Residual Analysis for Music Genre Classification

Salamun Nuhin and Yutong Xia
Boston College
{nuhin, xiapq}@bc.edu

***Abstract*— Classifying music genres using computational methods has gained considerable attention, driven by the need to organize and analyze the ever-growing volume of digital music. Central to this research is the GTZAN dataset, introduced by George Tzanetakis and Perry Cook in 2002 [1]. Despite its widespread use, the dataset's inherent issues, such as mislabeling, artist overlap, and audio distortions, have posed significant challenges for achieving reliable results [4].**

This study explores the use of ResNet, a deep convolutional neural network known for its residual learning framework, to tackle these challenges. Unlike traditional classification methods, ResNet's architecture allows for effective feature extraction by addressing gradient vanishing issues and enabling deeper network designs [7]. By applying transfer learning with ImageNet-pretrained weights, the model benefits from generalized features learned on a larger dataset, accelerating convergence and enhancing classification accuracy.

The preprocessing pipeline involved resampling audio files to sound waves, generating Mel Spectrograms, and normalizing these spectrograms for consistency. To further improve the model's robustness, SpecAugment [5] was employed for data augmentation, introducing variations in the training data without altering its core characteristics.

Through extensive experiments, the ResNet model achieved an accuracy of 83.8% on the GTZAN dataset. These results highlight the effectiveness of combining advanced preprocessing, transfer learning, and data augmentation for music genre classification. This approach offers a robust framework for future research, addressing dataset limitations while leveraging modern deep learning techniques to improve classification performance.

I. INTRODUCTION

The classification of music genres has long been a focus of research in music information retrieval, aiming to sort audio tracks into predefined categories based on intrinsic audio features. The original research is based upon classical machine learning techniques. This was from the work of George Tzanetakis and Perry Cook in 2002, where they employed a K-Nearest Neighbor (k-NN) classifier to analyze audio features such as beats per minute and Mel Spectrogram attributes [1]. Their model achieved an accuracy of 61%, a result they argued was comparable to human classification abilities based on trials with college students. This study led to the creation of the widely used GTZAN dataset which consists of 1,000 audio tracks, each 30 seconds long, distributed equally across ten genres, providing a widely adopted reference for comparative studies. Despite the widespread use, the dataset has significant flaws. As highlighted by Bob L. Sturm in 2013 [4], the dataset suffers from artist overlap, where certain genres are dominated by tracks from the same artists, limiting its ability to generalize across broader music styles. Also, audio distortions and other inconsistencies make it even more unreliable.

In this work, the ResNet model, a Convolutional Neural Network (CNN) known for its efficiency and residual learning framework, was utilized. Introduced by He et al. in 2015, ResNet-18 employs skip connections to facilitate deeper networks by mitigating vanishing gradient issues [7]. Its balance between depth and computational efficiency makes it particularly effective for medium-sized datasets like GTZAN. The model processes Mel Spectrograms, a time-frequency representation of audio, to extract temporal and spectral features, enabling robust genre classification.

Transfer learning played a crucial role in this study, leveraging pretrained ResNet weights from ImageNet to accelerate training and improve accuracy. The preprocessing pipeline included resampling audio to 16 kHz, generating Mel Spectrograms, and applying normalization. Data augmentation techniques, including SpecAugment [5], were incorporated to enhance the model's generalization capabilities. Experiments systematically explored various hyperparameter settings, augmentation methods, and evaluation strategies to optimize performance.

Although progress in music genre classification has been substantial, challenges persist, such as the subjective nature of genre definitions and the lack of universally representative datasets. Larger datasets often remain unlabeled, limiting their applicability for supervised learning. Despite these limitations, significant milestones have been achieved, such as the 90% accuracy on the Magnatune dataset (5,500 tracks across nine genres) reported in 2019 [4]. On GTZAN, the highest accuracy of 83.9% was achieved using a masked modeling approach, as listed on the Paperswithcode leaderboard [2].

This study seeks to refine the ResNet-18 architecture for music genre classification, addressing the GTZAN dataset's limitations and exploring techniques to improve the model's robustness and generalization. By building on the strengths of residual learning and systematic experimentation, this research aims to contribute to the development of more accurate and reliable genre classification systems.

II. Data Preprocessing and Augmentation

This section will discuss how the data was processed prior to the training of the models.

A. Performance Benchmarks

Table I highlights the performance of state-of-the-art models on the GTZAN dataset based on the Paperswithcode leaderboard.

Rank	Model	Year	Accuracy (%)
1	Masked Modelling Duo ratio = 0.7	2022	83.9
2	Masked Modelling Duo ratio = 0.6	2022	83.3
3	Jukebox	2021	79.7
4	CLMR	2021	68.6

TABLE I

Paperswithcode GTZAN Leaderboard [?]

The highest reported accuracy of 83.9% was achieved in 2022 by a masked modeling approach utilizing a duo ratio of 0.7, closely followed by a slightly modified version with a duo ratio of 0.6 achieving 83.3% accuracy. These models leverage advanced masking strategies to improve feature extraction and generalization.

Jukebox, a model primarily designed for music generation, achieved an accuracy of 79.7% in 2021, demonstrating its ability to model high-level musical features. Contrastive Learning of Musical Representations (CLMR), another 2021 approach, achieved 68.6% accuracy, emphasizing its utility in learning representations from unlabeled data.

While these models reflect significant advancements, the gap between their accuracy and Sturm's estimated maximum highlights ongoing challenges such as artist overlap, labeling errors, and dataset distortions. These results underscore the need for further research to refine methodologies and address these limitations. Future work could explore novel architectures, enhanced data preprocessing, and augmentation strategies to close the gap and advance the field of music genre classification.

III. Methods

3.0 Data List – The dataset was obtained from Kaggle [3] and was divided into training, validation, and testing sets with the following distribution: 800 files for training, 100 files for validation, and 99 files for testing. The split ensured that all subsets maintained a balanced distribution of the ten genres present in the GTZAN dataset.

3.1 Data Preprocessing – The GTZAN dataset was preprocessed to optimize its use for classification tasks. The preprocessing involved: 1. Conversion of audio files to a uniform sample rate of 16 kHz to standardize the dataset and align with audio analysis tools. 2. No division of the original 30-second audio files was performed; instead, the full tracks were preserved for spectrogram generation to retain all temporal data. 3. Audio files were saved with a bit depth of 16 bits per sample to maintain consistency across the dataset.

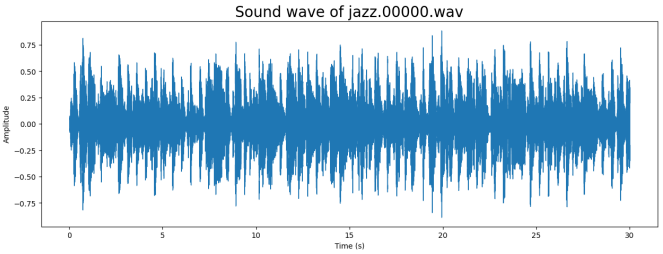

Fig. 1. Figure of the sound waves for the audio clip of a jazz sample file

3.2 Mel Spectrogram Generation – Spectrograms were generated using the Mel Scale, which maps frequencies to a scale approximating human auditory perception. The spectrograms were computed using the Fast Fourier Transform (FFT) algorithm with the following parameters:

- Sampling rate: Maximum amplitude for each audio file
- FFT length/number of bins: 2048
- Hop Length: 512

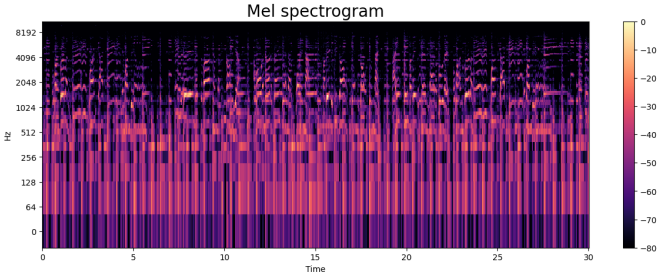

Fig. 2. Figure of a sample mel-spectogram

Each spectrogram was then normalized to have a mean of 0 and a standard deviation of 1. Normalization ensured consistency across the dataset, reducing the influence of amplitude variations and enhancing the generalization capability of the model.

3.3 Data Augmentation: Several experiments incorporated data augmentation techniques to improve the model's robustness. The following methods were applied:

1) **Spectral Augmentation:** Spectral Augmentation, specifically SpecAugment [5], was utilized to introduce variations in the input data by masking random frequency bands and time intervals in the Mel Spectrogram[5]. This technique involves two key operations:
   - **Frequency Masking:** A random range of consecutive Mel frequency bins (up to a maximum width of F bins) is set to zero. This simulates missing frequency information and encourages the model to generalize by learning features from the unmasked regions. In this study, up to 48 frequency bins out of 128 were masked for each spectrogram[5].

- **Time Masking:** A random range of consecutive time frames (up to a maximum width of T frames) is set to zero. This simulates temporal noise or dropout, encouraging the model to focus on robust temporal patterns. Here, up to 192 time frames out of 1024 were masked[5].

These augmentations were applied dynamically during training to the best-performing ResNet-18 model. By ensuring varied inputs in each epoch, the approach enhanced the model's ability to handle real-world variations in audio data while improving its generalization capability.

2) **Normalization:** Normalization was applied to the Mel Spectrograms to standardize their values. The mean was adjusted to 0, and the standard deviation to 1, ensuring consistent scaling of features. This assumes that each dataset follows a non-Gaussian distribution, improving model training stability and performance.

3.4 Evaluation Protocol – The preprocessed and augmented dataset was used for training a ResNet-18 model, with evaluation based on single-class classification accuracy. The model predictions were validated using the 100 validation samples, while the final accuracy was reported on the 99 test samples.

3.5 Summary – The preprocessing pipeline ensured the dataset was normalized and standardized, with balanced splits for training, validation, and testing. The generated Mel Spectrograms provided a robust representation of the audio signals, and the augmentation techniques further strengthened the model's ability to generalize.

IV. Experiments

4.0 Experiment Overview – Multiple experiments were conducted to evaluate the performance of ResNet-based models on the GTZAN dataset. These experiments included freezing layers, comparing ResNet-18 and ResNet-50 models, applying Squeeze and Excitation blocks, and incorporating Spec Augmentation techniques. The dataset split consisted of 800 training samples, 100 validation samples, and 99 testing samples.

4.1 Baseline Models – The baseline model used was ResNet-18 with all layers frozen except the last two layers. The model was trained for 200 epochs using the Adam optimizer with a learning rate of 0.0001, achieving an accuracy of 80.8%.

Model	Description	Epochs	Accuracy (%)
ResNet-18	All except last two	200	80.8
ResNet-18	Without any modifications	20	76.76

TABLE II

Baseline Model Results

4.2 ResNet-18 vs. ResNet-50 – To compare the performance of deeper ResNet models, ResNet-50 was evaluated alongside ResNet-18. ResNet-50, despite its higher capacity, demonstrated overfitting on the dataset and achieved lower accuracy than ResNet-18.

Model	Description	Epochs	Accuracy (%)
ResNet-18	18 layers	20	81.8
ResNet-50	50 layers	20	79.8

TABLE III

Comparison of ResNet-18 and ResNet-50

4.3 Layer Freezing – The effect of freezing layers on model performance was analyzed. For ResNet-18, freezing the first 8 layers yielded the highest accuracy of 81.8%. ResNet-50 achieved its best accuracy of 79.8% when 8 layers were frozen, though still lower than ResNet-18 due to overfitting.

Model	Frozen Layers	Accuracy (%)
ResNet-18	6 layers frozen	78.8
ResNet-18	8 layers frozen	81.8
ResNet-50	10 layers frozen	73.7
ResNet-50	8 layers frozen	79.8

TABLE IV

Layer Freezing Results

4.4 Squeeze and Excitation Blocks:

Squeeze and Excitation (SE) blocks are a neural network architectural enhancement introduced to improve the model's ability to selectively emphasize informative features while suppressing less relevant ones. Proposed by Hu et al. [3] in the SENet architecture, these blocks refine channel-wise feature maps by explicitly modeling the interdependencies between channels. SE blocks operate in two main stages:

- **Squeeze:** Global spatial information from each feature map is aggregated into a single value using global average pooling. This operation captures the global context of each channel, essentially summarizing its importance relative to the input data.
- **Excitation:** A fully connected layer processes the squeezed outputs to learn a set of channel-wise weights. These weights are applied via multiplication to scale the original feature maps, enhancing important channels and diminishing irrelevant ones.

In this study, SE blocks were integrated into the ResNet-50 architecture to improve channel-wise feature selection, particularly for the Mel Spectrogram inputs. By focusing on the most critical features in the spectrograms, the SE blocks aimed to reduce redundancy and improve the model's generalization to diverse audio patterns.

Despite the added complexity, this modification achieved an accuracy of 76.8% after 20 epochs. While the accuracy was lower than that of ResNet-18, this result highlights the potential of SE blocks to enhance feature representation, suggesting that further optimization of their integration could yield improved performance in future work.

4.5 Spec Augmentation – Spec Augmentation was applied to the best-performing ResNet-18 model (8 layers frozen) to simulate real-world audio variations. This technique involved frequency masking, time masking, and time warping, which improved the model's robustness and achieved a final accuracy of 83.8%.

Experiment	Model/Configuration	Epochs	Accuracy (%)
SE	ResNet-50 SE blocks	20	76.8
Spec Aug	ResNet-18 (8 layers frozen)	20	83.8

TABLE V

SQUEEZE AND EXCITATION VS. SPEC AUGMENTATION RESULTS

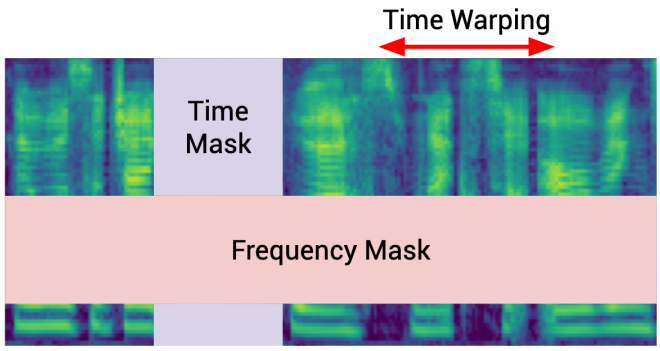

Fig. 3. Simple Visual for How Spec Augmentation Works

4.6 Summary of Experiments – The experiments demonstrated that ResNet-18 consistently outperformed ResNet-50 on the GTZAN dataset due to the latter's tendency to overfit on small datasets. Spec Augmentation further enhanced ResNet-18's performance, achieving the highest accuracy of 83.8%, with stable validation loss.

Epoch	Train Loss	Train Accuracy (%)	Val Loss	Val Accuracy (%)	Time (s)
1	0.002	100.0	0.582	82.0	29.46
2	0.001	100.0	0.596	82.0	29.55
3	0.001	100.0	0.593	82.0	29.34
4	0.001	100.0	0.568	82.0	29.45
5	0.000	100.0	0.569	81.0	29.43
6	0.000	100.0	0.579	81.0	29.33
7	0.000	100.0	0.610	84.0	29.47
8	0.000	100.0	0.601	80.0	29.29
9	0.000	100.0	0.617	83.0	29.50
10	0.000	100.0	0.589	82.0	29.44
11	0.000	100.0	0.598	83.0	29.41
12	0.000	100.0	0.579	82.0	29.50
13	0.000	100.0	0.624	83.0	29.41
14	0.000	100.0	0.594	82.0	29.36
15	0.000	100.0	0.600	84.0	29.38
16	0.000	100.0	0.586	82.0	29.56
17	0.000	100.0	0.599	80.0	29.33
18	0.000	100.0	0.592	81.0	29.43
19	0.000	100.0	0.589	83.0	29.42

TABLE VI

TRAINING AND VALIDATION RESULTS ACROSS EPOCHS FOR RESNET-18 WITH 8 LAYERS FROZEN UNDER SPECAUGMENTATION

V. CONCLUSION

This study demonstrated the application of ResNet-based architectures for music genre classification on the GTZAN dataset, with a focus on understanding the impact of architectural choices, layer freezing, and augmentation techniques on model performance. The experiments revealed that while deeper models like ResNet-50 possess greater representational capacity, they are more prone to overfitting when applied to smaller datasets such as GTZAN. Overfitting occurs when a model learns to memorize the training data rather than generalize from it, resulting in poor performance on unseen data. This was evident in the performance of ResNet-50, which, despite its increased depth, achieved lower accuracy compared to ResNet-18 due to its inability to generalize effectively.

The primary reason for overfitting in this context is the limited size and diversity of the GTZAN dataset. With only 1,000 audio tracks split across ten genres, the dataset provides insufficient variety to fully leverage the capacity of a deeper model like ResNet-50. Additionally, issues such as artist overlap and mislabeled files exacerbate the risk of overfitting, as the model may learn spurious patterns specific to these dataset artifacts rather than generalizable features of the genres.

To mitigate overfitting and improve model performance in future work, the following strategies could be explored:

- **Data Augmentation:** Employ more advanced augmentation techniques, such as time stretching, pitch shifting, and Spec Augmentation, to artificially expand the diversity of the dataset and reduce the model's reliance on specific patterns in the training data.
- **Transfer Learning:** Leverage larger, more diverse datasets like Audioset for pretraining before fine-tuning on GTZAN. Pretrained models can generalize better by starting with features learned from a broader distribution of audio data.
- **Regularization Techniques:** Introduce regularization methods such as dropout, weight decay, or label smoothing to prevent the model from overfitting to the training data.
- **Dataset Expansion:** Augment the GTZAN dataset by incorporating additional labeled tracks from other publicly available datasets, ensuring balanced genre representation and increased data variety.
- **Architecture Modifications:** Incorporate components like Squeeze and Excitation blocks or attention mechanisms to focus the model's learning on relevant features while reducing overfitting to irrelevant details.

In conclusion, while ResNet-18 proved to be more effective than ResNet-50 for this task, future improvements in data quality, augmentation, and model architecture hold the potential to further enhance performance. Addressing the inherent limitations of the GTZAN dataset and adopting robust generalization techniques will be critical for advancing music genre classification models.

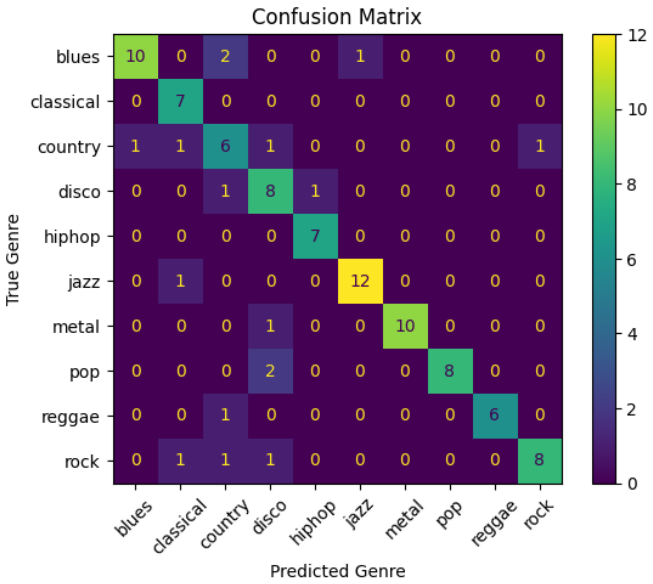

Fig. 4. Confusion Matrix for the Predicted Genres for the Best Model

REFERENCES

[1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," Speech and Audio Processing, IEEE Transactions on, vol. 10, no. 5, pp. 293–302, Jul 2002. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1021072.

[2] Papers with Code GTZAN leaderboard. Retrieved from: https://paperswithcode.com/sota/music-genre-classification-on-gtzan.

[3] He, Kaiming, et al. "Mask R-CNN." *arXiv preprint arXiv:1709.01507*, 2017, https://arxiv.org/abs/1709.01507.

[4] Sturm, Bob L. "The GTZAN Dataset: Its Contents, Its Faults, Their Effects on Evaluation, and Its Future Use." *arXiv preprint arXiv:1306.1461*, 2013, Lists posted on https://github.com/coreyker/dnn-mgr.

[5] Park, D.S., et al. "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition." *Proceedings of Interspeech 2019*, 2019, pp. 2613-2617, doi:10.21437/Interspeech.2019-2680.

[6] Tzanetakis, G., and Cook, P. "GTZAN." July 2002. Retrieved March 24, 2024, from https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification.

[7] He, Kaiming, et al. "Deep Residual Learning for Image Recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778, https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf.