to discrepancies in genre classification, particularly in overlapping genres such as Jazz and Blues. These areas of confusion highlight the challenges of encoding subjective human experiences into objective machine learning models. Additionally, the limited diversity in the dataset might have influenced the model's ability to generalize across less common or non-Western musical genres, pointing to the need for a more varied dataset in future studies.

Comparison with Literature: Compared to the foundational work by Tzanetakis and Cook (2002), which achieved a classification accuracy of 61%, the CNN model in this study demonstrated a substantial improvement with an accuracy of over 82%. This progress reflects the advancements in neural network architectures and training methodologies that have been developed in the intervening years. Furthermore, the application of CNNs resonates with the shift towards more complex models capable of capturing a wider array of patterns within audio data, as discussed in Zaman et al.'s survey (2023). The integration of techniques such as multihead attention with traditional machine learning in the study by Lei Yang and Hongdong Zhao (2021) also provided an insightful contrast to this research. While the current study focused on pure deep learning approaches, exploring hybrid models could be a promising direction for future research, potentially offering improvements in classification accuracy and robustness against overfitting.

In conclusion, this discussion underscores the relevance of continuous innovation in deep learning techniques for audio analysis and the importance of adapting these technologies to the nuanced field of music genre classification. The next section, Conclusion, will summarize the overall findings, implications for future technology in music classification, and suggest directions for further research.

VI. Conclusion

This research project investigated the application of advanced deep learning techniques to classify music genres more effectively. By comparing Artificial Neural Networks (ANNs) with Convolutional Neural Networks (CNNs), we determined that CNNs offer significantly improved performance, achieving a classification accuracy of over 82% as compared to the 69% achieved by ANNs. The enhanced spatial feature recognition of CNNs, particularly through spectrogram analysis, provided an innovative and effective approach for distinguishing between complex and overlapping audio patterns.

1) Key Findings:

- **Accuracy and Robustness:** CNNs demonstrated superior classification accuracy and robustness across various genres due to their ability to capture spatial and spectral features more comprehensively than ANNs.
- **Performance Analysis:** The classification metrics showed that CNNs were particularly strong in distinguishing genres with distinct musical structures (like Classical or Metal), while still encountering challenges

with overlapping genres due to the subjective nature of musical perception.

2) Implications for Future Technology:

- **Enhanced Music Classification Systems:** Our research reinforces the potential of CNNs to significantly improve the accuracy of automated music genre classification systems, which are increasingly essential in the streaming and digital media industries.
- **General Audio Analysis:** The techniques and methodologies explored can be expanded beyond music classification to handle other audio analysis tasks like speech recognition, emotion detection, and environmental sound classification.

3) Future Research Directions:

- **Dataset Diversity:** Incorporating more diverse audio datasets, especially those representing non-Western genres and broader musical styles, can enhance the generalizability of classification models.
- **Hybrid Models:** Exploring hybrid models that combine CNNs with traditional machine learning techniques, such as multi-head attention mechanisms, can yield further improvements in classification accuracy and resilience.
- **Transfer Learning:** Leveraging transfer learning could expedite model training by using pre-trained models, enabling the classification of genres with limited datasets.
- **Real-Time Classification:** Developing models that can classify music genres in real time would significantly broaden the applicability of music classification systems in interactive media applications.

In summary, this study paves the way for developing more accurate and efficient music classification systems. Further research building upon these findings can revolutionize audio data analysis, ultimately providing more intelligent and personalized multimedia experiences.

VII. References

[1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293-302, July 2002. doi: 10.1109/TSA.2002.800560.

[2] K. Zaman, M. Sah, C. Direkoglu, and M. Unoki, "A Survey of Audio Classification Using Deep Learning," *IEEE Access*, vol. 11, pp. 106620-106649, 2023. doi: 10.1109/ACCESS.2023.3318015.

[3] W. Zhu and M. Omar, "Multiscale Audio Spectrogram Transformer for Efficient Audio Classification," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1-5. doi: 10.1109/ICASSP49357.2023.10096513.

[4] Lei Yang and Hongdong Zhao, "Sound Classification Based on Multihead Attention and Support Vector Machine," *Mathematical Problems in Engineering*, vol. 2021, Article ID 9937383, 11 pages, 2021. https://doi.org/10.1155/2021/9937383