Universiteit Leiden

Master Computer Science

Effect of Audio augmentations on Genre Classification using Contrastive learning

Name:	Piyush Dash
Student ID:	3671097
Date:	25/11/2024
Specialisation:	Data Science
1st Supervisor:	Dr. Erwin Bakker
2nd supervisor:	Prof. dr. M.S. Lew

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Effect of Audio augmentations on Genre Classification using Contrastive learning

Master Thesis

Piyush Dash

December 2, 2024

1 ABSTRACT

Self-supervised learning (SSL) and contrastive learning are essential for extracting meaningful representations from vast amounts of unlabeled audio data, significantly enhancing classification performance. While most contrastive learning research focuses on images and spectrograms, there are fewer works focused on raw audio inputs. We propose an enhanced framework using SampleCNN [14] for feature extraction directly from raw audio, enabling efficient transfer learning. Our approach builds on Spijkervet et al.'s CLMR framework[21], adapting SimCLR for music by introducing novel audio augmentations and architectural modifications like residual networks in the linear evaluation.

Our study explores the impact of audio augmentations, including harmonic distortion, dynamic range compression, and simulated radio effect, on music genre classification. We benchmark our approach on the fault-filtered GTZAN dataset and assess scalability using subsets of the Free Music Archive (FMA) dataset [6].

2 Introduction

Traditionally, methods which used CNNs and supervised learning have been successful in tasks like key detection, and music recommendation. However, these depend a lot on the availability of labeled datasets, which are difficult and costly to curate. Whereas, in the context of music, there is a lot of raw, unlabeled data readily available [12]. This challenge has led to an increased interest in self-supervised learning (SSL) techniques, which aim to learn meaningful representations directly from unlabeled data. SSL's ability to generalize within smaller datasets makes it very attractive and suitable for music classification, where a shortage of labeled data can make it hard to proceed. Recent work shows SSL good for the task of genre classification, because it learns to build robust representations that capture important audio features even without relying on labeled data [19].

Historically, music genre classification has relied on spectrogram-based methods such as mel-spectrograms and MFCCs to represent audio in time-frequency domains [22, 16]. While effective, these methods often aggregate information over time, potentially losing subtle temporal dynamics crucial for genre differentiation, particularly in genres characterized by distinctive dynamic ranges or harmonic features. Models like ResNet and VGG, initially designed for image data, have been adapted to process spectrograms for musical tasks [9, 5], though this adaptation may compromise the temporal precision needed for nuanced genre classification. A prominent approach within SSL is

contrastive learning, which encourages models to learn invariant representations by distinguishing between augmented versions of the same sample (positive pairs) and other samples (negative pairs)[24]. By leveraging augmentations like pitch shifts or time stretching, contrastive learning enables models to focus on essential genre characteristics that remain consistent across different acoustic and recording environments [27]. Recent work, such as Spijkervet's Contrastive Learning of Musical Representations (CLMR) [21], has shifted focus toward using raw audio waveforms directly, avoiding the transformation of raw audio to image spectrogram construction. By analyzing the original signal, methods that work with raw waveforms might potentially extract richer features, allowing for a deeper comprehension of the signal's properties in both the frequency and temporal domains. The SampleCNN architecture in CLMR processes raw audio using small convolutional kernels, learning features directly from the data without needing preprocessing like log-scaling or normalization[14]. This method preserves both temporal and frequency details, often lost in spectrograms, making it effective for music genre classification.

The fault-filtered GTZAN dataset, a curated version of the original dataset is a widely recognized benchmark in music genre classification[23]. It has been pivotal in evaluating the effectiveness of various algorithms for music genre classification. Data augmentation is crucial in contrastive learning, as it generates positive pairs that simulate real-world variations in audio data. Beyond common augmentations like pitch shifting and time stretching, our study introduces three distinct augmentations—Dynamic Range Compression (DRC), Harmonic Distortion, and Radio Effect—each inspired by real-life recording environments. DRC mimics volume compression to balance dynamic levels, potentially enhancing genre classification by stabilizing amplitude variations[17], especially in genres with wide dynamic ranges like classical and metal. Harmonic Distortion, by introducing controlled overtones, enriches the harmonic content, making it valuable for genres where timbral characteristics are significant, such as rock and jazz. Radio Effect, which adds band-pass filtering and background noise, emulates lower-quality transmission audio, offering robustness to timbral deterioration and improved generalization in a variety of playback settings and environments. We discuss further in Section 8 how we simulate these effects for our experiments.

In this study, we leverage the CLMR model, which was originally trained on the MagnaTagATune dataset [13]. Inspired by the SimCLR framework, it extracts meaningful representations directly from raw audio data, enabling downstream tasks like music genre classification[21]. This process effectively combines self-supervised learning on large-scale audio datasets with a lightweight classifier for fine-tuned evaluation.

Our contributions include introducing novel, real-life recording environments and human experience-inspired augmentations aimed at enhancing contrastive learning performance for genre classification. We aim to determine whether genre-specific augmentations can reduce mis-classifications in challenging genres and improve classification robustness across all genres. Additionally, we evaluate the impact of more complex architectures in the linear evaluation stage, exploring the effect of additional layers and residual blocks, as discussed in the section.This study addresses two key questions:

- **How do novel data augmentation techniques impact genre classification performance?** We investigate the effects of augmentations like harmonic distortion, dynamic range compression, and formant shifting on genre classification accuracy. While the augmentations in the original work yielded promising results on certain genres like Classical and Metal, our goal is to assess if our novel augmentations inspired by real life recording environments hypothesized to further enhance model robustness by working well across genres where the model fails to classify correctly.

- **What improvements can enhanced classification architectures bring to genre classification?** We experiment with more complex linear classification architectures beyond the basic multi-layer perceptron, incorporating additional layers and residual blocks to examine if we can achieve better accuracy and genre wise performance.

The rest of this paper is organized as follows: Section 3 reviews related work in music genre classification. Section 4 discusses the datasets used, including subsets of the Free Music Archive (FMA) dataset [6]. Section 5 introduces the main definitions and metrics used in this study. Section 6 describes the baseline network CLMR, while Section 7 details additional network architectures. Section 8outlines the experimental setup, including parameters and procedures, and Sections 9 and 10 present the results and conclusions.

3 Related Work

Representation learning focuses on discovering features that simplify prediction tasks and improve robustness to complex variations in natural data [2]. In the context of supervised learning for music genre classification, several methods have achieved notable success on the GTZAN dataset. Zhang et al. [26] reported an accuracy of 87.4% using a 10-layer CNN combined with classical data augmentation techniques. They observed that cutting songs into smaller 3-second clips significantly improved classification accuracy, consistent with earlier findings by Gjerdingen and Perrott [8]. Liu et al. [15] utilized a bottom-up broadcasting architecture to process time-frequency information from mel-spectrograms, achieving a state-of-the-art accuracy of 93.9%. Additionally, K-nearest neighbors (K-NN) with 3-second input features achieved an accuracy of 92% [18]. These results highlight the importance of advanced architectures and diverse training data for improving classification performance.

In contrast, unsupervised and self-supervised learning (SSL) approaches aim to extract meaningful representations without labeled data. Generative modeling and likelihood-based methods [11, 10] attempt to reconstruct observations from learned representations. Among SSL techniques, contrastive learning has proven particularly effective in audio tasks. Contrastive Predictive Coding (CPC) [24] introduced a universal contrastive learning approach for tasks such as speaker and phoneme classification using raw audio. Saeed et al. [20] developed the COLA framework, which processes audio segments into log-mel spectrograms and applies a multi-class cross-entropy loss (N-pair loss). COLA achieved 73% accuracy on music genre classification, demonstrating the potential of contrastive methods for extracting genre-relevant features. Similarly, CLAR [1], inspired by SimCLR, applied augmentations like time-stretching, pitch shifting, and noise injection, leading to improved representation quality and training efficiency.

Using pretrained masked autoencoders, M2D achieved 83% accuracy on GTZAN by encoding visible patches and predicting masked patch representations. Castellón et al. explored contrastive pre-training using OpenAI's Jukebox model, achieving approximately 68% accuracy with a million-parameter setup and 79% accuracy with a billion-parameter setup on audio classification benchmarks[3]. The CLMR framework [21], which directly operates on raw audio waveforms, achieved 55% accuracy on the GTZAN dataset using eight audio-inspired augmentations under contrastive learning, when pretrained on MagnaTagATune dataset. The study highlighted misclassifications in several genres like Pop, Jazz, Rock and Blues, emphasizing the need to have robustness acrros closely related genres. Prior research indicates that applying compression to test data can improve classification accuracy as demonstrated by [17].

Motivated by these advancements, our study explores novel augmentations inspired by real-world acoustic environments, such as dynamic range compression, harmonic distortion, and radio effects, to improve model performance. These augmentations address the need for greater robustness in genre classification, especially for closely related genres such as blues and jazz, pop, and hiphop. While harmonic distortion emphasizes overtones characteristic of genres like rock and metal, radio effects simulate low-fidelity playback environments. By evaluating these strategies, we aim to contribute to the development of more robust music genre classification models. On top of the simple linear classifier, we introduce 2 classifiers; one with additional layers and one with Residual blocks.

4 Datasets

This section provides an overview of the primary datasets utilized in this study: MagnaTagATune, FMA, and the fault-filtered version of GTZAN. Each dataset offers unique characteristics pertinent to music genre classification and analysis.

- **Pretraining Dataset MagnaTagATune** : The MagnaTagATune dataset comprises over 25,000 music clips, each approximately 29 seconds long, annotated with multiple tags describing various attributes such as genre, instrumentation, and mood. Players of the TagATune game contributed insightful tags for the music clips, which allowed for the collection of these annotations. The dataset is a useful tool for music information retrieval problems because it consists of a variety of genres, including rock, jazz, and classical [13]. The self-supervised pretraining on this dataset allows the model to learn general audio representations.
- **FMA Dataset:** The Free Music Archive (FMA) dataset is an extensive collection of 106,574 tracks from 16,341 artists and 14,854 albums, organized into a hierarchical taxonomy of 161 genres[7]. It provides full-length, high-quality audio files along with pre-computed features and metadata, including track and user-level information, tags, and free-form text such as artist biographies. We will utilise the Small dataset of the FMA for our purposes. It contains 8000 songs of balanced 8 genres of 1000 songs each.
- **GTZAN Fault-Filtered Dataset:** The GTZAN dataset is a widely used benchmark for music genre classification, containing 1,000 audio tracks each 30 seconds long, evenly distributed across 10 genres[23]. However, it has been identified to contain several faults, including repetitions, mislabelings, and distortions. A fault-filtered version of GTZAN has been curated to address these issues, removing problematic tracks and ensuring no artist repetition across training, validation, and test sets. This refined version contains 443 and 290 files in the training and test dataset [23].

5 Fundamentals

In this chapter, the key fundamentals for this thesis are introduced, including concepts such as convolutional neural networks, transfer learning, data augmentations, and evaluation metrics. These are basic concepts and definitions used in the methods and approaches discussed in later chapters.

5.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are powerful architectures widely used in image recognition and have been adapted for audio classification tasks as well. CNNs use convolutional layers with

trainable kernels that slide over the input, producing feature maps that highlight important characteristics like edges or, in our case, musical patterns. Max-pooling layers are often applied to reduce the feature map dimensions, and dropout layers help mitigate overfitting by randomly setting certain activations to zero during training.

5.2 Transfer Learning

Transfer learning is a technique in which knowledge gained through one task or dataset is used to improve model performance on another related task and/or different dataset. So, transfer learning uses what has been learned in one setting to improve generalization in another setting.

5.2.1 SampleCNN as Encoder

SampleCNN is a specific CNN model architecture optimized for music audio data, operating directly on raw audio waveforms. It employs small 1D convolutional kernels to capture patterns in the audio signal. In the context of this thesis, SampleCNN serves as the encoder within the CLMR (Contrastive Learning of Musical Representations) framework, processing audio inputs to generate robust embeddings that are later used for genre classification tasks.

5.3 Data Augmentations

Data augmentation is a technique to enhance the variability of training data and improve model generalization. In the context of music, augmentations may include random cropping, polarity inversion, gain adjustments, and time/frequency manipulations. By making it more difficult for models to detect subtle changes in audio, these variations promote the extraction of generalizable features and aid in the learning of more robust representations.

5.4 Evaluation Metrics

To assess model performance, two key metrics are used: accuracy and confusion matrices.

5.4.1 Accuracy

Accuracy is defined as the ratio of correct predictions (True Positives and True Negatives) to the total number of predictions. In this study, accuracy reflects the proportion of correctly classified genres out of the total classifications:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP represents True Positives, TN represents True Negatives, FP represents False Positives, and FN represents False Negatives.

5.4.2 Confusion Matrix

The confusion matrix provides a detailed view of the model's performance by displaying the true versus predicted class labels. This matrix helps identify specific areas where the model may misclassify similar genres, enabling targeted improvements. An example confusion matrix for Jazz and Rock classifications is shown in Table 1.

Predicted \ Actual	Jazz	Rock
Jazz	5	3
Rock	7	3

Table 1: Confusion matrix example for genres Jazz and Rock.

In this example, Jazz and Rock genres show notable overlaps in misclassifications, The first row shows that 5 out of 12 Jazz classifications were correctly classified, but in many cases, Jazz and Rock are confused with each other.

6 Baseline : CLMR

6.1 Contrastive Learning

Contrastive learning is a self-supervised method that trains models by distinguishing between similar and dissimilar data points without needing explicit labels. In music genre classification, it involves learning genre-relevant features by pulling together representations of similar audio segments (such as augmented versions of the same track) and pushing apart those of different segments(can be considered as negative samples), allowing the model to directly extract important features from the audio input, such as rhythm and harmony.

6.2 Contrastive Loss Function

The contrastive loss function is a key component in contrastive learning frameworks, where the goal is to learn embeddings that bring similar pairs of samples closer together while pushing dissimilar pairs farther apart in the feature space[4]. This is achieved by minimizing the contrastive loss, which measures how well the model can distinguish between positive (similar) and negative (dissimilar) pairs[24].

The contrastive loss function can be defined as:

$$\mathcal{L}_{i,j} = -\log \frac{\exp\left(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp\left(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau\right)} \tag{2}$$

[24]where:

- $\mathbf{z}_i$ and $\mathbf{z}_j$ represent the embeddings of the anchor and positive samples, respectively.
- $\text{sim}(\cdot)$ denotes the similarity measure between the embeddings, typically computed as the cosine similarity:

$$\text{sim}(\mathbf{z}_i, \mathbf{z}_j) = \frac{\mathbf{z}_i \cdot \mathbf{z}_j}{\|\mathbf{z}_i\| \|\mathbf{z}_j\|} \tag{3}$$

- τ is a temperature parameter that controls the scale of the logits, influencing the sharpness of the similarity scores.
- N is the batch size, and the denominator in the contrastive loss includes all negative pairs (i.e., samples that are not the anchor-positive pair), making the task harder by increasing the number of negative samples.

- $\mathbb{1}_{[k \neq i]}$ is an indicator function that equals 1 when $k \neq i$, ensuring that the positive pair is excluded from the denominator.

The contrastive loss is often referred to as the *Normalized Temperature-scaled Cross-Entropy* (NT-Xent) loss. By using this loss function, the model is encouraged to maximize the similarity of the positive pair $\text{sim}(\mathbf{z}_i, \mathbf{z}_j)$, while minimizing the similarity between the anchor $\mathbf{z}_i$ and all other negative examples in the batch $\mathbf{z}_k$ for $k \neq i$[24].

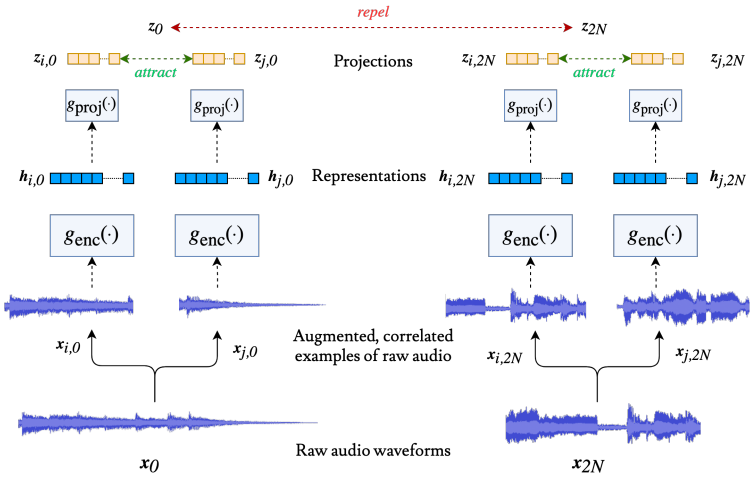


Figure 1: Illustration of the CLMR framework: it processes raw audio waveforms, applying contrastive learning in the latent space of augmented audio pairs to learn meaningful musical representations. Image source[21]

This objective enables the model to learn embeddings that capture important semantic similarities in the data, which can be leveraged in downstream tasks such as classification or retrieval. In this thesis, transfer learning allows us to use a model pre-trained on the MagnaTagATune dataset and apply it to genre classification on the GTZAN dataset. For our baseline, we utilize CLMR (Contrastive Learning of Musical Representations)[21] to compare the out-of-domain applicability for genre classification on the GTZAN dataset with our method. Augmentations used in Original experiment were : 8 augmentations with set prtobabilities applied stochastically. These augmentations were applied in a sequence which was hypothesized to achieve contrastive representations of the audio samples which will help the model learn about different genres better.

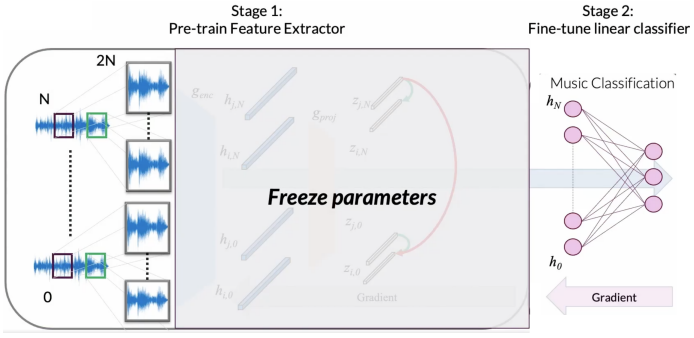


Image source[25]

Figure 2: Overview of the CLMR process

With an encoder built to handle raw audio waveforms with contrastive learning, our CLMR baseline configuration is based on Spijkervet and Burgoyne's implementation and is especially tested on the GTZAN dataset [25]. Each of the encoder's nine 1D convolutional layers is paired with max pooling, ReLU activation, and batch normalization. Inspired by the SimCLR architecture created by Chen et al.[4], these layers are followed by a projection head that converts the encoded representations into a 128-dimensional latent space where contrastive loss is applied. The original CLMR setup for SampleCNN uses NT-Xent contrastive loss for optimization, a batch size of 96, and an input sample rate of 22,050 Hz[21].

Table 2 provides the parameter configurations for this baseline network:

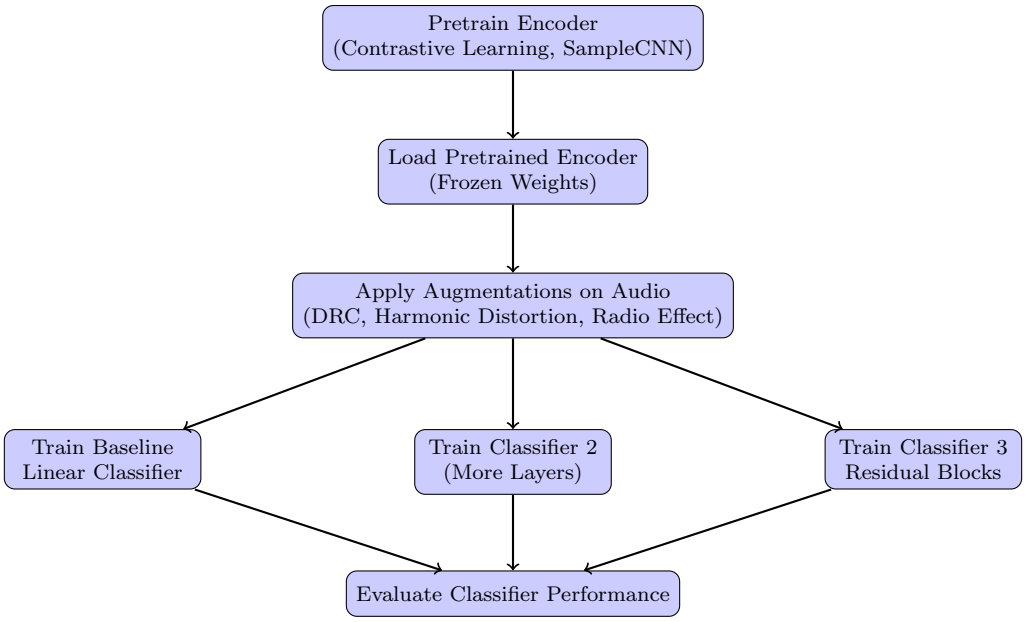
Parameter	**Value**
Input size	59,049 samples at 22,050 Hz
Batch size	96
Optimizer	Adam
Contrastive Loss	NT-Xent with temperature scaling (0.5)
Encoder Parameters	2.5 million

Table 2: Parameter configurations for the baseline CLMR network.

In the linear evaluation phase, the learned representations from the frozen encoder are classified using a linear classifier on GTZAN, assessing the model's transferability without in-domain data fine-tuning. This baseline allows us to evaluate the generalization capacity of self-supervised CLMR when faced with a new dataset and provides insights into CLMR's applicability in cross-domain genre classification task.

7 CLMR +

Our method investigates the impact of three novel audio augmentations—Dynamic Range Compression, Harmonic Distortion, and Radio Effect—on the training set for music genre classification. The performance is evaluated using a baseline linear classifier, a modified classifier with additional layers, and another with residual blocks. The results demonstrate how these augmentations enhance feature learning and classification accuracy across different classifier model architectures.


Pretrain Encoder
(Contrastive Learning, SampleCNN)
Load Pretrained Encoder
(Frozen Weights)
Apply Augmentations on Audio
(DRC, Harmonic Distortion, Radio Effect)
Train Baseline
Linear Classifier
Train Classifier 2
(More Layers)
Train Classifier 3
Residual Blocks
Evaluate Classifier Performance

Pipeline Overview:

- **Pretraining Encoder with Contrastive Learning on MagnaTagATune:** The SampleCNN encoder is pretrained on the MagnaTagATune dataset using contrastive self-supervised learning. Each audio sample is augmented twice to create a positive pair, while other samples in the batch serve as negatives. The contrastive loss ensures the encoder learns meaningful musical representations, capturing general musical features like rhythm, pitch, and timbre.
- **Loading Pretrained Encoder Weights (Frozen):** The pretrained weights of the encoder are loaded, and the encoder is kept frozen by setting `requires_grad = False`. This ensures the pretrained representations are used without being modified during training.
- **Applying Data Augmentation on GTZAN:** The GTZAN dataset undergoes various augmentations during training to enhance data variability. These include :
  - **Harmonic Distortion:** Aims to improve accuracy in genres with prominent harmonic content, such as Metal and Rock, while avoiding over-distortion in subtler genres like Jazz and Blues.
  - **Dynamic Range Compression:** Helps manage dynamic variations in genres like Classical and Metal, ensuring consistent amplitude features across tracks.
  - **Radio Effect:** Focuses on midrange frequencies critical for instruments and vocals, improving performance in genres like Blues and Jazz where timbral and melodic elements dominate.

  Validation and test datasets use minimal augmentations, such as normalization and fixed-length padding.
- **Training Linear Classifier on Extracted Representations:** The `LinearEvaluation` class implements the linear classifier, which maps the frozen encoder's representations to GTZAN genre labels. A single linear layer maps representations (hidden_dim = 512) directly to output classes (output_dim = 10). Cross-Entropy Loss is used as the criterion, with optimization handled by the Adam optimizer. The ReduceLROnPlateau scheduler reduces the learning rate when validation loss plateaus.
- **Evaluating Classifier Performance:** The `evaluate` function computes predictions on the test set and calculates key metrics like Accuracy, Classification Report and Confusion Matrix to visualize the classifications. Evaluation ensures the quality of the pretrained representations and the classifier's performance in the genre classification task.

Linear Classifier Architecture: The last projection layer of the encoder of 128 dimensions is removed and instead the penultimate 512-dimension layer is utilised by the classifier. The linear classifier processes the fixed-length 512-dimensional representations extracted by the frozen encoder. Its simplicity allows direct evaluation of the pretrained features' quality without introducing additional complexity. By experimenting with both configurations, we assess the robustness and generalization of the learned representations for music genre classification. We propose two enhanced linear classifiers in addition to the baseline simple linear evaluator. The first, Classifier 2, introduces additional layers with batch normalization and dropout to increase the capacity for capturing intricate patterns in the data. The second, Classifier 3, integrates residual blocks inspired by ResNet, enabling deeper feature extraction while maintaining gradient flow. Our method, CLMR+, will first be evaluated on the baseline linear classifier and subsequently on these two advanced classifiers to explore their effectiveness in music genre classification.

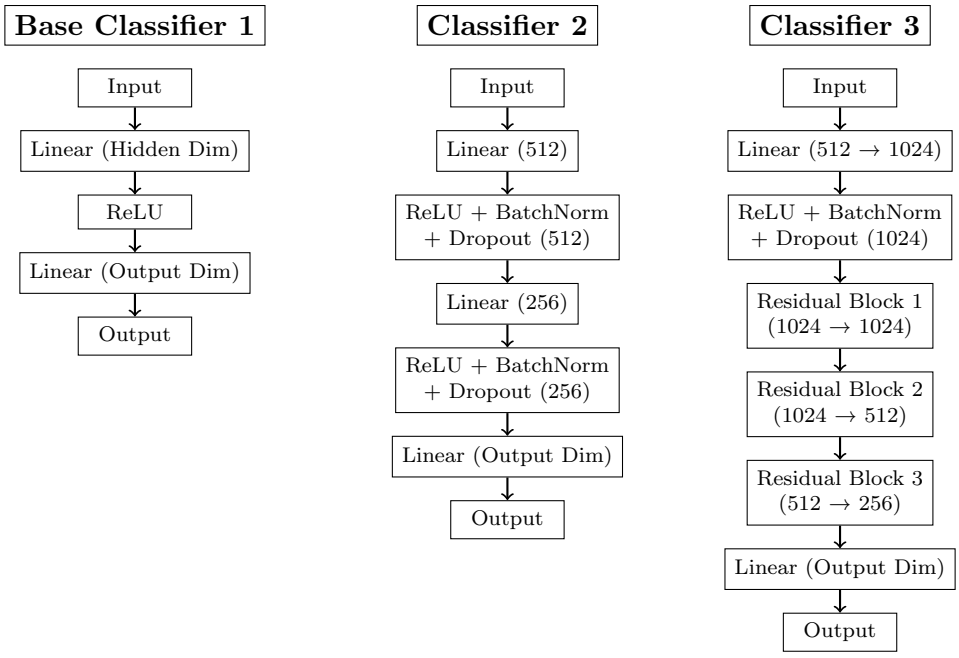


Figure 3: Architectures of Base Classifier 1, Classifier 2 with additional layers, and Classifier 3 with residual blocks.

- **Classifier 2 (Additional Layers):**
  - By introducing extra dense layers with batch normalization and dropout, this classifier enhances the model's ability to capture more intricate patterns while reducing overfitting.
  - The architecture consists of three fully connected layers, each followed by ReLU activation, batch normalization, and dropout for regularization.
- **Classifier 3 (Residual Blocks):**
  - Residual connections improve gradient flow and allow for deeper architectures, enabling the model to learn more complex representations without degradation.
  - The architecture includes a linear layer for initial processing, followed by 2 residual blocks and a final linear layer for classification.

8 Experimental Setup

In this section, we outline the experimental setup designed to investigate the impact of audio augmentations on genre classification performance. Our approach utilizes a pretrained encoder, frozen during linear evaluation, and systematically applies augmentations to the GTZAN dataset in a controlled manner. Each augmentation is tested under a base setting and two alternative settings, with application probabilities of $p = 0.2$, $p = 0.5$, and $p = 0.8$.

8.1 Base Setting : CLMR+

A foundational augmentation, Random Cropping, was applied across all experiments to ensure uniformity in input length and prevent overfitting to specific segments of audio. This served as a

base augmentation, over which 3 additional augmentations were layered. On top of the base setting, we designed two alternative configurations to represent lower and higher effects of the augmentation. These configurations were applied with varying probabilities p, allowing us to systematically analyze how different augmentation intensities and their probabilities influenced the audio features and, subsequently, the genre classification performance. We will then take the best values of each augmentation setting, based on imapct on genre classification and overall accuracy and use them to assess the 2 classifiers we proposed in later sections.

8.1.1 Harmonic Distortion

Harmonic Distortion introduces controlled nonlinearities, adding harmonic richness to the signal to create harmonic overtones. It emulates genre-specific effects like overdrive and guitar distortion.

Algorithm 1 Harmonic Distortion

Input: Audio waveform $\mathbf{w}$, distortion level δ
Output: Distorted waveform $\mathbf{w}_{\text{distorted}}$
$\mathbf{w} \leftarrow \mathbf{w} \cdot (1 + \delta)$ ▷ Apply gain scaling
$\mathbf{w}_{\text{distorted}} \leftarrow \tanh(\mathbf{w})$ ▷ Apply soft clipping
Return $\mathbf{w}_{\text{distorted}}$

Settings:

- Base: $\delta = 0.5$
- Alternative 1: $\delta = 0.25$ (milder distortion for subtler harmonic enrichment)
- Alternative 2: $\delta = 0.7$ (higher distortion for prominent harmonic features)

8.1.2 Dynamic Range Compression (DRC)

Dynamic Range Compression attenuates louder parts of the signal, enhancing consistent feature extraction. This augmentation is particularly useful for handling dynamic amplitude variations in genres like Classical and Pop.

Algorithm 2 Dynamic Range Compression

Input: Audio waveform $\mathbf{w}$, threshold θ_{dB}, ratio r
Output: Compressed waveform $\mathbf{w}_{\text{compressed}}$
$\theta \leftarrow 10^{(\theta_{\text{dB}}/20)}$ ▷ Convert threshold from dB to amplitude
mask $\leftarrow |\mathbf{w}| > \theta$
$\mathbf{w}_{\text{compressed}}[\text{mask}] \leftarrow \text{sign}(\mathbf{w}) \cdot (\theta + (|\mathbf{w}| - \theta)/r)$
Return $\mathbf{w}_{\text{compressed}}$

Settings:

- Base: $\theta_{\text{dB}} = -20.0$, $r = 2:1$
- Alternative 1: $\theta_{\text{dB}} = -25.0$, $r = 1.5:1$ (gentler compression for preserving dynamics)
- Alternative 2: $\theta_{\text{dB}} = -15.0$, $r = 3:1$ (stronger compression for louder dynamics)

8.1.3 Radio Effect

The Radio Effect applies a band-pass filter and Gaussian noise to simulate the frequency response of radio transmissions. This augmentation focuses on midrange frequencies critical for genres like Rock and Jazz.

Algorithm 3 Radio Effect

Input: Audio waveform $\mathbf{w}$, sample rate f_s, noise level ν, band-pass limits $[f_{\text{low}}, f_{\text{high}}]$
Output: Radio-augmented waveform $\mathbf{w}_{\text{radio}}$
$\mathbf{w}_{\text{filtered}} \leftarrow \text{BandPass}(\mathbf{w}, f_{\text{low}}, f_{\text{high}}, f_s)$
$\mathbf{w}_{\text{noisy}} \leftarrow \mathbf{w}_{\text{filtered}} + \nu \cdot \text{GaussianNoise}()$
$\mathbf{w}_{\text{radio}} \leftarrow \text{clip}(\mathbf{w}_{\text{noisy}}, -1, 1)$
Return $\mathbf{w}_{\text{radio}}$

Settings:

- Base: $f_{\text{low}} = 200\,\text{Hz}$, $f_{\text{high}} = 4000\,\text{Hz}$, $\nu = 0.02$
- Alternative 1: $f_{\text{low}} = 300\,\text{Hz}$, $f_{\text{high}} = 3000\,\text{Hz}$, $\nu = 0.02$ (narrower band to emphasize midrange frequencies)
- Alternative 2: $f_{\text{low}} = 150\,\text{Hz}$, $f_{\text{high}} = 4500\,\text{Hz}$, $\nu = 0.01$ (broader range with reduced noise)

9 Results

9.1 CLMR+

To evaluate the impact of different augmentation configurations on genre classification performance, we conducted experiments using RandomCrop as the base augmentation technique, supplemented with additional augmentations such as Radio Effect, Dynamic Range Compression (DRC), and Harmonic Distortion. Each augmentation was applied with varying probabilities to assess their effectiveness individually and in combination with the base augmentation. The following analysis compares the classification performance of the baseline model (RandomCrop) against models augmented with one additional technique (base +1 aug) across overall metrics and genre-specific metrics. The confusion matrix in Figure 4 displays the genre classification performance using the original set of augmentations in the CLMR framework, as implemented by Spijkervet [21]. This baseline configuration reveals how well the model distinguished between different music genres under the original augmentation settings. High precision in specific genres, such as Metal and Classical, suggests robust feature extraction for certain music styles. However, noticeable misclassifications in genres like Blues and Rock indicate potential overlaps in the learned representations for these categories.

The primary goal of these experiments was to explore the effectiveness of various audio augmentations on genre classification, benchmarked against the GTZAN dataset. The results are summarized in Tables 3, 5, and 6. Table 3 compares the performance of our best model configurations with prior state-of-the-art methods. Our approach demonstrated competitive accuracy and F1-scores, highlighting the effectiveness of tailored augmentations with specific parameter settings in improving genre classification.

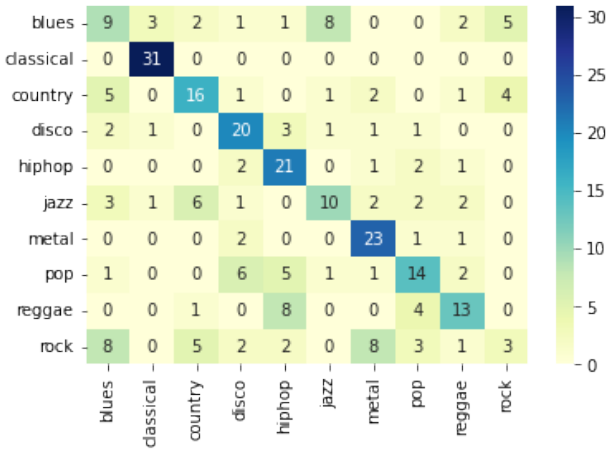


Figure 4: Confusion Matrix for Genre Classification using Original CLMR Augmentations by Spijkervet[25]

Table 3: Comparison with Prior Work and Our Best Augmentation Settings

Method	**Accuracy (%)**	**Precision**	**F1-Score**
Spijkervet et al. [CLMR]	55.2	0.52	0.51
Baseline (Random Cropping)	59.0	0.58	0.58
CLMR+ (Harmonic Distortion, $\delta = 0.7$, $p = 0.5$)	**62.1**	**0.62**	**0.60**

Table 4: CLMR Classification Report

Genre	**Precision**	**F1-Score**	**Common Misclassifications**
Blues	0.321	0.305	Rock, Jazz
Classical	0.861	0.925	None significant
Country	0.533	0.533	Blues, Rock
Disco	0.571	0.625	Hip-hop
Hip-hop	0.512	0.618	Pop, Reggae
Jazz	0.476	0.417	Country, Pop
Metal	0.605	0.707	Rock
Pop	0.519	0.491	Disco, Hip-hop
Reggae	0.565	0.531	Hip-hop, Pop
Rock	0.250	0.136	Metal, Blues
Macro Average	0.520	0.521	-
Weighted Average	0.553	0.539	-

9.1.1 Best Settings by Genre CLMR+

The effectiveness of different augmentations varied across genres. Table 5 highlights the best augmentation settings for each genre, including specific parameter values and probabilities, with the optimal settings marked in bold. Results demonstrate that specific augmentations, such as Dynamic Range Compression (DRC) and Harmonic Distortion, were particularly beneficial for genres like Metal and Jazz when appropriate parameter values were used.

Table 5: Best Augmentation Settings for each genre with parameter Values for CLMR+

Genre	Augmentation	Parameters	Precision	F1-Score
Blues	Radio Effect	$f_{\text{low}} = 300\,\text{Hz}$, $f_{\text{high}} = 3000\,\text{Hz}$, $\nu = 0.02$, $p = 0.8$	0.31	**0.33**
Classical	Radio Effect	$f_{\text{low}} = 300\,\text{Hz}$, $f_{\text{high}} = 3000\,\text{Hz}$, $\nu = 0.02$, $p = 0.5$	1.00	**1.00**
Country	DRC	$\theta_{\text{dB}} = -20\,\text{dB}$, $r = 2:1$, $p = 0.8$	0.69	**0.71**
Disco	DRC	$\theta_{\text{dB}} = -25\,\text{dB}$, $r = 1.5:1$, $p = 0.2$	0.63	**0.69**
Hip-Hop	DRC	$\theta_{\text{dB}} = -20\,\text{dB}$, $r = 2:1$, $p = 0.8$	0.80	**0.77**
Jazz	DRC	$\theta_{\text{dB}} = -20\,\text{dB}$, $r = 2:1$, $p = 0.8$	0.43	**0.48**
Metal	Harmonic Distortion	$\delta = 0.5$, $p = 0.5$	0.96	**0.92**
Pop	Harmonic Distortion	$\delta = 0.25$, $p = 0.2$	0.67	**0.59**
Reggae	Harmonic Distortion	$\delta = 0.7$, $p = 0.5$	0.61	**0.67**
Rock	Radio Effect	$f_{\text{low}} = 300\,\text{Hz}$, $f_{\text{high}} = 3000\,\text{Hz}$, $\nu = 0.02$, $p = 0.8$	0.79	**0.48**

9.1.2 Best Augmentations with Random Cropping

The baseline performance using only Random Cropping achieved an overall accuracy of 59%. Table 6 outlines the genre-wise performance under this configuration.

Table 6: Genre-wise Performance : No augmentation (Random Cropping)

Genre	Precision	Recall	F1-Score
Blues	0.15	0.13	0.14
Classical	0.94	1.00	0.97
Country	0.62	0.67	0.65
Disco	0.55	0.76	0.64
Hip-Hop	0.67	0.81	0.73
Jazz	0.38	0.52	0.44
Metal	0.85	0.85	0.85
Pop	0.62	0.53	0.57
Reggae	0.58	0.54	0.56
Rock	0.50	0.19	0.27

Our augmentation strategies contributed to balanced improvements across multiple genres, enhancing the model's generalization capabilities and some enhanced performance compared to CLMR results.

- **Harmonic Distortion ($\delta = 0.7, 0.5$, $p = 0.5$):** This augmentation worked exceptionally well for genres like Metal and Rock, where distortion is a natural characteristic of the music. For Disco, it also improved performance significantly under higher distortion levels ($\delta = 0.7$). Additionally, Pop benefited under a slightly lower distortion setting ($\delta = 0.5$), demonstrating the augmentation's adaptability across genres. Overall, Harmonic Distortion contributed positively to multiple genres.
- **Dynamic Range Compression (DRC) ($\theta_{\text{dB}} = -20\,\text{dB}$, $r = 2:1$, $p = 0.8$):** It was particularly effective for improving performance in genres like Reggae, Pop, Country, and Hip-Hop. The lower threshold ($\theta_{\text{dB}} = -20$) allowed the model to capture quieter musical elements, which enhanced recognition in subtler genres like Blues . A gentler ratio (2:1) maintained the natural dynamics of these genres. A variant of DRC ($\theta_{\text{dB}} = -25$, $r = 1.5:1$, $p = 0.2$) was highly effective for Disco, Pop, and Reggae, delivering better accuracy overall for these genres.

- **Radio Effect ($f_{\text{low}} = 300\,\text{Hz}$, $f_{\text{high}} = 3000\,\text{Hz}$, $\nu = 0.02$, $p = 0.8$):** This augmentation, which focuses on midrange frequencies, was especially effective for Classical and Hip-Hop, where melodic and vocal elements dominate. The band-pass filter emphasized critical frequency ranges, leading to significant improvements. It performed well for Jazz and Rock under settings with noise level $\nu = 0.02$, low cutoff $f_{\text{low}} = 300\,\text{Hz}$, and high cutoff $f_{\text{high}} = 3000\,\text{Hz}$, with $p = 0.5$, achieving accuracy of 62%. However, under $p = 0.8$, it excelled for Blues and Rock, highlighting its adaptability across genres.
  Genres like Blues, Jazz, and Rock continued to pose classification challenges, with persistent misclassifications such as Blues being classified as Jazz and vice versa. However, certain augmentation settings led to improved accuracy for both genres, and overall, the classification performance for Rock showed notable improvement as seen in 7.

Table 7: Comparison of Precision and F1-Score: CLMR vs CLMR+

Genre	CLMR Precision	CLMR F1-Score	CLMR+ Precision	CLMR+ F1-Score
Blues	0.321	0.305	0.31	0.33
Classical	0.861	0.925	1.00	1.00
Country	0.533	0.533	0.69	0.71
Disco	0.571	0.625	0.63	0.69
Hip-Hop	0.512	0.618	0.80	0.77
Jazz	0.476	0.417	0.43	0.48
Metal	0.605	0.707	0.96	0.92
Pop	0.519	0.491	0.67	0.59
Reggae	0.565	0.531	0.61	0.67
Rock	0.250	0.136	0.79	0.48

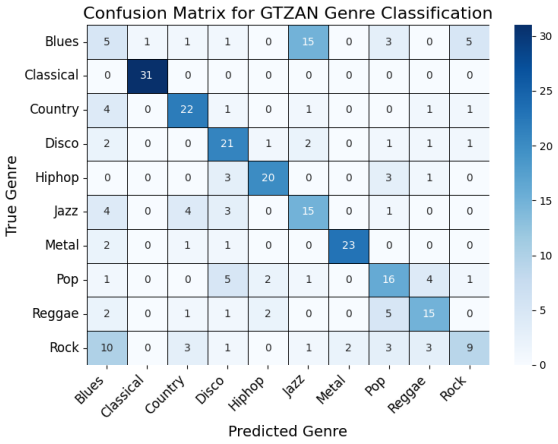


Figure 5: Confusion Matrix for CLMR+Classifier1 with DRC

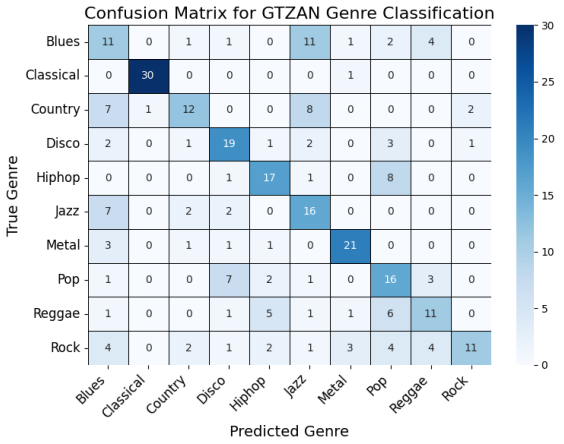


Figure 6: Confusion Matrix for CLMR+Classifier1 with Radio Effect

By systematically testing different configurations with varying parameter values and probabilities, we get the best settings for each augmentation and proceed to use them.

9.2 CLMR+ Classifier 2 and 3

Classifier 2: We test the best performances of the 3 augmentations using the 2 proposed classifiers with the best performing parameter values from the previous experiments and compare them with the CLMR+ base classifier results. Under the Radio effect, we see the overall accuracy improved to 63%, with Classical achieving near-perfect performance. Blues, Jazz, and Rock remain challenging, showing frequent misclassifications such as Blues as Jazz or Pop, and Rock as Blues or Pop. Notable improvements were observed in Country, Disco, and Hip-Hop. With the Dynamic range compression, we also achieve comparable performances.

Classifier3 :The DRC augmentation achieved an accuracy of 58%, with Classical performing best (F1-score: 0.98), but struggled with challenging genres like Blues and Jazz. Misclassification trends included Blues frequently being classified as Jazz or Pop and Rock as Blues or Pop. The Radio Effect augmentation slightly improved accuracy to 60%, again excelling in Classical . However, Pop (F1-score: 0.47) and Rock (F1-score: 0.42) remained challenging, with Blues often misclassified as Pop or Disco, and Rock confused with Blues or Pop. Similarly, the Harmonic Distortion augmentation achieved 59% accuracy, Blues often classified as Pop or Disco, and Rock as Blues or Reggae.

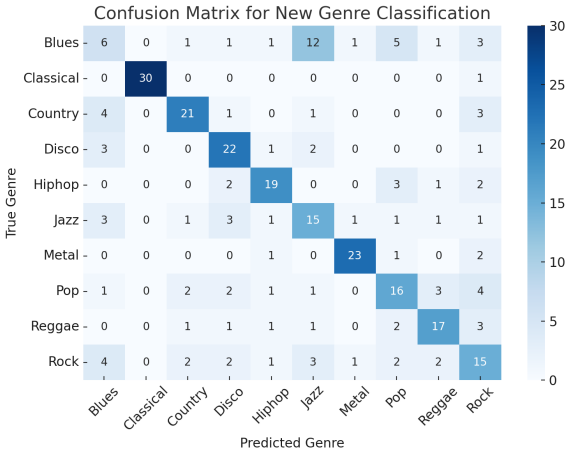


Figure 7: Confusion Matrix for CLMR+Classifier2 with Radio effect

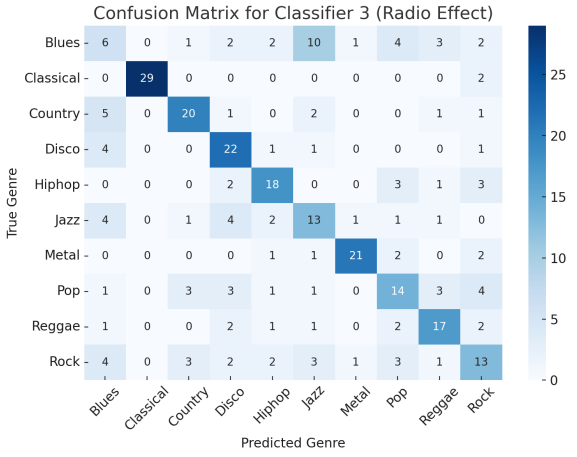


Figure 8: Confusion Matrix for CLMR+Classifier3 with Radio Effect

Table 8: Genre wise performance for CLMR+ Classifier 2

Genre	F1-Score (Best)	Precision	Augmentation Method
Blues	0.19	0.20	Radio Effect ($f_{\text{low}} = 300\,\text{Hz}$, $f_{\text{high}} = 3000\,\text{Hz}$, $\nu = 0.02$, $p = 0.8$)
Classical	0.99	0.98	Radio Effect ($f_{\text{low}} = 300\,\text{Hz}$, $f_{\text{high}} = 3000\,\text{Hz}$, $\nu = 0.02$, $p = 0.8$)
Country	0.67	0.65	Harmonic Distortion ($\delta = 0.7$, $p = 0.5$)
Disco	0.68	0.63	DRC ($\theta_{\text{dB}} = -20\,\text{dB}$, $r = 2:1$, $p = 0.8$)
Hiphop	0.68	0.65	Radio Effect ($f_{\text{low}} = 300\,\text{Hz}$, $f_{\text{high}} = 3000\,\text{Hz}$, $\nu = 0.02$, $p = 0.8$)
Jazz	0.49	0.45	DRC ($\theta_{\text{dB}} = -20\,\text{dB}$, $r = 2:1$, $p = 0.8$)
Metal	0.87	0.87	Harmonic Distortion ($\delta = 0.7$, $p = 0.5$)
Pop	0.58	0.63	DRC ($\theta_{\text{dB}} = -20\,\text{dB}$, $r = 2:1$, $p = 0.8$)
Reggae	0.63	0.61	Harmonic Distortion ($\delta = 0.7$, $p = 0.5$)
Rock	0.34	0.45	Radio Effect ($f_{\text{low}} = 300\,\text{Hz}$, $f_{\text{high}} = 3000\,\text{Hz}$, $\nu = 0.02$, $p = 0.8$)

Table 9: Genre wise performance for CLMR+ Classifier 3

Genre	F1-Score (Best)	Precision	Augmentation Method
Blues	0.26	0.30	Harmonic Distortion ($\delta = 0.7$, $p = 0.5$)
Classical	0.98	1.00	DRC ($\theta_{\text{dB}} = -20\,\text{dB}$, $r = 2:1$, $p = 0.8$)
Country	0.69	0.71	Radio Effect ($f_{\text{low}} = 300\,\text{Hz}$, $f_{\text{high}} = 3000\,\text{Hz}$, $\nu = 0.02$, $p = 0.8$)
Disco	0.70	0.65	Radio Effect ($f_{\text{low}} = 300\,\text{Hz}$, $f_{\text{high}} = 3000\,\text{Hz}$, $\nu = 0.02$, $p = 0.8$)
Hiphop	0.72	0.68	DRC ($\theta_{\text{dB}} = -20\,\text{dB}$, $r = 2:1$, $p = 0.8$)
Jazz	0.48	0.41	Radio Effect ($f_{\text{low}} = 300\,\text{Hz}$, $f_{\text{high}} = 3000\,\text{Hz}$, $\nu = 0.02$, $p = 0.8$)
Metal	0.87	0.88	DRC ($\theta_{\text{dB}} = -20\,\text{dB}$, $r = 2:1$, $p = 0.8$)
Pop	0.53	0.52	DRC ($\theta_{\text{dB}} = -20\,\text{dB}$, $r = 2:1$, $p = 0.8$)
Reggae	0.67	0.65	Radio Effect ($f_{\text{low}} = 300\,\text{Hz}$, $f_{\text{high}} = 3000\,\text{Hz}$, $\nu = 0.02$, $p = 0.8$)
Rock	0.45	0.48	Harmonic Distortion ($\delta = 0.7$, $p = 0.5$)

Table 10: Comparative Results for Different Models and Classifierswith F1 scores

Genre	CLMR	CLMR+	CLMR+Classifier2	CLMR+Classifier3
Blues	0.305	**0.33**	0.19	0.26
Classical	0.925	**1.00**	0.99	0.98
Country	0.533	**0.71**	0.67	0.69
Disco	0.625	0.69	0.68	**0.70**
Hip-Hop	0.618	0.77	0.68	**0.72**
Jazz	0.417	**0.48**	0.49	0.48
Metal	0.707	**0.92**	0.87	0.87
Pop	0.491	**0.59**	0.58	0.53
Reggae	0.531	**0.67**	0.63	0.67
Rock	0.136	**0.48**	0.34	0.45

9.3 FMA Experiments

Based on the previous experiments, we choose to go with the CLMR+base classifier1 as it did decently well with most genres to test the scalability across a bigger dataset. There are some new genres like Folk, International, Experimental, with some classic ones like Pop, Rock & Hiphop in the FMA_Small dataset. Across all augmentations, the best overall accuracy achieved was 53%. The top-performing genres included International and Hip-hop, with F1-scores ranging from 0.62 to 0.69 across Dynamic Range Compression (DRC), Radio Effect, and Harmonic Distortion augmentations. While International consistently excelled, Hip-hop also demonstrated strong performance with balanced precision and recall. However, certain genres remained challenging: Rock and Electronic showed low F1-scores (0.35–0.38) under DRC and Harmonic Distortion, while Folk and Instrumental struggled with moderate F1-scores (0.52–0.55) under the Radio Effect and Harmonic Distortion.

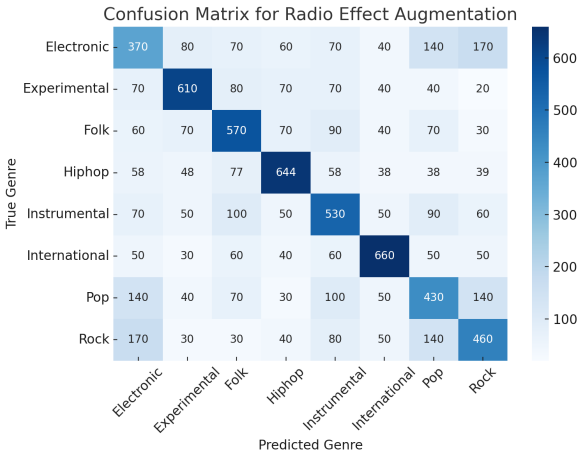


Figure 9: FMA_Small with CLMR+Classifier1 with Radio effect

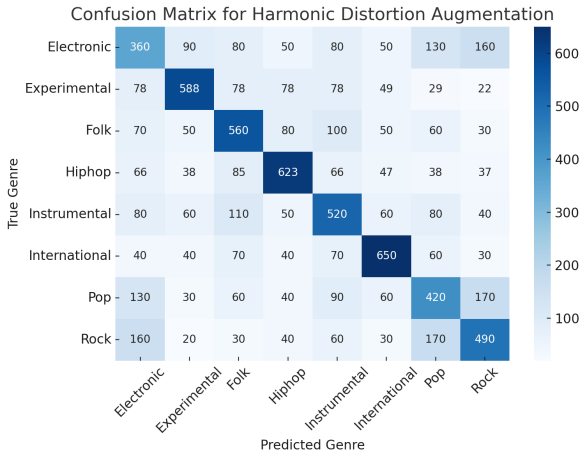


Figure 10: FMA_Small with CLMR+Classifier1 with Harmonic Distortion

Table 11: Genre-wise Best Results for FMA Small Dataset Across Augmentations

Genre	**F1-Score**	**Augmentation**	**Parameters**
Electronic	0.26	DRC	$p = 0.5$, $\theta_{\text{dB}} = -20$
Experimental	0.48	Radio Effect	$f_{\text{low}} = 300\,\text{Hz}$, $f_{\text{high}} = 3000\,\text{Hz}$
Folk	0.42	Radio Effect	$p = 0.8$, $\nu = 0.02$
Hiphop	0.52	DRC	$p = 0.5$, $\theta_{\text{dB}} = -20$
Instrumental	0.38	Harmonic Distortion	$\delta = 0.7$, $p = 0.5$
International	0.51	DRC	$p = 0.5$, $\theta_{\text{dB}} = -20$
Pop	0.46	Radio Effect	$f_{\text{low}} = 300\,\text{Hz}$, $f_{\text{high}} = 3000\,\text{Hz}$
Rock	0.44	Radio Effect	$p = 0.8$, $\nu = 0.02$

10 Conclusion

The experiments conducted with augmentations and classifiers demonstrate the importance of well considered augmentation methods, advanced classifier architectures, and parameter adjustment for music genre classification tasks. Augmentations like *Radio Effect*, *Dynamic Range Compression (DRC)*, and *Harmonic Distortion* were essential for improving performance across challenging genres, especially when employed with intermediate probability ($p = 0.2$ to $p = 0.5$). For instance, Harmonic Distortion ($\delta = 0.7$) raised F1-scores for distortion-heavy genres like metal and rock, while DRC ($\theta_{\text{dB}} = -20\,\text{dB}, r = 2:1$) enhanced categorization for dynamic genres like hip-hop and reggae.

In all setups, Classical consistently produced near-perfect results, which reflected its unique audio characteristics, but Blues, Rock, and Jazz greatly benefited from residual-based classifiers and tailored augmentations. The residual blocks in CLMR+Classifier3 worked best for complicated genres like jazz and rock, capturing nuanced features and improving F1-scores. Meanwhile, Classifier2 excelled in rhythmic genres like Disco and Hiphop by leveraging dropout and additional layers to enhance generalization. Augmentations applied at optimal probabilities introduced variability and improved model robustness without overwhelming genre-specific features, while over-augmentation led to performance degradation. These findings emphasize the necessity for deeper representations in difficult classification tasks and the importance of selecting augmentation kinds and parameters

with genre features.

Overall, the combination of genre-specific augmentations and advanced classifiers like CLMR+Classifier3 enhanced accuracy and generalization, which could be used for further research on real-world applications in music genre classification.

11 Future Work

We could focus on expanding the diversity and size of pre-training datasets could improve the generalization of self-supervised models, enabling them to capture more nuanced audio representations. Developing genre-specific augmentations tailored to the unique characteristics of challenging genres like Blues, Jazz, and Rock could help mitigate persistent misclassification trends. Based on the experiments, we could try a more comprehensive search for the parameters for the augmentations and the experiments involving them to get the best results. Though we were able to simulate the desired audio effects, we could try a more precise and perhaps a more scientific implementations of the augmentations used in the experiments. Additionally, integrating these augmentations during the pre-training phase may lead to better-aligned representations that capture genre-specific features more effectively. Finally, improving the model's robustness to real-world variations, such as background noise, recording inconsistencies, and live performances, could significantly enhance its practical utility and reliability in diverse audio environments.

References

[1] Al-Tahan, H., and Mohsenzadeh, Y. CLAR: Contrastive Learning of Auditory Representations. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics* (Virtual Conference, April 2021), vol. 130 of *Proceedings of Machine Learning Research*, PMLR, pp. 2530–2538. [Online].

[2] Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence 35*, 8 (2013), 1798–1828.

[3] Castellon, R., Donahue, C., and Liang, P. Codified audio language modeling learns useful representations for music information retrieval. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)* (2021), pp. 88–96.

[4] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)* (2020).

[5] Choi, K., Fazekas, G., and Sandler, M. Transfer learning for music classification and regression tasks. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)* (2017).

[6] Defferrard, M., Benzi, K., Vandergheynst, P., and Bresson, X. FMA: A dataset for music analysis. In *18th International Society for Music Information Retrieval Conference (ISMIR)* (2017).

[7] Defferrard, M., Benzi, K., Vandergheynst, P., and Bresson, X. Fma: A dataset for music analysis. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)* (2017).

[8] Gjerdingen, R. O., and Perrott, D. Scanning the dial: The rapid recognition of music genres. *Journal of New Music Research 37*, 2 (2008), 93–100.

[9] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., et al. Cnn architectures for large-scale audio classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2017).

[10] Jing, L., and Tian, Y. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering 35*, 1 (2023), 1–20.

[11] Kingma, D. P., and Welling, M. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)* (2014).

[12] Koops, H. V., de Haas, W. B., Burgoyne, J. A., Bransen, J., Kent-Muller, A., and Volk, A. Annotator subjectivity in harmony annotations of popular music. *Journal of New Music Research 48*, 3 (2019), 232–252.

[13] Law, E., von Ahn, L., Dannenberg, R. B., and Crawford, M. Evaluation of algorithms using games: The case of music tagging. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)* (2009).

[14] Lee, J., Park, J., Kim, K. L., and Nam, J. Samplecnn: End-to-end deep convolutional neural networks using very small filters for music classification. *Applied Sciences 8*, 1 (2018), 150.

[15] Liu, C., Feng, L., Liu, G., Wang, H., and Liu, S. Bottom-up broadcast neural network for music genre classification. *Multimedia Tools and Applications 80*, 5 (2021), 7313–7331.

[16] Logan, B. Mel frequency cepstral coefficients for music modeling. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)* (2000).

[17] Madsen III, A. R. Dynamic range compression and its effect on music genre classification. In *Proceedings of the 2024 International Conference on Audio, Language, and Image Processing (ICALIP)* (2024), IEEE, pp. 253–258.

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

[19] Qiu, L., Li, S., and Sung, Y. Dbtmpe: Deep bidirectional transformers-based masked predictive encoder approach for music genre classification. *Mathematics 9*, 5 (2021), 530.

[20] Saeed, A., Grangier, D., and Zeghidour, N. Contrastive learning of general-purpose audio representations. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2021), IEEE, pp. 3875–3879.

[21] Spijkervet, J., and Burgoyne, J. A. Contrastive learning of musical representations. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)* (2021), pp. 673–681.

[22] Stevens, S. S., Volkmann, J., and Newman, E. B. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America 8*, 3 (1937), 185–190.

[23] Sturm, B. L. The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use. *Journal of New Music Research 43*, 3 (2014), 147–172.

[24] van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)* (2018).

[25] Won, M., Spijkervet, J., and Choi, K. Pytorch tutorial — music classification: Beyond supervised learning. https://music-classification.github.io/tutorial/part5_beyond/self-supervised-learning.html#simclr, 2021. Accessed: 2024-11-26.

[26] Zhang, L.-C., and et al. End-to-end learning for music audio classification: From raw waveforms to features and classifiers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 24*, 9 (2016), 1522–1535.

[27] Zhao, X., et al. Contrastive learning for music genre classification using unlabeled data. *IEEE Transactions on Multimedia 24* (2022), 219–229.