SC4001 - Neural Networks and Deep Learning

Group Project Report

Project Idea: Music Genre Classification through Multi-modal fusion

Name	Matriculation Number
Park Yumin	U2123691G
Park Junseo	N2402401D
Lee Sanghyun	U2020666G

Table of Contents

1. Introduction

Music genre classification is a foundational task in Music Information Retrieval(MIR) which enables automatic organization, recommendation, and indexing of vast digital music libraries. Traditional approaches have largely relied on acoustic features extracted from the raw audio files using tools like librosa. While effective, we hypothesize that these features alone cannot fully capture the complex dynamics of multi-faceted musical expression. This research is driven by the question: Can integrating multiple representations of audio data lead to a more robust and accurate genre classifier?

We propose a multi-modal feature fusion framework to address this limitation. Different audio representations like aggregated librosa features and spectrogram images capture complementary information. The spectrogram, when treated as an image, excels at visually representing time-frequency energy distribution, making it ideal for processing with a Convolutional Neural Network (CNN) to capture timbral and temporal patterns [1]. Conversely, statistical acoustic features from librosa summarize various psychoacoustic properties, suitable for dedicated processing architectures.

The core of our approach is to systematically compare the effectiveness of two fusion strategies in combining the two musical feature modalities for genre classification. We process distinct input modalities, extracted from CNN and Multilayer Perceptron (MLP), and apply two different fusion methods, namely feature-level and logit-level fusion to integrate the extracted outputs. We anticipate that two strategies of multi-modal integration will improve the classification accuracy over single-modality approaches, and provide empirical evidence for the necessity of holistic audio data representation in MIR.

2. Dataset Selection and Preprocessing

2.1. Dataset Selection

The Free Music Archive (FMA) dataset, developed by researchers at EPFL and released in conjunction with the International Society for Music Information Retrieval Conference (ISMIR) [2], is chosen as the dataset for this work. FMA data consists of 30 seconds of 106,574 raw audio MP3 files and 161 unbalanced genres. In addition to the audio files, the dataset includes comprehensive metadata of artist, album, and genre labels, along with pre-extracted acoustic features that allows us to conduct systematic evaluation of multi-modal and fusion approaches to music genre classification.

2.2. Data Preprocessing

To ensure a fair and focused evaluation of genre classification strategies, we preprocessed a series of steps adjusting to the hierarchical structure of our dataset. Each track in the dataset is labeled with one or more genres, organized in a two-level hierarchy consisting of 16 parent (root) genres and 147 child genres. Since our study aims at single-label classification models, we restricted the dataset to tracks assigned to single genre, resulting in 26,723 tracks. This filtering of eliminating the label ambiguity enables more interpretable and standard evaluations of model performance.

Next, to reduce class imbalance, we mapped all child genres to their corresponding root genres based on the hierarchical relationship in metadata, specifically genre.csv. For example, tracks labeled 'Punk' (child genre) were reassigned to the 'Rock' root genre. Not only reassigning the genres, but also we limited our final dataset to tracks within top 5 genres (Rock, Electronic, Hip-Hop, Folk, Pop) which comprised approximately 73.6% of the entire tracks, reducing skewed class distributions and focusing on genres with

sufficient representation. After all filtering and genre mapping, our working set comprised 18,238 uniquely labeled audio tracks.

2.3. Data Splitting Method

The entire dataset (18,238 tracks) was partitioned using a 40:40:20 stratified split, creating three distinct, non overlapping subsets. These were designated as a Baseline Training Set (40%), a Fusion Training Set (40%), a Test Set (20%). Stratification ensured the genre distribution was preserved across all three sets.. This step was critical in ensuring that there is no data leakage in the two stages of our report: 1) MLP&CNN hyperparameter tuning and training & 2) Fusion strategy evaluation using designated validation and test sets. This rigorous preprocessing pipeline establishes a robust foundation for developing and validating the multi-modal feature fusion framework for music genre classification.

3. Model Selection and Implementation

3.1. Baseline Model Architecture Selection

Spectrograms transform audio signals into 2D images that encode time-frequency information, making them well-suited for CNN architectures which excel at spatial pattern recognition. Among CNN architectures, the EfficientNet-B0 is selected by its proven ability to efficiently extract complex hierarchical features from visual representations of audio data [3]. EfficientNet-B0 balances network depth, width, and resolution through compound scaling, achieving high accuracy with lower computational cost and fewer parameters compared to larger CNN variants [4].

For the acoustic features extracted via librosa, a Multi-Layer Perceptron (MLP) is employed owing to its flexibility and effectiveness in modeling structured statistical feature vectors. Librosa provides a rich set of audio descriptors such as MFCCs, chroma features, and spectral contrast, which summarize psychoacoustic properties in fixed-length vectors. MLPs are well-suited to learn complex nonlinear relationships in such tabular data and have been widely used in audio classification tasks where feature extraction precedes modeling [5]. By using an MLP for these statistical features, the model can capture complementary information to the spectrogram-based CNN while maintaining computational simplicity.

This two-branch of neural approach leverages the strengths of fusing both spatial and feature-based representations to improve overall genre classification performance.

3.2. Baseline Model Hyperparameter Optimization

Before the baseline models can be utilized as stepping stones for feature fusion, they must first be trained effectively. In order to get to feature fusion, we first optimize the respective models' performance on a proxy task of 5-class music genre classification.

This process is guided by the core assumption that a model exhibiting high accuracy in classification is also proficient at learning and extracting the most discriminative features from its respective input data [6]. To identify the optimal set of hyperparameters for both the CNN and MLP, we employed a 3-fold cross validation grid search methodology. The objective metric for this search was classification accuracy.

The following sections detail the specific hyperparameter space and the optimization results of the two baseline models.

3.2.1. CNN EfficientNet-B0

The first baseline model is an EfficientNet-B0, configured to process spectrogram inputs of the music audio data. This model was pre-trained on the ImageNet dataset with fixed core architectural hyperparameters (depth, width, resolution). The batch size is fixed to 64 during grid search to balance stable gradient estimation and computational efficiency.

The grid search for the CNN focused on two key training parameters:

- Dropout Rate: We evaluated dropout rate of [0.3, 0.4, 0.5] to identify the optimal overfitting prevention value for our CNN architecture.
- Learning Rate: We tested learning rates [0.0001, 0.0005, 0.001] to determine best optimisation speed.

The 3-fold cross validation is visualized in Figure 1 which indicates all 9 sets of models in the search grid. The learning rate of 0.001 with 0.5 dropout rate is selected as the optimal hyperparameter for CNN by achieving the highest average cross validation accuracy of 0.6781, as summarized in Table 1.

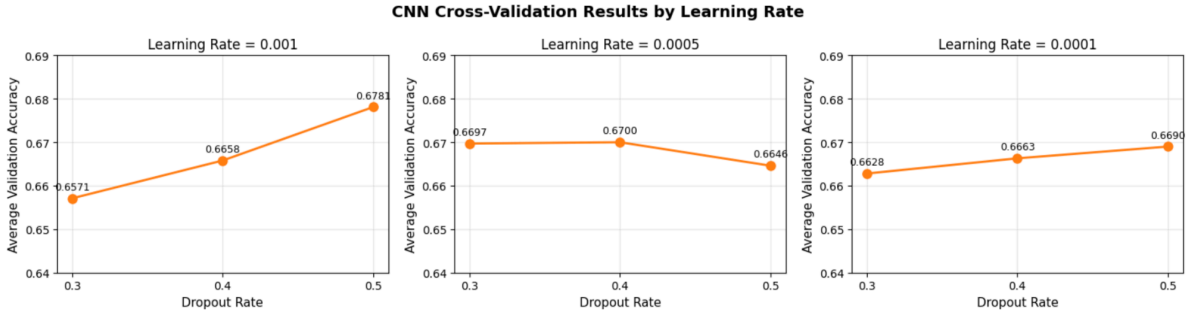

Figure 1. Average CV Accuracy Plot against Learning Rate

Learning Rate	Dropout Rate	Average Accuracy	Standard Deviation
0.001	0.5	0.6781	0.0048

Table 1. Optimal Hyperparameter Combination for CNN

To complete the results for the CNN baseline, we additionally report the final test performance after hyperparameter optimization and cross-validation. The test accuracy and test loss of best CNN configuration is shown in Table 2.

Test Loss	Test Accuracy
1.0938	68.46%

Table 2. Final Test Results for CNN

3.2.2. MLP

The second baseline model is a Multi-Layer Perceptron(MLP) designed to process the Librosa aggregated features. Given the structured, vector-based nature of the data, the MLP's architecture is a critical factor in its performance. Similarly we maintain the batch size of 64 to ensure the consistency with CNN and set the dropout rate to 0.4 to balance the overloading grid search and its computational feasibility.

The grid search for the MLP focused on both architectural and training parameters:

- Network Width: We explored different widths such as [64, 128, 256] to find the optimal network capacity.
- Network Depth: We additionally explored different depths such as [2, 3, 4]. The combinatory grid-search helped to identify the optimal network architecture suitable for the task at hand.
- Learning Rate: Similar to CNN, a range of learning rates [0.0001, 0.0005, 0.001] were tested.

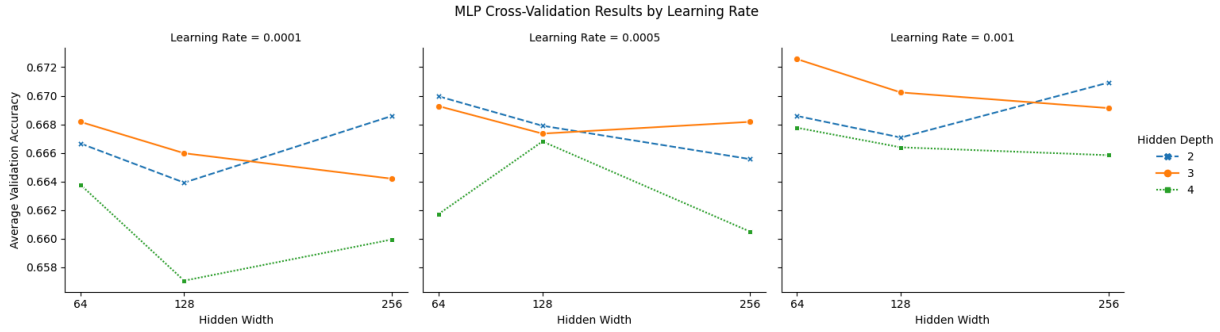

Figure 2. Average CV Accuracy Plot against Learning Rate

Altogether, the grid search has trained and evaluated 27 sets of models, and Figure 2 captures the 3-fold cross validation result. The optimal configuration, identified through the 3-fold cross validation is summarized in Table 3.

Network Width	Network Depth	Learning Rate	Average Accuracy	Standard Deviation
64	3	0.0010	0.6726	0.0067

Table 3. Optimal Hyperparameter Combination for MLP

Similarly, for the MLP baseline, we report the final test accuracy and loss based on the optimal hyperparameter identified through cross-validation. The results, presented in Table 4, provide a benchmark for comparing the effectiveness of genre classification.

Test Loss	Test Accuracy
0.8601	68.54%

Table 4. Final Test Results for MLP

4. Fusion Strategies

To leverage the complementary strengths of each modality, we implement and compare two fusion approaches: Logit-level and Feature-level fusion

4.1 Logit-Level Fusion

In neural networks, logits are the raw, unnormalized prediction scores produced by the final layer of the model before any activation function is applied [7]. Building on this, logit-level fusion is a late-stage integration strategy where the logits from independently trained models, such as a CNN on spectrogram images and an MLP on engineered audio features, are combined through concatenation or similar aggregation before final decision making. This approach enables the model to leverage complementary, modality-specific knowledge from each branch while preserving their distinct learning processes [8]. Since the outputs of both the CNN and MLP in our study are treated as modality-specific logits, we anticipate that logit-level fusion will enhance genre classification accuracy by maximizing the unique strengths and representations captured by each network prior to decision aggregation.

To evaluate logit-level fusion, we designed a multi-stage pipeline leveraging pre-trained unimodal networks and fusion classifiers. As we split the dataset, half of the train dataset is utilized to train the logit-level fusion and the final test is conducted with the test dataset. Both the CNN (EfficientNet-B0, trained on spectrogram images) and MLP (trained on engineered librosa features) are pretrained and their weights are frozen. Logits are extracted from both models for every track. The “LogitFusionModel” is constructed to concatenate these logits (each modality producing 5 logits for the five genres) into a joint representation, which feeds into a secondary MLP classifier, and again determined through a 3-fold cross validation adopting the grid-search strategy. The batch size for training and evaluation was also fixed at 64 to ensure stable gradients and computational efficiency.

The grid search for the logit-level fusion conducted to identify optimal value for two parameters:

- Hidden Width: [16, 32, 64] is trained as candidate values to find the optimal network capacity.
- Hidden Depth: Depths of [1, 2, 3] are explored to figure optimal versatility for logit-level fusion.

A total of 9 sets of models are trained and results are displayed in Figure 3.

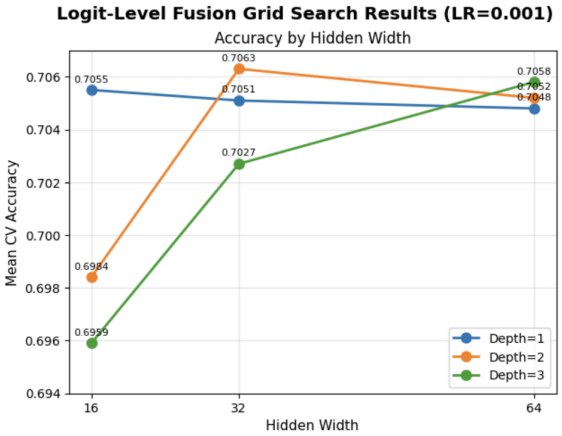

Figure 3. Average CV Accuracy Plot against Hidden Width

Hidden Width	Hidden Depth	Average Accuracy	Standard Deviation
32	2	0.7063	0.017

Table 5. Optimal Hyperparameter Combination for Logit-Level Fusion

After identifying the best fusion classifier hyperparameters as summarized in Table 5, final results are evaluated using the unseen test dataset. The test loss and accuracy in Table 6 demonstrates that notable improvement over single-modality models. The increase of test accuracy indicates that combining prediction at logit-level effectively captures complementary information from different feature modalities.

Test Loss	Test Accuracy
0.8287	70.10%

Table 6. Final Test Results for Logit_Level Fusion

4.2 Feature-Level Fusion

Feature-level fusion, also known as intermediate fusion [9], integrates representations from multiple modalities at the feature extraction stage, typically by concatenating or merging intermediate outputs from separate neural network branches before classification [10]. In our approach, feature-level fusion combines the feature vectors learned respectively by a CNN from spectrogram images and an MLP from librosa features; these vectors are fused and then passed to a joint MLP classifier for genre prediction. This method is designed to exploit the complementary information embedded in intermediate layers, aiming to capture richer cross-modal relationships than logit-level fusion.

Since the unbalanced dimensional feature occurs in biased fusion when it is concatenated and fed into the final classifier, the MLP outputs are linearly projected to match the CNN feature dimensionality of 1280. These projected MLP features and CNN features are then concatenated into a single fused feature vector of 2560 dimension. This fused vector is passed to a joint MLP classifier, whose architecture is controlled by the identical fusion hyperparameters with logit-level fusion:

- Hidden Width: Due to increased fused feature dimension, the width values also increased to [128, 256, 512] to ensure the balanced fusion.
- Hidden Depth: Same value of depths as [1, 2, 3] are implemented.

Our grid search evaluated a total of 9 combinations of these hyperparameters at the learning rate of 0.001. During all training and evaluations steps, all feature extraction models are set to evaluation mode with weights frozen to ensure only the fusion classifier is updated throughout training and testing. The result of 3-fold cross-validation is shown in Figure 4 and the optimal hyperparameter combination is summarized in Table 7.

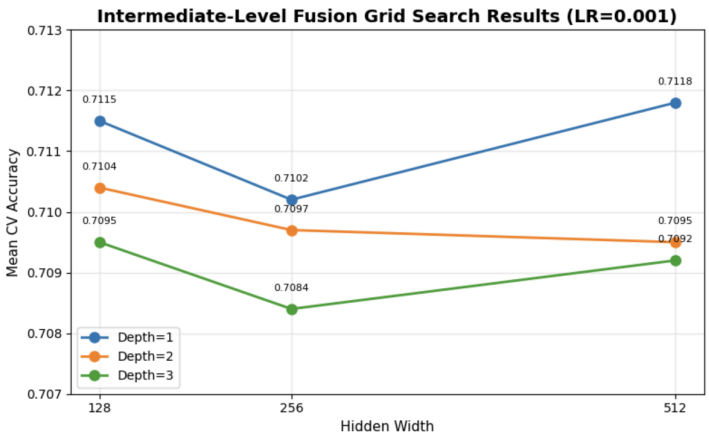

Figure 4. Average CV Accuracy Plot against Hidden Width

Hidden Width	Hidden Depth	Average Accuracy	Standard Deviation
512	1	0.7118	0.018

Table 7. Optimal Hyperparameter Combination for Feature-Level Fusion

As shown in Table 8, feature-level fusion achieved a test accuracy of 70.46% with a test loss of 0.8864, overperforming both the single-modality baselines and the logit-level fusion. This result clearly demonstrates that feature-level fusion better leverages the complementary information at the representation stage than aggregating the predictions at the decision level. The observed performance highlights that integrating features prior to classification provides more discriminative power for music genre prediction than relying solely on single-modal network or logit-level fusion.

Test Loss	Test Accuracy
0.8864	70.46%

Table 8. Final Test Results for Feature-Level Fusion

5. Results

The comparative evaluation of our music genre classification models are summarized in Table 9 and Figure 5

Model	Test Accuracy
CNN	68.46%
MLP	68.54%
Logit-Level Fusion	70.10%

Feature-Level Fusion	70.46%

Table 9. Accuracy Comparison among 4 models

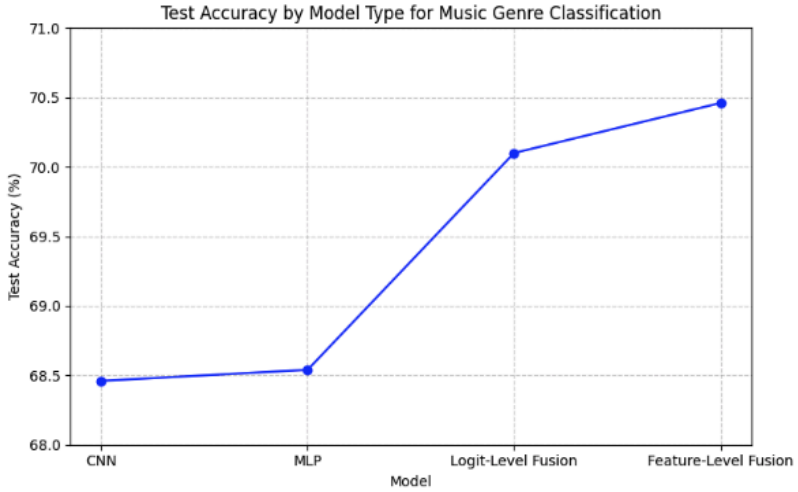

Figure 5. Line Chart of Test Accuracy by Model Type

The result clearly demonstrates the benefit of multi-modal fusion strategies over single-modality baselines. Both single models of CNN (68.46%) and MLP (68.54%) achieve relatively high test accuracy, reflecting the discriminative power of both spectral and audio features individually. However, integrating the two modalities with logit-level fusion leads to clear performance improvement to 70.10%. Feature-level fusion even further improves and scores 70.46%, the highest test accuracy among the models.

Our model accuracy is quite high compared to other models. In a related work [11] on deep neural networks for music genre classifiers, that used single modality data, achieved the maximum accuracy of 56.39%. Although their model is trained to predict 3 extra genres, the relatively high accuracy of our model and fusion strategy shows that it is comparable to other models that solely rely on a single mode of input to classify genre.

6. Conclusion

In this project, we explored music genre classification using a fusion technique that aims at capturing the complementary strengths of audio representations across modalities. As such, we integrated a Multi Layered Perceptron for aggregated statistical features and a Convolutional Neural Network for spectrogram spatial features. Our experiments on the Free Music Archive dataset revealed that this deeper integration of modalities enhances the discriminative capacity of the model, leading to improved music genre classification accuracy as backed by the improvement of classification accuracy from around 68% in single-modalities classifiers to above 70%. However, a few challenges remain, such as dealing with a wider range of genres or even further, classification among sub-genres which tend to display similar characteristics. Addressing these granular classification challenges through enhanced feature engineering and more complex fusion architectures represents the next step for this research.

7. References

[1] Zheng, W., Mo, Z., Xing, X., & Zhao, G. (2018). CNNs-based Acoustic Scene Classification using Multi-Spectrogram Fusion and Label Expansions. arXiv preprint arXiv:1809.01543. https://arxiv.org/pdf/1809.01543

[2] Defferrard, M., Benzi, K., Vandergheynst, P., & Bresson, X. (2017). FMA: A Dataset For Music Analysis. arXiv preprint arXiv:1612.01840. https://arxiv.org/abs/1612.01840

[3] Grasso, C., Carvalho, T., Amaral, J. F., Coelho, P., Oliveira, R., & Olivera, G. (2025). Optimizing Musical Genre Classification Using Genetic Algorithms. In *Proceedings of the 27th International Conference on Enterprise Information Systems (ICEIS 2025)* (Vol. 1, pp. 881-887). SCITEPRESS. https://www.scitepress.org/Papers/2025/134182/134182.pdf

[4] Morfi, V., Mavromatis, N., & Tsoumakas, G. (2021). Music genre classification using deep neural networks and data augmentation. *Artificial Intelligence*, *293*, Article 103456. https://doi.org/10.1016/j.artint.2020.103456

[5] Suhail, M. S. K., Kumar, J. G. V., Varma, U. M., Vege, H. K., & Kuchibhotla, S. (2020). MLP model for emotion recognition using acoustic features. *International Journal of Emerging Trends in Engineering Research*, *8*(5), 1702–1708. https://www.warse.org/IJETER/static/pdf/file/ijeter34852020.pdf

[6] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

[7] Moon Technologies Labs. (n.d.). What is logits in machine learning? A quick guide. Retrieved November 11, 2025, from https://www.moontechnolabs.com/qanda/logits-machine-learning/

[8] Karystinaios, E., Hentschel, J., Neuwirth, M., & Widmer, G. (2025). AnalysisGNN: Unified music analysis with graph neural networks. arXiv preprint arXiv:2509.06654. https://arxiv.org/pdf/2509.06654.pdf

[9] Li, Y., El Habib Daho, M., Conze, P.-H., Zeghlache, R., Le Boité, H., Tadayoni, R., Cochener, B., Lamard, M., & Quellec, G. (2024). A review of deep learning-based information fusion techniques for multimodal medical image classification. arXiv preprint arXiv:2404.15022. https://arxiv.org/html/2404.15022v1

[10] Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2016). Multimodal fusion for multimedia analysis: A survey. *Information Fusion*, *27*, 156–175. https://doi.org/10.1016/j.inffus.2015.06.002

[11] Kostrzewa, D., Kaminski, P., & Brzeski, R. (2021). Music genre classification: looking for the Perfect Network. *International Conference on Computational Science (ICCS 2021) Proceedings*. https://www.iccs-meeting.org/archive/iccs2021/papers/127420059.pdf