Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MERT-SpecAMGCNet: Attention-Modulated Gated Convolutional Network for Robust Music Genre Classification with Cross-Dataset Generalization

Agrima Malviya

230304002@HBTU.AC.IN

Harcourt Butler Technical University

Rashi Agarwal

Harcourt Butler Technical University

Research Article

Keywords: music genre classification, cross-dataset transfer, music foundation models, MERT, frequency-aware spectral modeling, gated convolution, attention fusion, GTZAN, FMA

Posted Date: August 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10335887/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

MERT-SpecAMGCNet: Attention-Modulated Gated Convolutional Network for Robust Music Genre Classification with Cross-Dataset Generalization

Agrima Malviya[1*] and Rashi Agarwal[1]

[1*]Department of Computer Science and Engineering, Harcourt Butler Technical University, Kanpur, Uttar Pradesh, India.

*Corresponding author(s). E-mail(s): 230304002@hbtu.ac.in;
Contributing authors: rashi@hbtu.ac.in;

Abstract

Music genre classification models generally achieve good performance on their own training sets, but fail to perform well on sets of varying quality, production style, artist composition, genre taxonomy and annotation practice. A dual-branch architecture is proposed in this paper, MERT-SpecAMGCNet, that combines the pretrained MERT waveform representations and a frequency-aware log-Mel spectral branch. The spectral branch uses multi-scale gated convolutions, channel recalibration and low-, mid-, high-frequency region fusion to capture local time–frequency evidence while gated cross-representation attention selectively fuses the spectral tokens with the temporal embeddings from MERT. This fused representation is further enhanced by temporal modelling using Conformer and multi-head attentive statistics pooling. The model is evaluated on the GTZAN 10-class classification, GTZAN-to-FMA Shared-3 zero-shot transfer, full target-domain fine-tuning (Shared-3) and FMA Small target-domain fine-tuning (8-class supervised classification). MERT-SpecAMGCNet achieved **89.60 ± 0.55%** accuracy and **89.90 ± 0.60%** macro-F1 on GTZAN, **65.00 ± 0.71%** accuracy and **62.77 ± 0.82%** macro-F1 in zero-shot GTZAN-to-FMA Shared-3 transfer, **89.00 ± 0.71%** accuracy and **88.75 ± 0.74%** macro-F1 after full FMA Shared-3 fine-tuning, and **81.60 ± 0.76%** accuracy and **81.44 ± 0.84%** macro-F1 on FMA Small. Baseline and ablation studies demonstrate that each of the four components (frequency-aware spectral modelling, gated fusion, temporal refinement, attentive pooling) makes a contribution to the performance. The results

show that explicit spectral evidence can be applied along with the pre-trained music embeddings for the transfer learning-based music genre classification.

Keywords: music genre classification, cross-dataset transfer, music foundation models, MERT, frequency-aware spectral modeling, gated convolution, attention fusion, GTZAN, FMA

1 Introduction

Music genre classification is a fundamental task in music information retrieval, supporting recommendation, playlist generation, search, catalog organization, and large-scale music analysis. Early approaches relied on handcrafted descriptors such as timbral texture, rhythm, pitch, and statistical audio features [1, 2]. With the growth of labeled datasets and deep learning, research has shifted toward learned representations from waveforms, spectrograms, and other time–frequency inputs [3–5].

Time–frequency representations remain important because they capture local information related to instrumentation, rhythm, bass activity, harmonic structure, vocal characteristics, and percussion. CNNs, recurrent networks, attention mechanisms, and CNN–Transformer hybrids have been widely applied for spectrogram-based genre classification [3, 6–8]. More recent studies have explored gated convolutions, channel attention, visual spectrogram learning, and data augmentation strategies [9–12].

However, strong performance on a single benchmark does not necessarily imply robust generalization. Music datasets differ in artist composition, recording quality, production style, class balance, genre definitions, and annotation practices. Consequently, genre labels reflect not only acoustic content but also editorial and cultural conventions [13–15]. While GTZAN remains a popular benchmark [1], the Free Music Archive (FMA) provides a larger and more diverse collection suitable for studying cross-dataset generalization [16].

Pretrained music representation learning offers a promising solution to dataset shift. Recent self-supervised and foundation-model approaches learn contextual audio representations from large-scale music corpora [17–19]. Among them, MERT produces frame-level embeddings that capture timbre, rhythm, melody, instrumentation, and temporal structure [18]. Nevertheless, these embeddings are not explicitly optimized to preserve localized genre-transfer cues such as bass patterns, percussion transients, vocal brightness, and harmonic texture.

To address this limitation, we propose *MERT-SpecAMGCNet*, a dual-branch architecture that combines pretrained MERT waveform embeddings with frequency-aware log-Mel spectral modeling. The spectral branch employs multi-scale gated convolutions, channel recalibration, and frequency-region fusion, while gated cross-representation attention enables temporal MERT tokens to selectively retrieve spectral evidence through a learnable residual gate. Unlike simple concatenation or ungated cross-attention, this design preserves pretrained representations while incorporating only the most informative spectral cues for genre discrimination under dataset shift.

The proposed model is evaluated on GTZAN 10-class classification, GTZAN-to-FMA Shared-3 zero-shot transfer, FMA Shared-3 full fine-tuning, and FMA Small eight-class supervised classification. The Shared-3 setting aligns the Hip-Hop, Pop, and Rock classes between GTZAN and FMA, enabling controlled cross-dataset evaluation.

The main contributions of this work are:

1. We propose MERT-SpecAMGCNet, a dual-branch architecture combining pretrained MERT embeddings with frequency-aware log-Mel spectral modeling.
2. We introduce gated cross-representation fusion using MERT temporal tokens and a learnable residual gate for selective spectral integration.
3. We evaluate the model on GTZAN-to-FMA Shared-3 transfer in both zero-shot and full fine-tuning settings.
4. We compare the proposed approach against spectral-only, MERT-only, concatenation-based, ungated cross-attention, and component-removal baselines.
5. We analyze the contributions of frequency-aware spectral modeling, gated fusion, Conformer refinement, and attentive statistics pooling through accuracy, macro-F1, confusion matrices, and ablation studies.

The remainder of this paper is organized as follows. Section 2 reviews related work, Section 3 presents the proposed method, Section 4 describes the experimental protocol, Section 5 reports the results and ablations, Section 6 discusses the findings and limitations, and Section 7 concludes the paper.

2 Related Work

Music genre classification is a fundamental task in music information retrieval (MIR). Early approaches relied on handcrafted timbral, rhythmic, pitch-based, and MFCC features combined with conventional machine-learning classifiers [1, 2, 20]. More recently, research has shifted toward deep representation learning from waveforms and time–frequency representations [4, 5, 21].

Spectrogram-based deep learning has become particularly successful because spectrograms capture genre-discriminative cues related to timbre, rhythm, harmony, instrumentation, and vocal characteristics. CNN-based models using spectrogram and Mel-spectrogram inputs have achieved strong performance across diverse genre-recognition tasks [3, 9, 22–25]. However, most spectrogram-based systems are evaluated on a single dataset, leaving cross-dataset generalization insufficiently explored.

To improve representation learning, researchers have proposed attention mechanisms, recurrent networks, CNN–RNN hybrids, wavelet-based models, Transformers, gated CNNs, channel-aware networks, and capsule-style architectures [6–8, 10, 11, 26–29]. Additional advances include optimization-aware architectures, data augmentation, and supervised contrastive learning [12, 30–32]. Despite these improvements, architectural advances alone do not guarantee robustness under dataset shift.

Pretrained and self-supervised audio representation learning has therefore gained increasing attention. Models such as BYOL-A, MERT, MusicFM, and other music foundation models learn transferable representations from large-scale pretraining [17–19]. Recent evaluations of MERT, MusicFM, MuQ, MULE, MusiCNN, Jukebox, and

EncodecMAE further demonstrate their value for transfer-oriented MIR tasks [33]. However, these embeddings may not explicitly preserve localized frequency-region cues that remain important for genre discrimination.

Cross-dataset evaluation is essential because genre recognition is influenced not only by audio content but also by annotation practices, cultural context, and dataset construction [13]. Although GTZAN remains widely used [1], performance on a single benchmark is insufficient evidence of robust genre recognition [14]. The Free Music Archive (FMA) provides a larger and more diverse benchmark for studying transfer and generalization [15, 16].

Another promising direction is representation fusion, where complementary feature streams are combined through attention, multi-view learning, or feature-level fusion [34–37]. However, simple fusion strategies often assume equal reliability across representations, which may be problematic under dataset shift.

Motivated by these challenges, we propose *MERT-SpecAMGCNet*, which combines pretrained MERT waveform embeddings with a frequency-aware log-Mel spectral branch. Multi-scale gated convolutions and channel recalibration capture localized spectral information, while gated cross-representation attention selectively integrates spectral evidence with contextual MERT embeddings. The framework is designed for robust GTZAN-to-FMA genre transfer and is evaluated in both zero-shot and target-domain fine-tuning settings.

3 Proposed Method

3.1 Overview

MERT-SpecAMGCNet, MERT-based Spectral Attention-Modulated Gated Convolutional Network, is a proposed dual-branch architecture for music genre classification. It integrates pretrained MERT waveform representations with frequency-aware log-Mel spectral modeling to jointly capture contextual musical information and localized time–frequency cues. The MERT branch provides contextual music embeddings learned through large-scale self-supervised pretraining [17–19], while the spectral branch captures localized time–frequency cues such as timbre, rhythm, bass energy, harmonic texture, percussion transients, and production style [3, 9, 22]. Both branches are projected into a shared token space and fused using gated cross-representation attention, followed by temporal refinement, attentive statistics pooling, and classification.

3.2 Input Representation

Each track is converted to mono, resampled to 24 kHz, and processed using five-second crops. During training, one random crop is sampled per track at each epoch. During validation and testing, deterministic evenly spaced crops are used, and crop-level logits are averaged for track-level prediction. For the spectral branch, the input is a log-Mel spectrogram

$$S = \log(\text{MelSpectrogram}(x) + \epsilon),$$

computed using 128 Mel bins, a 2048-point FFT, a hop length of 512 samples, and a frequency range of 20 Hz to 12 kHz.

3.3 MERT Representation Branch

The waveform branch uses pretrained MERT to extract frame-level contextual music representations [18]. Instead of using only the final hidden state, the model learns a weighted combination of the final four MERT layers. The resulting representation is normalized and projected to a shared token dimension $d = 320$, producing the MERT token sequence $T_M \in \mathbb{R}^{N_M \times d}$. This allows the model to use both contextual musical structure and relatively local acoustic information from the upper MERT layers.

3.4 Frequency-Aware Spectral Branch

The spectral branch processes the log-Mel spectrogram through a convolutional stem followed by three multi-scale attention-modulated gated convolutional stages with 128, 192, and 256 output channels. Each stage contains gated convolutional blocks with 3×3, 5×5, and 7×7 kernels, where feature and sigmoid gate pathways are combined through residual connections, normalization, activation, and dropout. This enables the extraction of genre-relevant patterns at multiple receptive-field scales [8, 10, 11].

Channel recalibration is applied after each stage using global average pooling and a bottleneck transformation. The final feature map is divided into low-, mid-, and high-frequency regions, which are processed independently, pooled, concatenated, and projected into the shared token space to produce spectral tokens $T_S \in \mathbb{R}^{N_S \times d}$. This preserves complementary frequency-specific cues related to bass activity, harmonic texture, percussion, brightness, and vocal timbre.

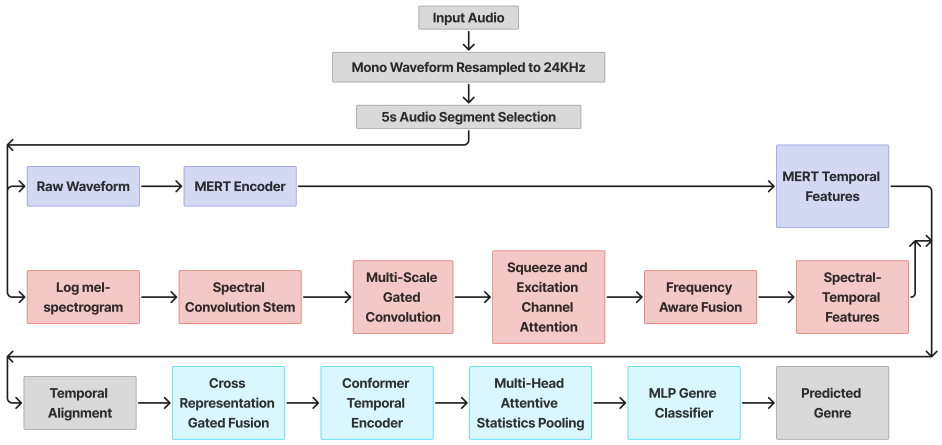


Fig. 1 Overview of the proposed MERT-SpecAMGCNet architecture.

3.5 Gated Cross-Representation Fusion

Because the MERT and spectral branches may produce different temporal lengths, the spectral token sequence is aligned to the MERT sequence using linear interpolation, yielding $\widetilde{T}_S \in \mathbb{R}^{N_M \times d}$. MERT tokens are used as queries and aligned spectral tokens as keys and values to compute $C = \text{MultiHeadCrossAttention}(Q = T_M, K = \widetilde{T}_S, V = \widetilde{T}_S)$. A residual gate $g = \sigma(W_g[T_M; C] + b_g)$ then regulates the injected spectral information, producing the fused representation $T_F = \text{LayerNorm}(T_M + g \odot C)$. This mechanism preserves pretrained MERT representations while selectively incorporating genre-relevant spectral evidence.

3.6 Temporal Refinement, Pooling, and Classification

The fused sequence is combined with sinusoidal positional encoding and passed through two Conformer-style temporal refinement blocks, which combine self-attention and temporal convolution. The refined sequence is summarized using multi-head attentive statistics pooling, where each head computes an attention-weighted temporal mean and standard deviation. The pooled representation is passed to an MLP classifier followed by a softmax layer.

The number of output classes is $K = 10$ for GTZAN, $K = 3$ for GTZAN-to-FMA Shared-3 transfer, and $K = 8$ for FMA Small supervised classification. The model is trained using multiclass cross-entropy loss. For full target-domain fine-tuning, label smoothing with factor $\lambda = 0.02$ is used.

3.7 Training Strategy

Training is performed in two stages. In Stage A, the pretrained MERT backbone is frozen while the layer-fusion weights, projection layer, spectral branch, fusion module, temporal refinement module, pooling module, and classifier are trained. In Stage B, the best Stage A checkpoint is used for partial fine-tuning, where only the final four MERT transformer layers are unfrozen and optimized using a smaller learning rate. For FMA Shared-3 fine-tuning, the model is initialized from the selected GTZAN source-domain checkpoint and trained using the same staged strategy.

Table 1 Configuration of MERT-SpecAMGCNet.

Parameter	**Value**	**Parameter**	**Value**
Sampling rate	24,000 Hz	Audio duration	5 s
Mel bins	128	FFT size	2,048
Hop length	512 samples	Frequency range	20–12,000 Hz
MERT checkpoint	m-a-p/MERT-v1-330M	MERT layers fused	Final 4 layers
MERT layers unfrozen	Final 4 layers	Shared token dimension	320
Spectral stem channels	64, 96	Spectral stage channels	128, 192, 256
Gated kernels	3×3, 5×5, 7×7	Cross-attention heads	5
Conformer-style blocks	2	Pooling heads	4
Classifier hidden dimension	512	Dropout	0.20

4 Experimental Protocol

4.1 Datasets and Tasks

The experiments use GTZAN and FMA Small to evaluate source-domain classification, cross-dataset transfer, and target-domain fine-tuning [1, 16]. GTZAN serves as the source-domain benchmark, while FMA Small is an independent target-domain collection. Three tasks are evaluated: GTZAN 10-class classification, GTZAN-to-FMA Shared-3 transfer, and FMA Small eight-class supervised classification.

For GTZAN 10-class classification, the split includes 800 training, 100 validation, and 100 test tracks. Because GTZAN contains known limitations such as possible duplicate recordings, recording artifacts, and artist overlap, it is used primarily as a source-domain diagnostic benchmark rather than standalone evidence of real-world robustness. Robustness is instead assessed mainly through GTZAN-to-FMA Shared-3 transfer, where source-trained models are evaluated on an independent target-domain dataset.

The cross-dataset experiment uses a Shared-3 taxonomy of Hip-Hop, Pop, and Rock, the common top-level classes in both datasets. The GTZAN Shared-3 source split contains 240 training, 30 validation, and 30 test tracks. The FMA Shared-3 target split contains 1,920 training, 240 validation, and 300 held-out test tracks. The FMA Small eight-class task uses the official top-level classes Electronic, Experimental, Folk, Hip-Hop, Instrumental, International, Pop, and Rock, with 6,400 training, 800 validation, and 800 test tracks.

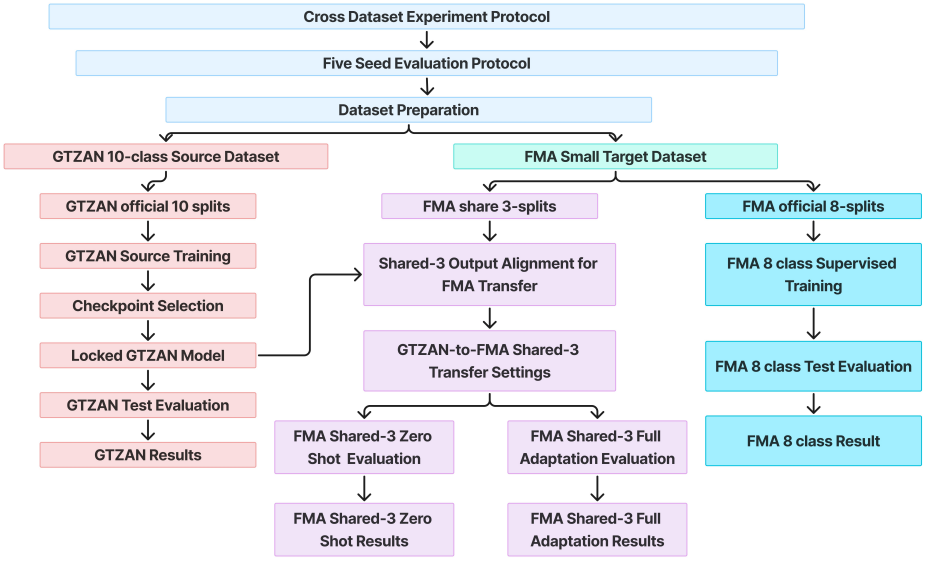


Fig. 2 Experimental protocol for GTZAN source-domain classification.

4.2 Shared-3 Transfer Setting

The Shared-3 transfer setting aligns the overlapping GTZAN and FMA top-level classes Hip-Hop, Pop, and Rock. All other classes are excluded from this protocol. The aligned labels are shown in Table 2.

Table 2 Shared-3 label alignment between GTZAN and FMA Small.

Label	GTZAN class	FMA class
0	Hip-Hop	Hip-Hop
1	Pop	Pop
2	Rock	Rock

Two transfer settings are evaluated. In the zero-shot setting, the model is trained only on GTZAN Shared-3 and evaluated directly on the FMA Shared-3 test set. No FMA labels are used for training, validation, calibration, checkpoint selection, or hyperparameter tuning. In the full fine-tuning setting, the selected GTZAN source checkpoint is fine-tuned using all 1,920 labelled FMA Shared-3 training tracks.

4.3 Preprocessing and Augmentation

All audio is converted to mono, resampled to 24 kHz, and processed using five-second crops. During training, one random crop is sampled per track at each epoch. During validation and testing, deterministic evenly spaced crops are used, and final track-level predictions are obtained by averaging logits over five test crops. In the zero-shot GTZAN-to-FMA setting, normalization statistics are estimated only from the GTZAN source training set.

Training-time augmentation includes random gain perturbation, additive Gaussian noise, frequency masking, time masking, and mixup [12, 38, 39]. Random gain is applied with probability 0.35 using a gain range of 0.85 to 1.15, Gaussian noise with probability 0.25, and mixup with probability 0.25 using a Beta distribution with $\alpha = 0.20$. Pitch shifting and time stretching are not used.

4.4 Baselines and Implementation

The proposed model is compared with spectral-only, MERT-only, and fusion-based baselines: Scratch AMGCNet, MERT-MLP Probe, MERT Temporal Adapter, Spectral AMGC Branch, MERT + Spectral Concat, MERT + Cross-Attention, and MERT-SpecAMGCNet. Ablations further examine the effect of frequency-aware fusion, Conformer-style temporal refinement, and attentive statistics pooling.

All models are implemented in PyTorch and optimized using AdamW with cosine learning-rate decay, linear warm-up, gradient clipping, dropout, and exponential moving average checkpointing [40, 41]. Source-domain training is performed for up to 120 epochs with early stopping patience of 35 epochs. Target-domain fine-tuning is performed for up to 80 epochs with early stopping patience of 35 epoch.

Table 3 Training configuration.

Setting	Value	Setting	Value
Optimizer	AdamW	Batch size	8
Source max epochs	120	Target max epochs	80
Source patience	35 epochs	Target patience	35 epochs
Source learning rate	3×10^{-4}	Target learning rate	1×10^{-4}
Warm-up epochs	8 source / 4 target	Weight decay	2×10^{-4}
Gradient clipping	1.0	Dropout	0.20
EMA decay	0.998	Target label smoothing	0.02
Validation crops	3 per track	Test crops	5 per track

4.5 Evaluation Metrics

Macro-F1 is the primary metric because it gives equal importance to each class. Accuracy, macro precision, macro recall, weighted F1, and confusion matrices are also reported for a fuller evaluation [42, 43]. Repeated experiments are reported as mean ± sample standard deviation across five seeds. GTZAN 10-class and FMA Small eight-class results are reported at track level. For FMA Shared-3 confusion matrices, five deterministic crops are evaluated for each of the 300 held-out test tracks, producing 1,500 crop-level decisions.

5 Results

5.1 Reporting Convention

Unless stated otherwise, results are reported as mean ± sample standard deviation across five random seeds. Macro-F1 is used as the primary metric because it gives equal importance to each class, while accuracy, macro precision, macro recall, weighted F1, and confusion matrices provide complementary views of model behavior.

The evaluation addresses three questions: whether MERT-SpecAMGCNet learns strong source-domain genre representations on GTZAN, whether these representations transfer to FMA Shared-3 without target-domain labels, and how much performance is recovered after full target-domain fine-tuning. GTZAN 10-class and FMA Small eight-class results are reported at track level. For FMA Shared-3 confusion matrices, five deterministic crops are evaluated for each of the 300 held-out test tracks, producing 1,500 crop-level decisions.

Table 4 Main classification and transfer results of MERT-SpecAMGCNet.

Setting	Acuracy	Macro Prec.	Macro Rec.	Macro-F1	Weighted F1
GTZAN-10	89.60 ± 0.55	91.65 ± 0.93	89.60 ± 0.55	89.90 ± 0.60	89.90 ± 0.60
FMA-3 ZS	65.00 ± 0.71	65.55 ± 0.88	65.00 ± 0.71	62.77 ± 0.82	62.77 ± 0.80
FMA-3 FT	89.00 ± 0.71	89.69 ± 0.78	89.00 ± 0.71	88.75 ± 0.74	88.75 ± 0.72
FMA-8	81.60 ± 0.76	82.10 ± 0.94	81.60 ± 0.76	81.44 ± 0.84	81.44 ± 0.84

5.2 GTZAN 10-Class Source-Domain Classification

MERT-SpecAMGCNet achieved $89.60 \pm 0.55\%$ accuracy and $89.90 \pm 0.60\%$ macro-F1 on GTZAN 10-class classification. This confirms that the proposed dual-branch architecture can learn strong genre-discriminative representations in the source domain. However, this result is treated as a diagnostic benchmark rather than as evidence of cross-dataset robustness, since strong GTZAN performance alone does not guarantee reliable transfer to a different music collection.

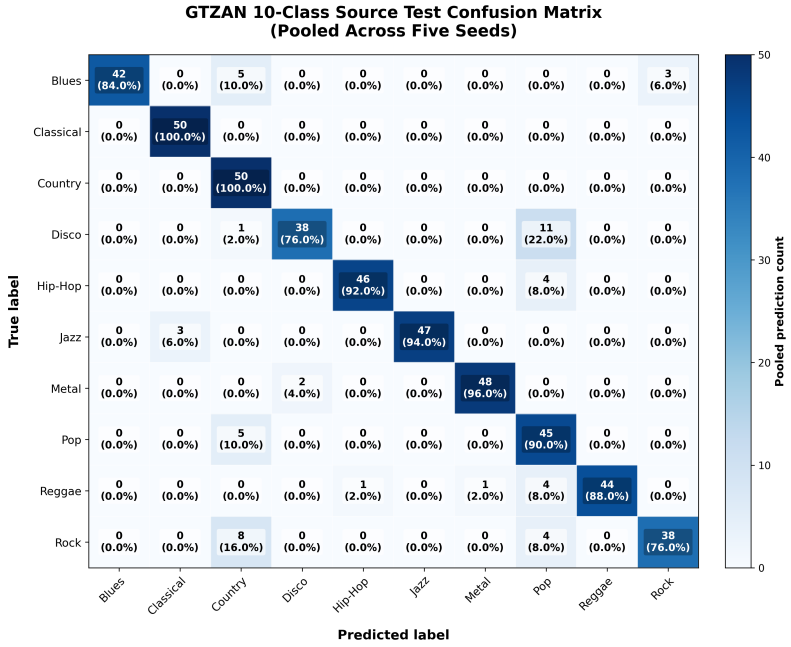
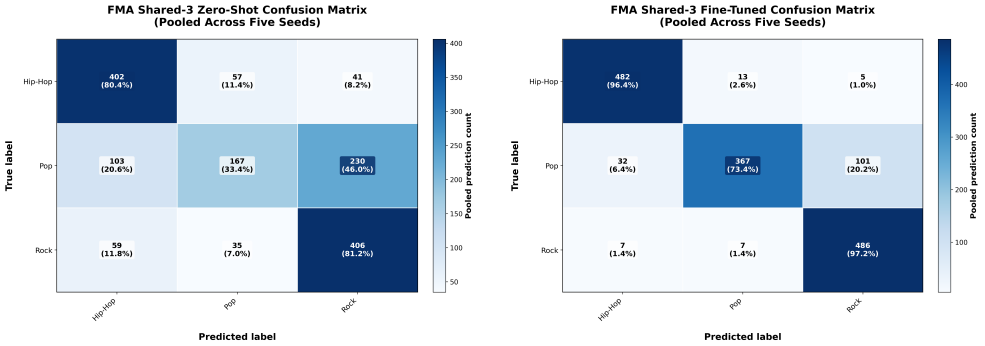


Fig. 3 Pooled five-seed confusion matrix for GTZAN 10-class classification. The matrix contains 500 pooled test predictions and 448 correct predictions, corresponding to 89.60% accuracy.

5.3 GTZAN-to-FMA Shared-3 Transfer

The GTZAN-to-FMA Shared-3 setting is the primary cross-dataset evaluation. In the zero-shot setting, the model is trained only on GTZAN Shared-3 and evaluated directly on FMA Shared-3. No FMA labels are used for training, validation, calibration, checkpoint selection, or hyperparameter tuning. Under this setting, MERT-SpecAMGCNet achieved $65.00 \pm 0.71\%$ accuracy and $62.77 \pm 0.82\%$ macro-F1.

This zero-shot result shows that the source-trained representation retains transferable genre information, but it also exposes the difficulty of cross-dataset generalization. After full FMA Shared-3 fine-tuning, performance increased to $89.00 \pm 0.71\%$ accuracy and $88.75 \pm 0.74\%$ macro-F1. The improvement of 24.00 percentage points in accuracy and 25.98 percentage points in macro-F1 demonstrates that the architecture adapts effectively when labelled target-domain data are available.



(a) Zero-shot transfer **(b) Full fine-tuning**

Fig. 4 FMA Shared-3 confusion matrices for zero-shot GTZAN-to-FMA transfer and full target-domain fine-tuning; the repeated-seed numerical results are reported in Table 4.

5.4 FMA Small Eight-Class Supervised Classification

On the broader FMA Small eight-class task, MERT-SpecAMGCNet achieved $81.60 \pm 0.76\%$ accuracy and $81.44 \pm 0.84\%$ macro-F1. This task is not directly comparable to the three-class Shared-3 transfer setting because it contains a broader taxonomy. Its purpose is to test whether the proposed architecture remains effective beyond the restricted Hip-Hop, Pop, and Rock transfer setup.

5.5 Baseline and Ablation Comparison

Table 5 compares MERT-SpecAMGCNet with MERT-only, spectral-only, fusion-based, and component-removal variants across all evaluation settings. The results show a consistent progression from single-branch models to fusion-based models and then to the complete architecture. MERT-only variants provide strong pretrained representations, while the spectral branch contributes complementary frequency-local evidence. Simple concatenation improves over single-branch variants, but attention-based fusion performs better by allowing interaction between MERT and spectral tokens.

The full MERT-SpecAMGCNet achieves the highest Macro-F1 across GTZAN 10-class classification, FMA Shared-3 zero-shot transfer, FMA Shared-3 full fine-tuning, and FMA Small eight-class classification. The component-removal variants further show that frequency-aware fusion, Conformer-style temporal refinement, and attentive statistics pooling each contribute to the final performance.

5.6 Training and Validation Behavior

The training and validation curves provide additional evidence about optimization behavior. The GTZAN curves show stable convergence during source-domain training. The FMA Shared-3 curves show the effect of adapting the source-trained model to the target-domain taxonomy, while the FMA Small eight-class curves show that the model remains trainable under a broader genre set. Validation monitoring is used for checkpoint selection to reduce overfitting during both source-domain training and target-domain fine-tuning.

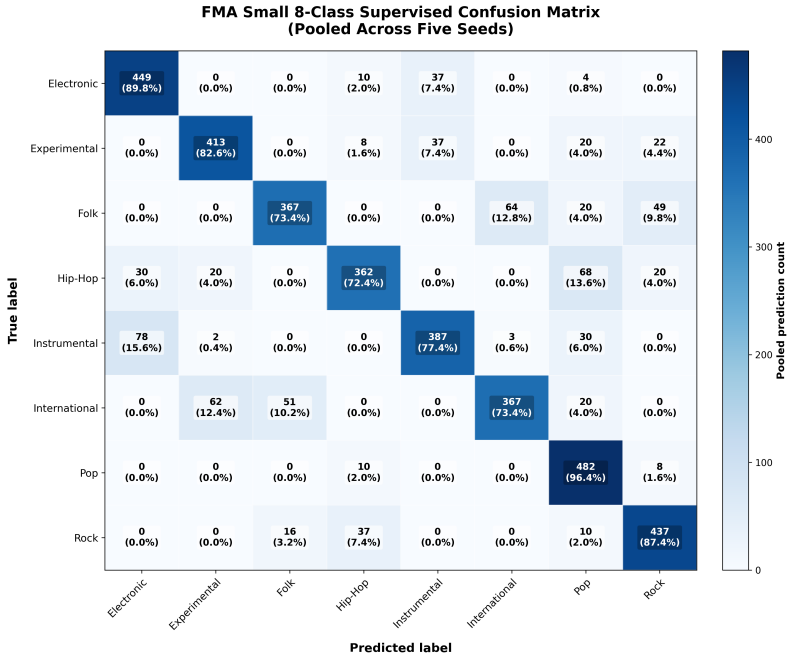


Fig. 5 Pooled five-seed confusion matrix for FMA Small eight-class supervised classification. The matrix contains 4000 pooled test predictions and 3264 correct predictions, corresponding to 81.60% accuracy.

Table 5 Baseline and ablation comparison across evaluation settings. Values are Macro-F1 percentages reported as mean ± standard deviation across five seeds. (FMA-3 ZS is FMA 3 class zero shot and FMA-3 FT is FMA 3 class fine tuned)

Model / Variant	GTZAN 10	FMA-3 ZS	FMA-3 FT	FMA-8
MERT-MLP Probe	83.35 ± 0.82	57.45 ± 0.95	82.35 ± 0.86	75.80 ± 1.02
MERT Temporal Adapter	84.60 ± 0.79	58.70 ± 0.92	83.65 ± 0.82	76.95 ± 0.98
Spectral AMGC Branch only	82.40 ± 0.88	56.80 ± 0.98	81.50 ± 0.90	74.60 ± 1.08
MERT + Spectral Concatenation	86.20 ± 0.74	60.05 ± 0.88	85.40 ± 0.78	78.45 ± 0.94
MERT + Cross-Attention	87.10 ± 0.71	60.95 ± 0.86	86.35 ± 0.76	79.20 ± 0.90
Without Frequency-Aware Fusion	87.85 ± 0.69	61.20 ± 0.85	87.05 ± 0.75	79.85 ± 0.88
Without Conformer Refinement	88.20 ± 0.68	61.55 ± 0.84	87.35 ± 0.74	80.10 ± 0.86
Mean Pooling	87.60 ± 0.72	60.75 ± 0.87	86.70 ± 0.77	79.50 ± 0.91
Full MERT-SpecAMGCNet	$\mathbf{89.90 \pm 0.60}$	$\mathbf{62.77 \pm 0.82}$	$\mathbf{88.75 \pm 0.74}$	$\mathbf{81.44 \pm 0.84}$

The results support three main findings. First, MERT-SpecAMGCNet learns strong source-domain representations on GTZAN. Second, the model transfers non-trivially to FMA Shared-3 in the zero-shot setting, showing that the learned representation retains cross-dataset genre information. Third, full target-domain fine-tuning substantially improves performance, indicating that the proposed architecture is adaptable when labelled FMA data are available. The FMA Small eight-class result

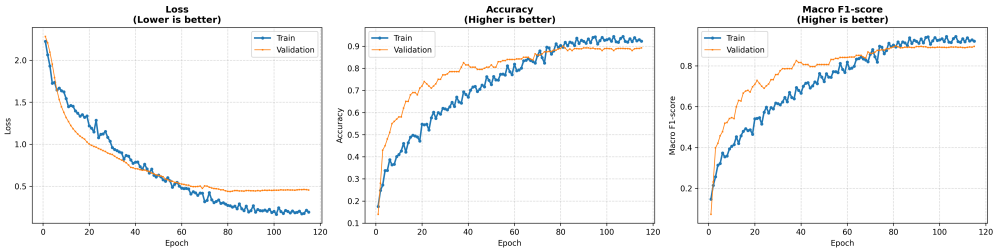


Fig. 6 Training and validation curves for GTZAN 10-class source-domain classification.

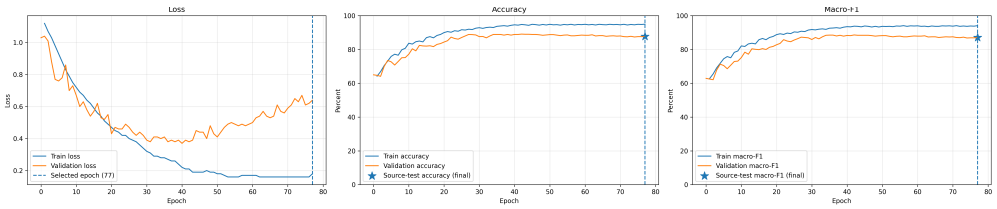


Fig. 7 Training and validation curves for full FMA Shared-3 target-domain fine-tuning.

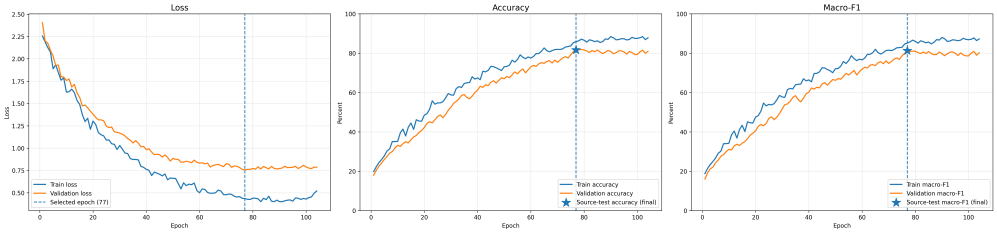


Fig. 8 Training and validation curves for the FMA Small eight-class supervised classification task.

further shows that the architecture remains effective under a broader target-domain taxonomy.

6 Discussion

The results show that MERT-SpecAMGCNet is effective as both a source-domain classifier and a transfer-oriented architecture. Its strong GTZAN performance confirms that combining pretrained MERT representations with frequency-aware spectral modeling produces discriminative genre representations. However, the more important result is the GTZAN-to-FMA Shared-3 evaluation, where the model is tested under dataset shift rather than only within a familiar benchmark.

In zero-shot GTZAN-to-FMA transfer, the model achieved $62.77 \pm 0.82\%$ macro-F1 without using labelled FMA data. This shows that the source-trained representation retains meaningful genre information across datasets. At the same time, the gap

between GTZAN performance and zero-shot FMA performance highlights the difficulty of cross-dataset genre classification. Genre labels are shaped not only by acoustic content, but also by production style, dataset construction, annotation practice, and cultural or editorial conventions.

Full FMA Shared-3 fine-tuning increased macro-F1 to $88.75 \pm 0.74\%$, showing that the proposed representation is highly adaptable when target-domain labels are available. This suggests that pretrained music embeddings provide a strong foundation, but target-domain supervision remains important when the deployment collection differs from the source dataset.

The baseline and ablation results support the architectural design. MERT-only variants confirm the value of pretrained music representations, while the spectral branch shows that localized time–frequency cues remain important for genre discrimination. Fusion-based variants improve over single-branch models, and the full MERT-SpecAMGCNet performs best across all settings. This indicates that the gains come from the combined effect of frequency-aware spectral modeling, gated cross-representation fusion, temporal refinement, and attentive statistics pooling.

The gated fusion mechanism is central to this behavior. Simple concatenation treats all feature streams as equally reliable, while the proposed residual gate controls how much spectral evidence is injected into the MERT representation. This is useful under dataset shift because some spectral cues reflect stable genre information, such as bass energy, harmonic texture, and percussion activity, while others may be dataset-specific.

The FMA Small eight-class result further shows that the model remains effective beyond the restricted three-class transfer setting. Nevertheless, the study is limited to GTZAN and FMA, and the main transfer experiment uses only the Shared-3 taxonomy. Future work should evaluate broader genre taxonomies, additional datasets, intermediate few-shot settings, and more efficient variants of the proposed architecture.

Overall, the findings support the central claim of this paper: robust music genre classification benefits from combining contextual pretrained music representations with explicit frequency-aware spectral evidence through selective gated fusion.

7 Conclusion

In this paper, a dual-branch architecture for music genre classification, called MERT-SpecAMGCNet, was introduced, which consists of a pretrained MERT branch and a frequency-aware log-Mel spectral modeling branch. The model selectively fuses the contextual music embeddings and the local spectral evidence with gated cross-representation fusion.

The experiments show that MERT-SpecAMGCNet performs well on GTZAN, a transfer task with a meaningful target domain FMA Shared-3 in the zero-shot setting, and a good performance after full target-domain fine-tuning and on the larger task FMA Small (8 classes). The results support the primary hypothesis that music representations based on a large dataset and explicit spectral features are complementary for effective genre classification in the presence of dataset shift.

In general, the results indicate that the pretrained music embeddings and frequency-aware spectral modeling models complementary aspects of genre information. Local information, such as bass energy, harmonic texture, percussion transients and vocal brightness, are maintained in the spectral branch, and the contextual musical structure is given by MERT. The gated fusion mechanism is crucial to enable the selective fusion of different sources of evidence when models are transferred between datasets that have different recording conditions, production styles and conventions for annotations.

Other music datasets, extended genre taxonomies and other cross-collection transfer settings will be explored in the future. Other issues that need to be explored further are track-level error pattern, computational efficiency, model compression and alternative fusion mechanism for robust music information retrieval.

Declarations

Funding

The authors received no specific funding for this work.

Conflict of interest

The authors declare that they have no conflict of interest.

Ethics approval and consent to participate

Not applicable. This study does not involve human participants, animals, or identifiable personal data.

Consent for publication

Not applicable.

Data availability

The datasets used in this study are publicly available. The GTZAN dataset and the Free Music Archive (FMA) dataset were obtained from their respective public repositories. No new datasets were generated during this study.

Materials availability

Not applicable.

Code availability

The code used in this study is available from the corresponding author upon request.

Author contributions

Agrima Malviya: Conceptualization, methodology, software, data curation, formal analysis, investigation, validation, visualization, writing—original draft preparation, and writing—review and editing.

Rashi Agarwal: Supervision, methodology review, validation, project administration, writing—review and editing, and overall research guidance.

All authors read and approved the final manuscript.

References

[1] Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. IEEE Transactions on speech and audio processing **10**(5), 293–302 (2002)

[2] Li, T., Ogihara, M., Li, Q.: A comparative study on content-based music genre classification. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 282–289 (2003)

[3] Costa, Y.M., Oliveira, L.S., Silla Jr, C.N.: An evaluation of convolutional neural networks for music classification using spectrograms. Applied soft computing **52**, 28–38 (2017)

[4] Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.-Y., Sainath, T.: Deep learning for audio signal processing. IEEE Journal of Selected Topics in Signal Processing **13**(2), 206–219 (2019)

[5] Ramírez, J., Flores, M.J.: Machine learning for music genre: multifaceted review and experimentation with audioset. Journal of Intelligent Information Systems **55**(3), 469–499 (2020)

[6] Ashraf, M., Abid, F., Din, I.U., Rasheed, J., Yesiltepe, M., Yeo, S.F., Ersoy, M.T.: A hybrid cnn and rnn variant model for music classification. Applied Sciences **13**(3), 1476 (2023)

[7] Chen, J., Ma, X., Li, S., Ma, S., Zhang, Z., Ma, X.: A hybrid parallel computing architecture based on cnn and transformer for music genre classification. Electronics **13**(16), 3313 (2024)

[8] Xie, C., Song, H., Zhu, H., Mi, K., Li, Z., Zhang, Y., Cheng, J., Zhou, H., Li, R., Cai, H.: Music genre classification based on res-gated cnn and attention mechanism. Multimedia Tools and Applications **83**(5), 13527–13542 (2024)

[9] Cheng, Y.-H., Kuo, C.-N.: Machine learning for music genre classification using visual mel spectrum. Mathematics **10**(23), 4427 (2022)

[10] Liu, Z., Bian, T., Yang, M.: Locally activated gated neural network for automatic music genre classification. Applied Sciences **13**(8), 5010 (2023)

[11] Ding, Y., Zhang, H., Huang, W., Zhou, X., Shi, Z.: Efficient music genre recognition using ecas-cnn: A novel channel-aware neural network architecture. Sensors **24**(21), 7021 (2024)

[12] Ba, T.C., Le, T.D.T., Van, L.T.: Music genre classification using deep neural networks and data augmentation. Entertainment Computing, 100929 (2025)

[13] Law, E., Brown, M.H.: Evaluation of music information retrieval systems: Issues, datasets, and challenges. In: Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009). International Society for Music Information Retrieval (ISMIR), Kobe, Japan (2009). https://www.ismir.net

[14] Sturm, B.L.: Classification accuracy is not enough: On the evaluation of music genre recognition systems. Journal of Intelligent Information Systems **41**(3), 371–406 (2013)

[15] Christodoulou, A.-M., Lartillot, O., Jensenius, A.R.: Multimodal music datasets? challenges and future goals in music processing. International Journal of Multimedia Information Retrieval **13**(3), 37 (2024)

[16] Defferrard, M., Bresson, X., Vandergheynst, P.: Fma: A dataset for music analysis. In: Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017). International Society for Music Information Retrieval (ISMIR), Suva, Switzerland (2017). https://www.audiolabs-erlangen.de/resources/MIR/FMA

[17] Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N., Kashino, K.: Byol for audio: Exploring pre-trained general-purpose audio representations. IEEE/ACM Transactions on Audio, Speech, and Language Processing **31**, 137–151 (2022)

[18] Li, Y., Yuan, R., Zhang, G., Ma, Y., Chen, X., Yin, H., Xiao, C., Lin, C., Ragni, A., Benetos, E., Gyenge, N., Dannenberg, R., Liu, R., Chen, W., Xia, G., Shi, Y., Huang, W., Wang, Z., Guo, Y., Fu, J.: MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training (2023)

[19] Won, M., Hung, Y.-N., Le, D.: A foundation model for music informatics. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2024)

[20] Bhalke, D.G., Rajesh, B., Bormane, D.S.: Automatic genre classification using fractional fourier transform based mel frequency cepstral coefficient and timbral features. Archives of Acoustics **42**(2), 213–222 (2017)

[21] Moysis, L., Iliadis, L.A., Sotiroudis, S.P., Boursianis, A.D., Papadopoulou, M.S., Kokkinidis, K., Volos, C., Sarigiannidis, P., Goudos, S.K.: Music deep learning: deep learning methods for music signal processing—a review of the state-of-the-art. Archives of Computational Methods in Engineering **30**(7), 4163–4177 (2023)

[22] Li, J., Han, L., Li, X., Zhu, J., Yuan, B., Gou, Z.: An evaluation of deep neural network models for music classification using spectrograms. Multimedia Tools and Applications, 1–27 (2022)

[23] Wu, P., Gao, W., Chen, Y., Xu, F., Ji, Y., Tu, J., Lin, H.: An improved vit model for music genre classification based on mel spectrogram. PloS one **20**(3), 0319027 (2025)

[24] Wang, J., Yang, Y.: Rigdelet neural network and improved partial reinforcement effect optimizer for music genre classification. Scientific Reports **14**(1), 29392 (2024)

[25] Prajuli, S., Karna, A., Dhakl, O.: Music Genre Classification: A Comparative Analysis of Classical Machine Learning and Deep Learning Approaches (2026)

[26] Yu, Y., Luo, S., Liu, S., Qiao, H., Liu, Y., Feng, L.: Deep attention based music genre classification. Neurocomputing **372**, 84–91 (2020)

[27] Gan, J.: Music feature classification based on recurrent neural networks with channel attention mechanism. Mobile Information Systems **2021**(1), 7629994 (2021)

[28] Jena, K.K., Bhoi, S.K., Mohapatra, S., Bakshi, S.: A hybrid deep learning approach for classification of music genres using wavelet and spectrogram analysis. Neural Computing and Applications **35**(21), 15447–15460 (2023)

[29] Jiang, L., Yang, L., Azimi, S.: Enhanced capsule neural network with advanced triangulation topology aggregation optimizer for music genre classification. Scientific Reports **15**(1), 41 (2025)

[30] Zhang, Y., Li, T.: Music genre classification with parallel convolutional neural networks and capuchin search algorithm. Scientific Reports **15**(1), 9580 (2025)

[31] Tian, Y.: Deep neural networks and fractional grey lag goose optimization for music genre identification. Scientific Reports **15**(1), 6702 (2025)

[32] Zhuang, T., Chang, D., Wang, X.: A hybrid transformer model with supervised contrastive loss for automatic music classification. Alexandria Engineering Journal **143**, 160–169 (2026) https://doi.org/10.1016/j.aej.2026.03.046

[33] Tamm, Y.-M., Aljanaki, A.: Adopting State-of-the-Art Pretrained Audio Representations for Music Recommender Systems (2026)

[34] Wang, Q., Su, F., Wang, Y.: Hierarchical attentive deep neural networks for semantic music annotation through multiple music representations. International Journal of Multimedia Information Retrieval **9**(1), 3–16 (2020)

[35] Liu, J., Wang, C., Zha, L.: A middle-level learning feature interaction method with deep learning for multi-feature music genre classification. Electronics **10**(18), 2206 (2021)

[36] Phan, H., Le Nguyen, H., Chén, O.Y., Pham, L., Koch, P., McLoughlin, I., Mertins, A.: Multi-view audio and music classification. In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 611–615 (2021). IEEE

[37] Sharma, D., Taran, S., Pandey, A.: A fusion way of feature extraction for automatic categorization of music genres. Multimedia Tools and Applications **82**(16), 25015–25038 (2023)

[38] Sarris, A.L., Vryzas, N., Vrysis, L., Dimoulas, C.: Investigation of data augmentation techniques in environmental sound recognition. Electronics **13**(23), 4719 (2024)

[39] Nguyen, K.T.A., Tran, N.H., Dinh, V.Q.: Fmmix: A novel augmentation strategy for enhancing feature diversity to boost model generalization for image data. IEEE Access (2024)

[40] Zhang, Z.: Improved adam optimizer for deep neural networks. In: 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS), pp. 1–2 (2018). IEEE

[41] Sun, R.-Y.: Optimization for deep learning: An overview. Journal of the Operations Research Society of China **8**(2), 249–294 (2020)

[42] Opitz, J.: A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. Transactions of the Association for Computational Linguistics **12**, 820–836 (2024)

[43] Susmaga, R.: Confusion matrix visualization. In: Intelligent Information Processing and Web Mining: Proceedings of the International IIS: IIPWM '04 Conference Held in Zakopane, Poland, May 17–20, 2004, pp. 107–116 (2004). Springer