Music Genre Classification Project Report

Wenxiao Bu
Department of Electrical and Computer Engineering
Western University
London, Canada
wbu2@uwo.ca

Jiaxin Yang
Department of Electrical and Computer Engineering
Western University
London, Canada
jyan682@uwo.ca

Yushi Gan
Department of Electrical and Computer Engineering
Western University
London, Canada
ygan29@uwo.ca

David Wei
Department of Electrical and Computer Engineering
Western University
London, Canada
xwei85@uwo.ca

***Abstract*—Music genre classification is a fundamental task in music information retrieval that enables personalized recommendations and enhances user experience in streaming platforms. This paper presents a novel CNN-GRU-Attention hybrid architecture with attention mechanism for automatic music genre classification using Mel-spectrogram features. Our model combines convolutional neural networks for spatial feature extraction with gated recurrent units for temporal sequence modeling, enhanced by an attention mechanism that focuses on the most discriminative temporal segments. We evaluate our approach on the Free Music Archive Small dataset, comprising 8,000 30-second audio clips across 8 balanced genres. Through systematic hyperparameter optimization and class-specified data augmentation, we demonstrate the effectiveness of combining spatial and temporal modeling for music genre classification. The proposed hybrid model achieves 62.71% validation accuracy and 63.65% test accuracy, outperforming other baseline architectures. The results show promising performance for real-world music recommendation systems and highlight the benefits of hybrid deep learning architectures in audio classification tasks.**

***Index Terms*—Music genre classification, Convolutional neural networks, Gated Recurrent Unit, Deep Learning**

I. Introduction

Music has become an integral part of modern digital life, with millions of new tracks released daily across various streaming platforms. As music libraries continue to expand exponentially, traditional manual classification methods are no longer feasible given the scale of modern music collections, necessitating the development of automated solutions for music genre classification.

Music genre classification represents a fundamental problem in Music Information Retrieval (MIR), serving as a cornerstone for numerous applications including music recommendation systems, playlist generation, and content organization. Meanwhile, recent advances in deep learning have demonstrated remarkable success in pattern recognition tasks across various domains, especially Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). These architectures have shown exceptional capability in extracting hierarchical or sequential features from complex data structures, making them particularly well-suited for audio signal processing and music analysis tasks.

This project aims to develop and evaluate a deep learning-based music genre classification model using CNN-based architectures. The remainder of this paper is organized as follows:

Section 2 reviews related work in music genre classification with deep learning approaches.

Section 3 describes the dataset and processing methodology.

Section 4 compares the model architecture and presents the training procedures and primary results.

Section 5 provides the systematic hyperparameter optimization strategy, along with class-specific enhancement techniques to improve overall model robustness.

Section 6 evaluates and discusses the experimental results.

Finally, Section 7 concludes with a summary of findings and directions for future work.

II. Related Work

In recent years, music genre classification has been an active area of machine learning [1]. Bahuleyan [2] conducted a comprehensive comparison of machine learning techniques for music genre classification, evaluating Support Vector Machines (SVM), Neural Networks (NN), and CNN on the Audio Set dataset. The study revealed that CNN-based approaches, particularly VGG-16, achieved superior performance with 64% test accuracy, highlighting the effectiveness of deep learning methods for audio classification tasks.

Building upon CNN architectures, researchers have explored hybrid models combining convolutional and recurrent networks. Ashraf et al. [3] investigated hybrid architectures combining CNN and RNN using the GTZAN dataset, where CNN and Bi-GRU using Mel-spectrogram achieved the best accuracy at 89.30%. Similarly, Gessle et al. [4] analyzed the performance of CNN and Long Short-Term Memory (LSTM) and indicated that CNN is superior in short segment audio recognition.

The choice of dataset substantially affects training results and model performance. The Free Music Archive (FMA) dataset, introduced by Defferrard et al. [5], provides 917 GiB and 343 days of Creative Commons-licensed audio from 106,574 tracks arranged in a hierarchical taxonomy of 161 genres, and has become a crucial benchmark for large-scale music genre classification research. Several studies have explored the effectiveness of CNN architectures on this dataset. Gorlongaiatto et al. [6] and Zhang et al. [7] proposed a deep convolutional neural network (CNN) model for music genre classification on the FMA dataset, demonstrating the potential of CNNs in extracting hierarchical features from audio spectrograms.

While significant progress has been made in applying deep learning to music genre classification, CNN-GRU hybrid models represent great potential for further improvement. This project aims to analyze and implement an optimized CNN-GRU-Attention hybrid model that effectively combines the spatial feature extraction capabilities of CNNs with the temporal modeling strengths of GRUs, specifically designed and tuned for the FMA Small dataset.

III. Data Processing

A. Dataset Overview

The Free Music Archive (FMA) dataset provides a comprehensive collection of Creative Commons-licensed audio tracks suitable for music information retrieval tasks. The dataset contains 106,574 tracks from 16,341 artists across 14,854 albums, totaling 917 GiB of audio data representing 343 days of music content.

For this study, we utilized the FMA Small subset, which contains 8,000 balanced tracks (approximately 1,000 per genre) from the 8 most popular top-level genres: Electronic, Experimental, Folk, Hip-Hop, Instrumental, International, Pop, and Rock. The specifications of the audio files are provided in Table I,

B. Feature Extraction

We employed the librosa library for audio feature extraction, investigating three different feature representations to determine their effectiveness for genre classification. Specifically, we extracted and compared the following audio features:

- Mel-frequency Cepstral Coefficients (MFCC): Captures spectral characteristics crucial for audio classification
- Mel-frequency Chroma Features (MFCG): Represents harmonic and melodic content
- Mel spectrogram (Mel-Spec): Provides time-frequency representation preserving temporal dynamics

Figure 1 shows a visualization of Mel-Spec. Table II summarizes the analyses of each feature type.

Based on validation accuracy comparison in Table II, mel-spectrogram demonstrated superior performance over compressed feature representations. MFCC feature also achieves high validation accuracy, but exhibits a noticeably reduced train-validation gap of 2.84%, requiring vigilance against overfitting. Consequently, mel-spectrogram was selected as the optimal feature representation for subsequent model training and evaluation.

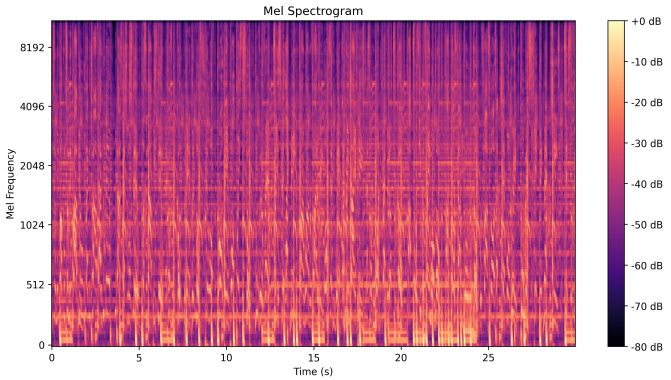

Fig. 1. Visualization of Food by AWSL, Hip-hop.

Notably, all experimental results demonstrated higher validation than training accuracy. Since we employed early stopping (patience=10) to guarantee complete training, this pattern cannot be explained by underfitting. We speculate this reflects artist and album effect, or confounding effects [8]: when different compositions by identical artists or albums are distributed across training and validation sets, models may potentially memorize creator-specific patterns rather than authentic style features for classification. We will address this effect in future studies.

C. Data Split and Augmentation

The dataset was split into training, validation, and test sets in a 7:2:1 ratio (5946: 850: 1700). Additionally, stratified sampling was employed to preserve class balance across all splits, guaranteeing equal representation of all 8 genres in each subset. This balanced distribution eliminates potential bias toward any particular genre during training.

Due to hardware limitations restricting us to the FMA Small dataset (8,000 samples), we implemented a data augmentation strategy on the training set to enhance model generalization and prevent overfitting. The augmentation applied four transformation types in equal proportions:

- Original data: Unmodified mel-spectrograms (25%)
- Noise augmentation: Gaussian noise addition (25%)
- Time shift: Temporal displacement along time axis (25%)
- Combined augmentation: Noise + time shift (25%)

Given the notably lower accuracy and recall performance of the Experimental genre, we implemented a targeted oversampling strategy specifically for this challenging category. The strategy and results will be evaluated in following sections.

IV. Model Structure

A. Baseline CNN Model

Initially, we designed a conventional four-layer CNN architecture as our baseline model. However, preliminary parameter calculations revealed that this approach would result in more than 1 million parameters, which could significantly reduce

TABLE I
AUDIO SPECIFICATIONS

Type	Format	Sampling rate	Bit rate	Channels	Duration
Standard	MP3	44,100 Hz	360 kbit/s	Stereo	30 s

TABLE II
AUDIO FEATURES COMPARISON

Feature	Dimensions	Size	Best Val Acc	Train-Val Gap
Origin	L × 1	7.42 GB	N/A	N/A
MFCC	13 coefficients	0.471 GB	56.13%	2.84%
MFCG	12 chroma bins	1.42 GB	54.37%	7.75%
Mel-Spec	128 × T[1]	4.44 GB	56.35%	5.09%

[1] T represents the number of time frames

training speed and increase GPU memory usage while also introducing substantial overfitting risks that might compromise generalization performance.

To address these limitations, we adopted depthwise separable convolutions as proposed by Howard et al. replacing standard convolution layers to create an Efficient CNN model. As shown in Figure 2, The model architecture consists of:

- One standard convolution layer (1→32 channels, 3×3 kernel)
- Two depthwise separable convolution blocks with batch normalization and ReLU activation
- Adaptive pooling and dropout regularization
- Final classification layer with softmax output

B. Hybrid Model

Beyond the Efficient CNN model, we explored hybrid architectures combining convolutional feature extraction with sequential modeling:

(a) CNN-GRU Hybrid Model. This architecture combines a CNN backbone for spatial feature extraction from mel-spectrograms with GRU layers for temporal sequence modeling and an attention mechanism for weighted feature aggregation and residual connections to prevent gradient degradation.

(b) CNN-LSTM Hybrid Model. This architecture is similar to the CNN-GRU-Attention model but utilizing LSTM units instead of GRU cells, offering enhanced memory capacity for capturing longer-term dependencies in audio sequences.

Table III presents the comparative analysis of our candidate models.

TABLE III
HYBRID MODEL PERFORMANCE COMPARISON

Model	Parameters	Training Time	Best Val Accuracy
Efficient CNN	16,456	720 s	52.59%
CNN-GRU	39,050	757 s	56.35%
CNN-LSTM	46,346	775 s	55.65%

The CNN-GRU hybrid demonstrated superior performance, achieving 56.35% validation accuracy while maintaining reasonable computational efficiency. We also noticed the incorporation of attention mechanisms [9] [10] has been proven to significantly enhance the performance of neural networks across various domains while suppressing irrelevant information, thereby improving learning efficiency.

Based on these results, we selected CNN-GRU-Attention hybrid model as our final architecture for comprehensive hyperparameter optimization and evaluation.

V. HYPERPARAMETER TUNING AND OPTIMIZATION

To maximize model performance while maintaining computational efficiency, we implemented a systematic five-phase hyperparameter optimization strategy. Our tuning strategy consists of the following phases:

1) Batch size optimization. Four batch sizes were evaluated: 16, 32, 48, and 64.
2) Learning rate optimization with Adam optimizer. Four learning rates were evaluated: 5e-5, 1e-4, 3e-4, and 1e-3.
3) Optimizer and scheduler selection. We compared Adam and AdamW with or without Scheduler.
4) Multi-parameter grid search. We conducted comprehensive grid search across beta2, CNN dropout, classifier dropout and weight_decay.
5) Class-specific enhancement for underperforming genre, i.e. Experimental genre.

All experiments were conducted using early stopping with a patience of 8 epochs and validation loss as the primary metric to prevent overfitting and reduce training time.

A. Batch Size Optimization

We first optimized the batch size [11] to establish a stable training foundation while considering GPU memory constraints. The baseline configuration used CNN dropout of 0.3, classifier dropout of 0.3. Since batch size 64 encountered GPU memory overflow issues and was excluded from further consideration, Table IV shows the comparison of other batch size.

TABLE IV
PERFORMANCES OF DIFFERENT BATCH SIZE

Batch Size	Val Acc	Training Time
16	59.88	42.3 min
32	58.35	28.0 min
48	58.82	28.2 min

While batch size 32 and 48 these configurations significantly improved training speed, they resulted in slightly inferior model performance. Batch size 16 achieved the best validation accuracy and was selected as the optimal configuration.

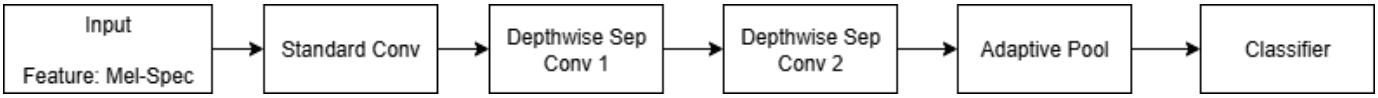

Fig. 2. Structure of Efficient CNN Model

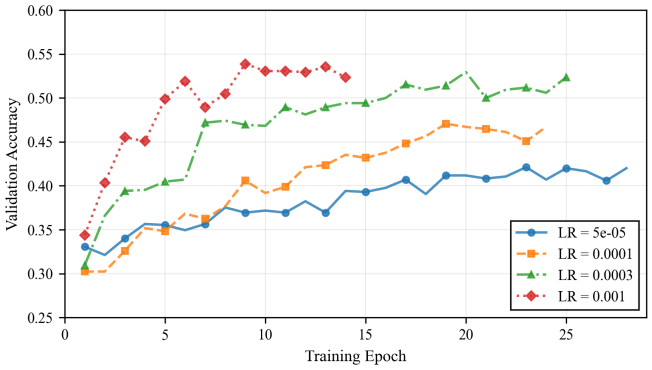

Fig. 3. Learning Rate Comparison

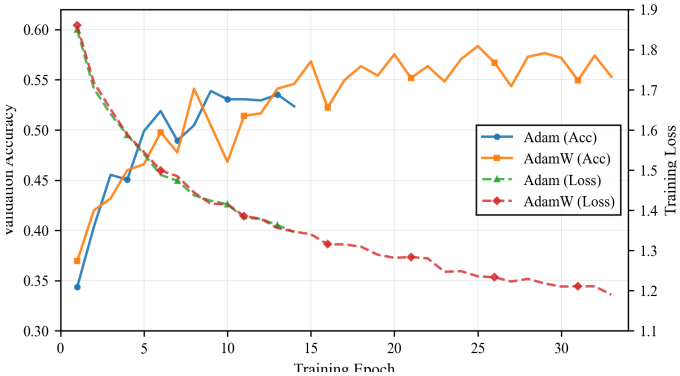

Fig. 4. Optimizer Comparison

B. Learning Rate Optimization

In the baseline model, we employed the Adam optimizer, but the initial learning rate setting still has a significant impact on model training results. Figure 3 shows the model performances under different initial learning rates.

The results demonstrate that within the tested range, higher learning rates lead to faster convergence and superior validation accuracy. Consequently, the learning rate of 1e-3 was chosen for subsequent phases.

C. Optimizer and Scheduler Selection

Building upon the optimal learning rate, we experimented with Adam and AdamW [12] optimizers to improve training stability and convergence.

As shown in Figure 4, the AdamW optimizer demonstrated superior performance with a 4.47% improvement in validation accuracy and significantly lower training loss, likely due to its improved weight decay regularization.

With AdamW optimizer, we compared training with and without the ReduceLROnPlateau scheduler. The results is listed in Table V.

TABLE V
SCHEDULER COMPARISON

Configuration	Val Acc	Train Acc	Train Loss	Val Loss
Without Scheduler	58.35%	59.20%	1.2356	1.2002
With ReduceLROnPlateau	58.35%	60.58%	1.1515	1.1787

While validation accuracy remained constant, the scheduler improved training accuracy and reduced both training loss and validation loss, indicating better optimization dynamics.

D. Multi-parameter grid search

Using AdamW optimizer with ReduceLROnPlateau scheduler, we conducted comprehensive grid search across key hyperparameters:

- **Adam beta2 parameter (β_2):** $\{0.98, 0.999\}$ [13]
- **CNN dropout rate (λ_{CNN}):** $\{0.2, 0.3, 0.4\}$
- **Classifier dropout rate (λ_{cls}):** $\{0.2, 0.3, 0.4\}$
- **Weight decay (λ_{wd}):** $\{0.01, 0.02, 0.03\}$

This resulted in 54 total combinations, requiring approximately 13 hours of training time on NVIDIA RTX 4060. Due to computational constraints, each configuration was trained for at most 30 epochs.

Figure 5 shows the Hyperparameter impact analysis, and Table VI shows the top 3 configurations from grid search.

TABLE VI
TOP 3 HYPERPARAMETER CONFIGURATIONS

Rank	Val Acc	β_2	λ_{CNN}	λ_{cls}	λ_{wd}
1	60.47%	0.999	0.2	0.4	0.03
2	60.12%	0.999	0.2	0.3	0.03
3	59.65%	0.98	0.3	0.2	0.02
Baseline	56.35%	N/A	0.3	0.3	N/A

The top configurations share the pattern of ($\beta_2 = 0.999, \lambda_{CNN} = 0.2, \lambda_{wd} = 0.03$), suggesting these three parameters form an optimal base configuration. The best-performing configuration achieved 60.47% validation accuracy, representing a 4.12% improvement over the baseline configuration.

E. Class-specific Enhancement

Analysis of the confusion matrix from our best-performing model revealed significant performance disparities across different music genres. While the model achieved satisfactory accuracy for most genres, it demonstrated notably poor performance in identifying "Experimental" music, with a recall of only 0.29.

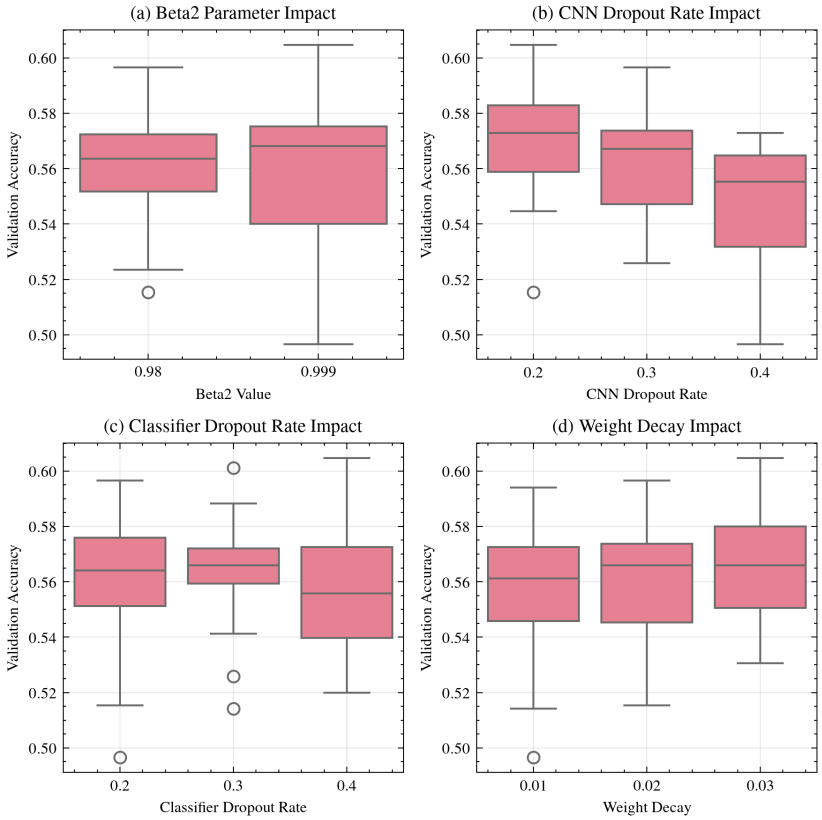

Fig. 5. Hyperparameter Impact Analysis

This performance degradation can be attributed to the inherent characteristics of experimental music. Unlike conventional genres with well-defined musical patterns, experimental music consists of ambiguous spectral features, making it challenging for automated classification systems.

To address the poor performance on the experimental genre, we implemented a class-specific enhancement strategy focused on targeted oversampling and augmented data generation. The details of this strategy is shown in Table VII, and the comparison of models is shown in Table VIII.

TABLE VII
DATA AUGMENTATION STRATEGIES COMPARISON

Augmentation Type	Standard	Strong Enhanced
Gaussian Noise (σ)	0.005	0.01 (2×)
Time Shifting	±10%	±15% (1.5×)
Spectral Masking	2-5%	5-10%
Oversampling Factor[1]	0	1
Applied to	All genres	Experimental only

[1] The factor represents the number of samples to generate additionally

The class-specific enhancement strategy demonstrates significant effectiveness in addressing the experimental genre classification challenge. With recall increased from 0.28 to 0.45 (+60.7%), directly addressing the model's primary weakness in identifying experimental music. Importantly, the overall validation accuracy improved from 61.06% to 62.71%, demonstrating that targeted enhancement of the underperforming class benefits the entire model without compromising performance on other genres.

TABLE VIII
PERFORMANCE COMPARISON FOR EXPERIMENTAL GENRE

Configuration	TF	Recall	F1-score	Overall Val Acc
Standard Enhanced	30	0.28	0.33	61.06%
Strong Enhanced	48	0.45	0.48	62.71%
Improvement	**+18**	**+0.17**	**+0.15**	**+1.65%**

VI. EVALUATION AND DISCUSSION

The final evaluation was conducted on a balanced test set comprising 1,700 music samples from the FMA Small dataset, with sample counts of [213, 213, 213, 212, 212, 212, 213, 212] for Electronic, Experimental, Folk, Hip-Hop, Instrumental, International, Pop, and Rock genres respectively.

A. Performance

Table IX and Figure 7 presents the detailed classification performance for each genre. The optimized CNN-GRU-Attention hybrid model achieved a test accuracy of 63.65% on the balanced test set with consistent performance across precision, recall, and F1-score metrics, indicating robust classification capabilities across the eight music genres.

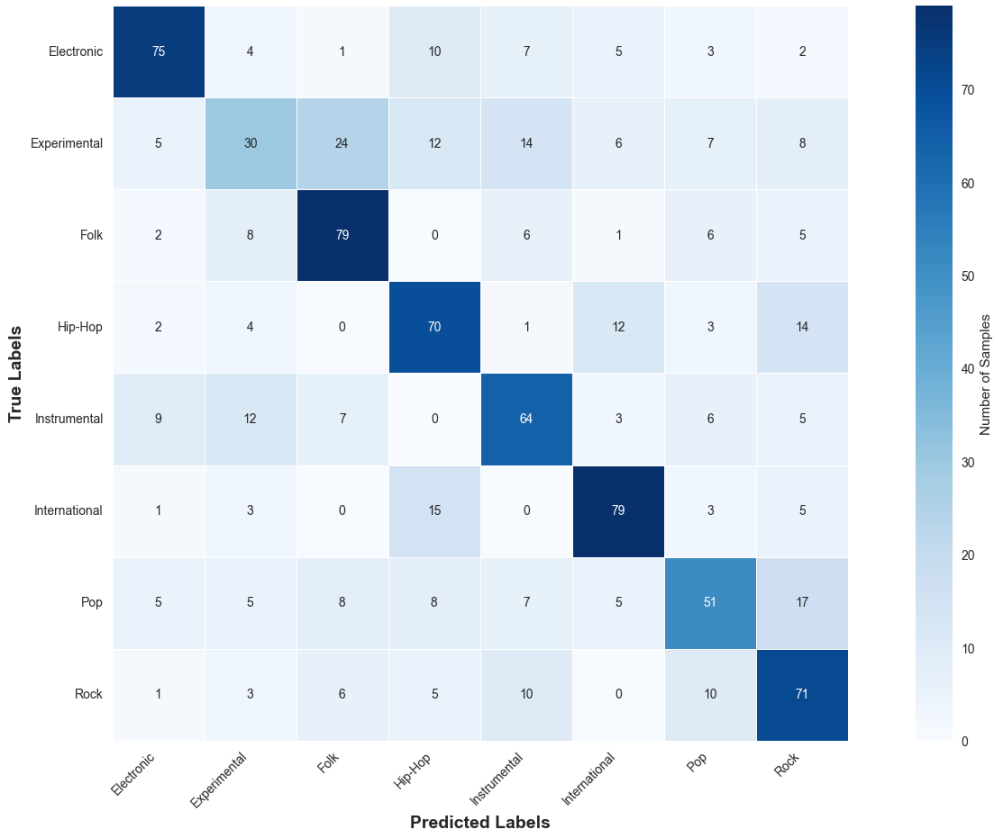

Fig. 6. Confusion Matrix of Best Configuration on Validation Dataset

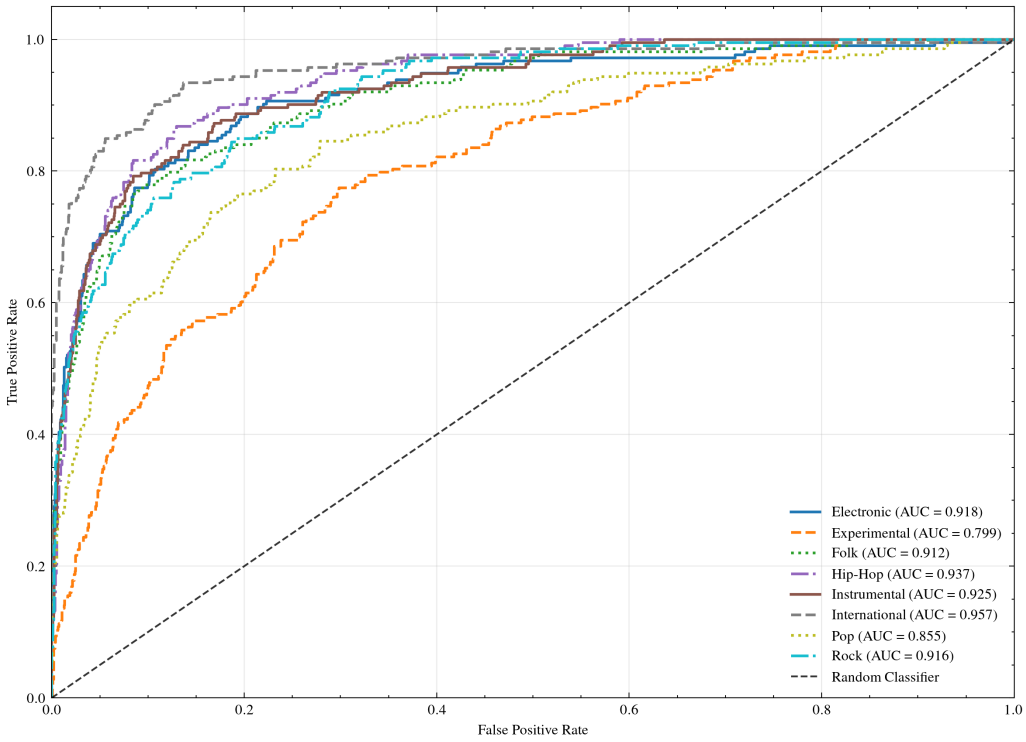

Fig. 7. ROC Curves for Final Model

TABLE IX
CLASSIFICATION PERFORMANCE BY GENRE

Genre	Precision	Recall	F1-Score	Support
Electronic	0.68	0.68	0.68	213
Experimental	0.44	0.44	0.44	213
Folk	0.63	0.65	0.64	213
Hip-Hop	0.64	0.71	0.67	212
Instrumental	0.70	0.65	0.68	212
International	0.76	0.80	0.78	212
Pop	0.57	0.56	0.56	213
Rock	0.66	0.59	0.63	212
Macro avg	0.64	0.64	0.64	1700
Weighted avg	0.64	0.64	0.64	1700

B. Discussion

The classification results reveal distinct performance patterns across different music genres, indicating varying levels of classification difficulty inherent to each category.

- International, Instrumental and Electronic demonstrate strong classification performance with high precision and recall, indicating their distinguishable spectral patterns.
- Hip-Hop, Folk and Rock show higher recall than precision, indicating the model successfully identifies most of these samples but with some false positives, possibly due to shared rhythmic patterns with other genres.
- Pop and Experimental exhibit relatively low performance, reflecting the genre's inherent diversity and tendency to incorporate elements from multiple other genres.

In summary, the model successfully captures mel-spectrogram patterns that distinguish culturally and instrumentally distinct music, and still requires specialized approaches for handling unconventional acoustic patterns.

VII. CONCLUSION

This project presents a comprehensive approach for music genre classification task using deep learning models on the FMA Small dataset. Our investigation demonstrates that depth-wise separable convolutions significantly reduce computational complexity while maintaining classification performance. The CNN-GRU hybrid outperformed both the CNN baseline and CNN-LSTM model, striking an optimal balance between model complexity and accuracy. Through hyperparameter optimization and additional data augmentation for experimental genre, we identified optimal configurations that achieved 63.65% test accuracy.

For future improvement, we have proposed two primary directions:

- Enhanced Data Augmentation Strategy. We plan to segment the original 30-second audio clips into smaller temporal fragments while designing specialized augmentation techniques for poorly performing categories, such as the Experimental and the Pop. This approach would increase training sample diversity, potentially improving model generalization for all genres.
- Advanced Architecture Exploration. Future work will investigate advanced model architectures from recent music classification literature to further enhance our model. This includes exploring deeper attention mechanisms, transformer-based approaches, and multi-scale feature fusion techniques that have shown promise in contemporary audio classification tasks.

These directions aim to address current limitations in genre-specific performance while advancing the overall effectiveness of deep learning approaches.

APPENDIX - 1 RELATED LINKS AND CONTRIBUTIONS

The dataset is available on: https://github.com/mdeff/fma.

The repository is available on: https://github.com/W-Bu-coder/music_classification.

Wenxiao Bu (wbu2@uwo.ca) handled data processing, baseline model comparison, and participated in hyperparameter optimization.

Yushi Gan (ygan29@uwo.ca) handled data preprocessing and cleaning.

Jiaxin Yang (jyan682@uwo.ca) handled dataset selection and model selection.

David Wei (xwei85@uwo.ca) handled model optimatization and hyperparameter tuning.

All members worked together on report drafting.

APPENDIX - 2 GENAI USAGE

In this project, the GenAI is only used for following goals:

- Results Visualization. Figures are generated by Claude to show our results clearly.
- Citation Generation. We used Claude to transfer different citations into IEEE format.
- Non-technical Issues. We encountered GPU out of memory issue when tuning with batch size = 48. Then we asked Claude to add garbage collection operations into our code and keep our project moving forward.

REFERENCES

[1] N. Ndou, R. Ajoodha, and A. Jadhav, "Music genre classification: A review of deep-learning and traditional machine-learning approaches," in *Proc. IEEE Int. IOT, Electronics and Mechatronics Conf. (IEMTRONICS)*, Toronto, ON, Canada, Apr. 2021, pp. 1–6.

[2] H. Bahuleyan, "Music genre classification using machine learning techniques," *arXiv preprint arXiv:1804.01149*, Apr. 2018.

[3] M. Ashraf, F. Abid, I. U. Din, J. Rasheed, M. Yesiltepe, S. F. Yeo, and M. T. Ersoy, "A hybrid CNN and RNN variant model for music classification," *Appl. Sci.*, vol. 13, no. 3, p. 1476, Jan. 2023.

[4] G. Gessle and S. Åkesson, "A comparative analysis of CNN and LSTM for music genre classification," M.S. thesis, Dept. Computer Science, Linköping Univ., Linköping, Sweden, 2019.

[5] C. Gorlongaiatto *et al.*, "Music genre classification with convolutional recurrent neural networks: An analysis on the FMA dataset," Neural Networks and Deep Learning Project, 2023. [Online]. Available: https://github.com/carlosgorlongaiatto/NNDL_Project

[6] Y. Zhang and T. Li, "Music genre classification with parallel convolutional neural networks and capuchin search algorithm," *Sci. Rep.*, vol. 15, p. 9580, 2025.

[7] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A dataset for music analysis," *arXiv preprint arXiv:1612.01840*, Dec. 2016.

[8] F. Rodríguez-Algarra, B. L. Sturm, and S. Dixon, "Characterising confounding effects in music classification experiments through interventions," *Trans. Int. Soc. Music Information Retrieval*, vol. 2, no. 1, pp. 52–66, 2019.

[9] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *arXiv*:1709.01507 [cs.CV], Sep. 2017.

[10] W. Xu, Y. Wan, and D. Zhao, "SFA: Efficient attention mechanism for superior CNN performance," *Neural Process. Lett.*, vol. 57, no. 38, pp. 1–21, 2025.
[11] S. Raschka, "No, we don't have to choose batch sizes as powers of 2," Sebastian Raschka's Blog, Jul. 2022. [Online]. Available: https://sebastianraschka.com/blog/2022/batch-size-2.html
[12] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, Nov. 2017.
[13] S. Gugger and J. Howard, "AdamW and Super-convergence is now the fastest way to train neural nets," *fast.ai*, Jul. 2018. [Online]. Available: https://www.fast.ai/posts/2018-07-02-adam-weight-decay.html