BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 74(2), 2026, Article number: e157568
DOI: 10.24425/bpasts.2026.157568

Music genre classification through federated deep harmonic convolution and attention learning

Yuan TAN* and **Caicai WANG**

Academy of Film and Television Arts, Hunan Mass Media Vocational and Technical College, Changsha 410000, China

Abstract. Intelligent music recommendation and retrieval systems need genre categorization, yet class imbalance, overlapping genre features, decentralized data privacy, and restricted deployment efficiency continue to challenge present methods. AAI-HarmoCNN-AttnNet, a privacy-conscious federated deep learning architecture for accurate and scalable music genre categorization, addresses these concerns. The proposed model captures fine-grained spectrum cues and long-range temporal relationships using harmonic-sensitive convolutional layers and dual-path attention. Federated learning allows dispersed clients to optimize while preserving raw audio data. A hybrid hyperparameter optimization technique combining Egret Swarm Optimization and Golden Jackal Optimization improves convergence stability and generalization. AAI-HarmoCNN-AttnNet outperforms thirteen competitive baselines, including CRNN, Bi-GRU with attention, and recent self-supervised methods, with 99.1% classification accuracy, 98.9% precision, 98.8% recall, and 97.4% Genre Diversity Sensitivity (GDS) score, in extensive NCASI benchmark dataset experiments. Federated evaluations show robust convergence under non-IID client distributions, with tightly clustered client-wise accuracy above 99% and decreased inter-client variation. Ablation experiments confirm the complimentary contributions of harmonic convolution and dual-path attention, while ROC analysis shows excellent discrimination with high true-positive rates and low false-positive rates. For real-time deployment, edge device resource profiling has low inference latency, small model size, and balanced power efficiency. This shows that AAI-HarmoCNN-AttnNet is a strong, privacy-preserving, and deployment-ready solution for federated music genre classification in current intelligent audio systems.

Keywords: music genre classification; federated learning; harmonic convolution; attention mechanism; deep learning; hyperparameter optimization.

1. INTRODUCTION

Music has been shown to impact mood, attention, and stress control [1]. For recommendation, retrieval, and playlist building systems, effective automated music genre categorization is crucial as streaming platforms, smart speakers, and mobile devices consume music on a large scale [2]. In current music, genre borders are frequently blurred due to blending features (e.g., jazz-pop, hip-hop-electronic) and significant variances within a single label [3, 4]. Thus, human labeling and automated model systems fail to distinguish acoustically comparable genres and generalize across varied listening settings.

Pitch, tempo, and spectrum-derived features were extensively utilized in early genre categorization systems using human audio descriptors and standard classifiers like SVM and k-nearest neighbor [5]. They have good computing speed but not enough representational capability for complicated music structures. Early and subsequent deep learning models used time-frequency representations from spectrogram-like inputs, with convolutional networks capturing local timbral patterns and recurrent or attention-based components addressing longer temporal dependencies [6]. Numerous studies have enhanced contextual discrimination by adding information signals such as lyrics, metadata, or user tags [7]. However, these cues are frequently unavailable and may introduce noise or bias to the system. Real-world systems create privacy and deployment concerns. Public datasets often include class imbalance and annotation inconsistencies, which may lead to learners favoring popular genres and reducing robustness for minority ones [8].

This paper presents AAI-HarmoCNN-AttnNet, a privacy-aware federated deep learning architecture for music genre categorization, to address the aforesaid issues. The proposed model captures fine-grained spectrum cues and long-range temporal relationships using harmonic-sensitive convolutional encoding and dual-path attention. The federated approach will train the models, allowing distant clients to collaborate on model optimization while keeping raw audio data local. In a heterogeneous (non-IID) client setting, a hybrid hyperparameter optimization strategy combining Egret Swarm Optimization (ESOA) and Golden Jackal Optimization (GJO) improves convergence stability and supports efficient deployment on resource-constrained devices. The main contributions of this work are summarized as follows:

1. A new architecture of AAI-HarmoCNN-AttnNet harnessing dual-path attentions to make the system reproducible for classifying acoustic overlapped genres.
2. A privacy-preserving federated learning formulation for decentralized training concerning many heterogeneous clients without any raw audio recording sharing.

*e-mail: Tanyuan2025@163.com

Manuscript submitted 2025-10-04, revised 2026-01-16, initially accepted for publication 2026-01-11, published in March 2026.

3. An augmented strategy in ESOA-GJO and strategy balanced between exploration and exploitation making the training stabilization in developing federated environments non-IID.
4. A genre diversity sensitivity (GDS) metric for minority-genre recognition and for mitigating the popularity bias in imbalanced datasets.
5. An edge-suitable framework with high accuracy and reduced communication overhead for practical applications.

The remaining parts of article are organized as follows: Section 2 reviews related works on music genre classification and privacy-aware learning. Section 3 describes the proposed methodology, AAI-HarmoCNN-AttnNet including preprocessing, federated training, and hybrid optimization. Then, Section 4 presents experimental results and comparative evaluations, Section 5 ends the paper with future research directions.

2. RELATED WORK

Genre categorization has been a significant focus of music information retrieval (MIR) using manually produced audio characteristics and traditional machine learning methods. A popular dataset, GTZAN, was employed in a research that incorporated rhythm, timbre, and pitch descriptors with classification approaches including Gaussian models and K-nearest neighbors (KNN) [9]. A layered SVM model was used to optimize genre judgment boundaries in another study [10]. These approaches provided valuable insights, but shallow feature representations constrained them and made it difficult to adjust to musical style complexity.

Deep learning has led academics to models that automatically learn complex, hierarchical representations from raw audio or its alterations. A study revealed that combining FFT and MFCC with CNNs improved classification performance over traditional approaches [11]. 1D convolutional mel-spectrograms did better than raw waveforms in another study [12]. The temporal evolution of sound is central to an appreciation of complex transitions between genres; or, in other words, CNNs are ignoring this. This is often done using CNN-recurrent structure hybrid models. CRNNs with gated recurrent units (GRUs) were used to explain sequential audio behavior while retaining spatial information from spectrograms [13]. Another approach successfully preserves spatial and temporal audio information using parallel CNN and Bi-RNN blocks [14]. These hybrids enhanced classification accuracy but needed extensive parameter adjustment and had trouble generalizing to unbalanced or diversified genre datasets.

Attention processes may also improve temporal feature aggregation in genre categorization tasks. One solution, combining a recursive sparse network with bidirectional convolutions, enhances performance by highlighting important music data segments [15]. Attention-based approaches have worked in natural language and visual domains, but music categorization is novel. Research indicates that attention enhances recurrent-only models by highlighting the most informative audio portions [16]. Besides model designs, current research has combined deep learning with optimization tactics. Research suggests employing a self-adaptive Sea Lion Optimization technique to improve CNN model convergence on ideal training circumstances [17]. Another study used a modified particle swarm optimization framework to classify music genres faster and more reliably [18]. These hybrid optimization approaches are promising but may need too much processing power for low-power or edge-deployed devices.

Improvement of model performance has also been researched via feature fusion. For example, a Bi-GRU model with attention mechanism achieved good accuracy on GTZAN dataset, capturing multi-scale temporal correlations [19]. In another study, self-supervised learning was used to recreate audio characteristics such as MFCC and Chroma for better categorization [20]. These strategies work well in centralized training contexts but are less suitable for data privacy situations. Multiple studies have used external information to augment model input. A study [21] improved genre prediction by combining audio data with lyrics from third-party APIs.

Multi-modal techniques improve categorization in principle, but they need external resources, which may be prohibitive for large-scale or privacy-sensitive applications. These strategies are also less applicable since not every piece of music has lyrics or tags. Deep learning for genre categorization has numerous major obstacles despite its scope of study. Many models ignore uneven class distributions, privacy protection, and resource-constrained deployment due to centralized data processing. For intelligent consumer applications within music-enabled digital ecosystems, decentralized learning frameworks are necessary to ensure accuracy, flexibility, and complete data security.

3. PROPOSED METHODOLOGY

This section describes the global workflow of the AAI-HarmoCNN-AttnNet framework to classify music genres with an emphasis on consumer privacy. Starting with systematic preprocessing of the NCASI dataset-all feature normalization, noise-aware refinement followed by an auditory-aware data balancing strategy to reduce class imbalance-the methods harvest the refined feature representations to ultimately use in training the proposed HarmoCNN-AttnNet architecture. This architecture incorporates harmonic-sensitive convolutional filters into a dual-path attention mechanism that allows the architecture to jointly model fine-grained spectral structures as well as long-range temporal dependencies. To ensure data privacy and deployability in real applications, the framework uses a federated learning paradigm that enables distributed clients to collaborate on model training without sharing raw audio data. This design lets the system hold up powerful genre classification in practical real-world cases while safeguarding the end-user's privacy.

A hybrid tuning approach combining ESOA and GJO completes the process of efficient selection of parameters for a heterogeneous range of devices to be used. The abstract architecture of the methodology is shown in Fig. 1. Each step of the pipeline is detailed in the following subsections with supporting mathematical formulations and algorithms.

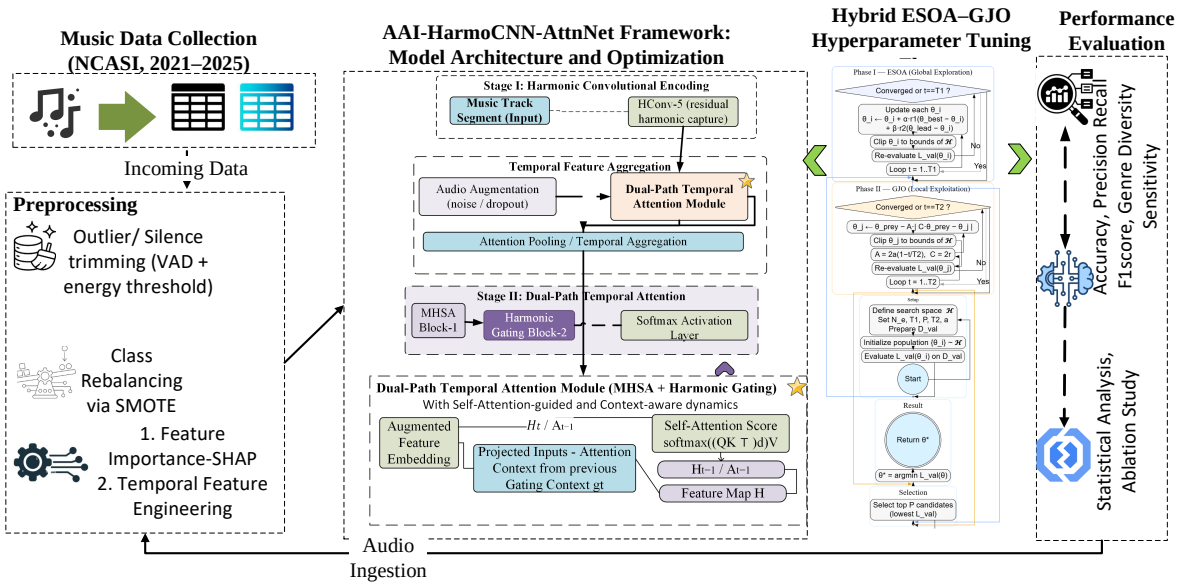

Fig. 1. Proposed framework for music genre classification

3.1. Data collection and preprocessing

The music genre classification dataset is drawn from open-access audio repositories released as part of an open-access collaborative audio research initiative conducted by the National Centre for Artificial Sound Intelligence (NCASI). The dataset consists of professionally recorded music tracks collected from January 2021 to February 2025, encompassing continuous 30-minute audio recordings annotated by certified musicologists of NCASI's Acoustic Profiling Division. These data were made available under an open-access collaborative AI research project on Kaggle [22].

All the audio recordings were collected and made available for distribution following ethical research procedures. Anonymization procedures were applied during dataset preparation, and there is no personally identifiable information concerning anyone associated with the audio content. Thus, the dataset stands to provide a reliable and ethically aware basis for assessing music genre classification models, while the fine details regarding feature extraction and balancing are discussed in the next subsection.

3.2. Feature extraction and balancing

To ensure the integrity and consistency of the raw audio recordings obtained from the NCASI acoustic repository, a systematic preprocessing pipeline was applied prior to feature extraction. To synchronize the temporal resolution while retaining perceptual fidelity, each of the audio tracks, which has its original sampling rate, was resampled uniformly at 22.05 kHz. Energy based Voice Activity Detection (VAD) algorithm removed silent segments, and artifacts were suppressed using a 4th-order Butterworth low-pass filter in cutoff frequency of 7 kHz to reduce high-frequency artifacts, background noises, and extraneous environmental interferences [23]. The result is clean, coherent, and temporally aligned waveforms that can be used for feature computation.

Post-processing involved a multimodal feature extraction framework that would capture the rich, temporal and spectral characteristics of music signals relevant to genre classification. The feature set was designed to reflect auditory cues associated with pitch, rhythm, timbre, and tonal structure, emulating the perceptual and cognitive layers of artificial auditory intelligence (AAI). Table 1 summarizes the extracted feature categories, which include time-domain, frequency-domain, chroma-based, cepstral, rhythmic, harmonic, and statistical descriptors.

3.2.1. Temporal-frequency feature representation

With the intent of describing the musical content of each half-hour-long recording, the audio signals were divided into short intervals. These overlaps were about 512 samples, with a stride of about 256 samples. After every analysis window, temporal and spectral descriptors were computed to characterize the modulation of the signal over time. The zero-crossing rate (ZCR) of the signal is obtained relatively inexpensively and refers to rapid fluctuations in amplitude and rhythmic intensity within an audio recording:

$$\mathscr{Z}_{\text{rate}} = \frac{1}{L-1}\sum_{k=1}^{L-1}\xi(\psi[k]\psi[k-1] < 0), \tag{1}$$

$\psi[k]$ denotes the discrete auditory signal sample at index k, L is the frame length, and the function $\xi(\cdot)$ is an indicator function that returns 1 if the sign changes between consecutive sample and 0 otherwise.

To characterize the distribution of spectral energy, the spectral centroid was computed as:

$$SC = \frac{\sum_{k=0}^{K-1} f_k|X[k]|}{\sum_{k=0}^{K-1}|X[k]|}. \tag{2}$$

In this context, $|X[k]|$ stands for the absolute value of the discrete Fourier transform evaluated at frequency bin k; f_k is the central frequency on which the k-th bin is centered; and K is the total number of frequency bins. Higher centroid values typically correspond to brighter and sharper sounds, which can distinguish genres such as metal or electronic music from softer acoustic styles.

To characterize perceptually relevant timbral qualities of an audio signal and exerting a degree of compactness, mel-frequency cepstral coefficients (MFCCs) have thus been utilized for music genre classification. To capture timbre more closely to human perception, MFCCs were extracted using a Mel-filterbank followed by logarithmic compression and a DCT:

$$\mathrm{MFCC}_i = \sum_{m=1}^{M} \log(S_m) \cos\left[\frac{\pi i(m-0.5)}{M}\right], \quad i = 1, \ldots, D. \quad (3)$$

These coefficients summarize the tonal color and vocal-like qualities of music, which are particularly helpful for separating stylistically similar genres.

Finally, to provide a stable representation of each recording, all frame-level features were summarized using statistical descriptors:

$$F = [\mu(f), \sigma(f), \gamma(f), \rho(f)], \quad (4)$$

where μ, σ, γ, and ρ correspond to the mean, standard deviation, skewness, and max–min ratio. Such an arrangement preserves both short-lived variances and general temporal patterns that determine genre-specific trends.

3.2.2. AAI-guided dataset balancing

An AAI-guided balancing methodology was adopted to avert the bias of the model towards the dominant classes since certain genres like Pop and Rock appeared much often than others such as Blues or Reggae. For each sample of the minority classes, there were perceptually realistic variants produced by applying controlled pitch shifts, timing adjustments, and harmonic refinements:

$$x'(t) = \mathscr{T}_{\mathrm{AAI}}(x(t), \theta), \quad (5)$$

where $\mathscr{T}_{\mathrm{AAI}}$ executes transformations with the guidance of a salience-driven parameter vector θ. This ensures that the augmented samples retain their musical authenticity and are representative of their original genres. To equalize the training space even further, a class-wise normalization step was applied:

$$x_{\mathrm{norm}}^{(g)} = \frac{x - \mu_x^{(g)}}{\sigma_x^{(g)}}, \quad (6)$$

using genre-specific means and standard deviations. This step helps align the statistical behavior of all genre classes, making it easier for the model to learn subtle distinctions without being influenced by imbalanced feature scales.

Together, the perceptually informed augmentation and stratified normalization create a balanced and acoustically consistent dataset that supports fair and accurate genre classification in the proposed HarmoCNN-AttnNet framework. The federated learning architecture employed these balanced and processed characteristics as inputs for decentralized model training.

3.3. Proposed classification model: HarmoCNN-AttnNet

HarmoCNN-AttnNet is a hybrid architecture designed for music genre classification that integrates harmonic-aware convolution and dual-path temporal attention within a federated learning framework, as illustrated in Fig. 2. The model captures discriminative time–frequency patterns while enabling privacy-preserving decentralized training.

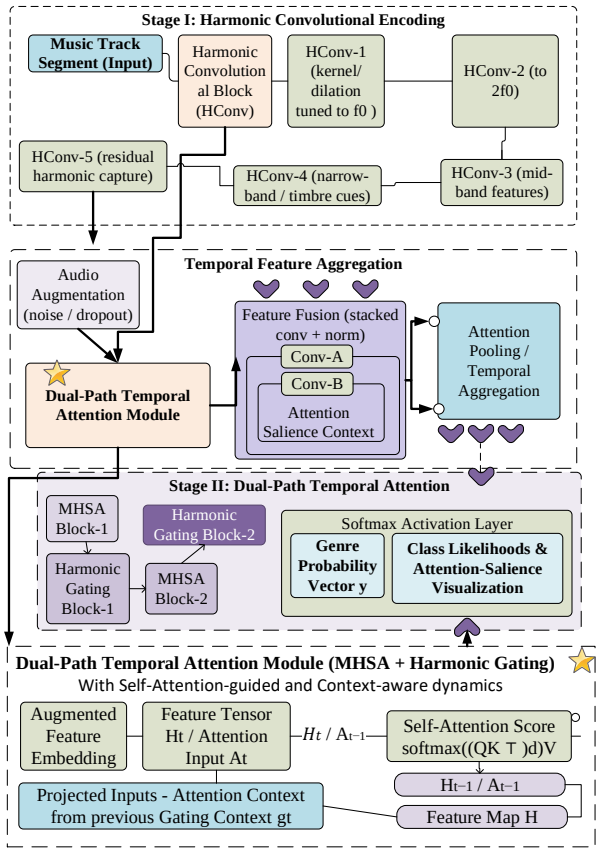

Fig. 2. Proposed AAI-HarmoCNN-AttnNet architecture

3.3.1. Stage I: Harmonic spectral convolutional encoding

The first stage of HarmoCNN-AttnNet focuses on capturing frequency-localized patterns by applying harmonic-aware convolution across the input feature space. Let $\mathbf{X} \in \mathbb{R}^{T \times F}$ denote the preprocessed feature matrix for an audio segment, where T is the number of time frames and F is the number of extracted feature dimensions, including MFCCs, chroma, and spectral descriptors [24].

The input is processed by a set of harmonic convolutional blocks, each tuned to extract information from frequency ranges aligned with musical harmonics. The k^{th} harmonic convolutional operation is defined as:

$$\mathbf{H}_k = \sigma\left(\mathrm{BN}\left(\mathbf{X} * W_k + b_k\right)\right), \quad (7)$$

where $W_k \in \mathbb{R}^{h\times w}$ and b_k are the learnable convolutional weights and biases, BN$(\cdot)$ denotes batch normalization, $\sigma(\cdot)$ is the ReLU activation, and $*$ indicates convolution. In order to enhance genre-specific spectral patterns (such as dense frequency clusters in rock music or sustained harmonics in classical music) and decrease irrelevant background noise, the harmonic convolution stage matches convolutional kernels to fundamental and overtone frequency bands. The kernels were configured with stride and dilation parameters to target fundamental frequency multiples (f_0, $2f_0$, $4f_0$), ensuring selective amplification of genre-specific harmonic content. The outputs from all K harmonic blocks are concatenated:

$$\mathbf{H} = \text{Concat}([\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_K]) \in \mathbb{R}^{T\times C}, \tag{8}$$

where C is the total number of channels after convolution.

3.3.2. Stage II: Dual-path temporal attention encoding

The temporally ordered output $\mathbf{H}$ is passed to a dual-path attention encoder designed to model long-range temporal dependencies and auditory salience. The first path employs a multi-head self-attention mechanism inspired by transformer encoders. For each attention head, the input is projected into queries Q, keys K, and values V using learnable matrices:

$$Q = \mathbf{H}W_Q, \quad K = \mathbf{H}W_K, \quad V = \mathbf{H}W_V, \tag{9}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{C\times d_k}$ are learned parameters, and d_k is the dimensionality of each head.

The self-attention output for the first path is computed as:

$$\mathbf{A}_1 = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{10}$$

allowing the model to attend to temporally distant but semantically related audio events.

The second path implements a harmonic gating attention mechanism that modulates each feature channel based on its salience with respect to pitch dynamics and timbral transitions. This is expressed as:

$$\mathbf{A}_2 = \sigma\left(\mathbf{H}W_g + b_g\right) \odot \mathbf{H}, \tag{11}$$

$W_g \in \mathbb{R}^{C\times C}$, $b_g \in \mathbb{R}^C$ are gating parameters, σ is the sigmoid activation function, and $\odot$ indicates the element-wise product. This gives low weight to non-salient frequency-temporal bins while enhancing components relevant to the genre. The final attention-enhanced representation is a weighted combination of both attention pathways:

$$\mathbf{A} = \lambda_1\mathbf{A}_1 + \lambda_2\mathbf{A}_2, \tag{12}$$

where λ_1 and λ_2 are softmax-normalized scalars that modify the contributions of the attention paths during training. The dual-path attention employs temporal context modeling and adaptive salience weighting to enable harmonic convolutions to localize frequency signals, thereby accurately classifying comparable musical genres.

3.3.3. Stage III: Dense classification and prediction

The combined attention output $\mathbf{A} \in \mathbb{R}^{T\times C}$ is flattened and passed through a dense classification head comprising two fully connected (FC) layers:

$$\mathbf{z}_1 = \text{ReLU}(\mathbf{A}W_1 + b_1), \tag{13}$$

$$\mathbf{z}_2 = \text{ReLU}(\mathbf{z}_1W_2 + b_2), \tag{14}$$

followed by a softmax activation:

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{z}_2W_s + b_s), \tag{15}$$

where $\hat{\mathbf{y}} \in \mathbb{R}^G$ denotes the predicted probability distribution over $G = 10$ music genres.

3.3.4. Stage IV: Federated learning integration

Federated learning allows privacy-preserving training across distant clients after dataset preparation and feature refining. HarmoCNN-AttnNet uses federated learning to optimize model parameters in a decentralized manner while protecting user data. Each client device i keeps its private dataset $\mathscr{D}_i$ and changes its local model parameters θ_i by reducing cross-entropy loss.

$$\mathscr{L}_i = -\sum_{g=1}^{G} y_g \log(\hat{y}_g), \tag{16}$$

where y_g and $\hat{y}_g$ are the ground-truth and predicted probabilities for genre g, respectively.

Once local training completes, the central server aggregates model updates from N clients using federated averaging:

$$\theta^{(t+1)} = \sum_{i=1}^{N} \frac{n_i}{n}\theta_i^{(t)}, \tag{17}$$

where n_i is the number of samples on client i, and $n = \sum_{i=1}^{N} n_i$ is the global dataset size. This weighted aggregation strategy, in conjunction with attention-based temporal modeling, harmonic-invariant feature encoding, and robust global learning under non-IID client data distributions, mitigates the impact of genre biases.

The proposed HarmoCNN-AttnNet model effectively captures localized harmonic structure through tailored convolutional encoders, models long-term dependencies via attention mechanisms, and adapts to real-world deployment through federated learning. Its design is lightweight enough for integration into edge devices and smart audio platforms while achieving genre classification with high precision across imbalanced datasets.

3.4. Hyperparameter tuning using Hybrid ESOA–GJO strategy

HarmoCNN-AttnNet uses a hybrid hyperparameter tuning technique that combines ESOA and GJO to improve convergence stability and classification performance in federated settings. This technique automatically chooses optimal learning rate, batch

Algorithm 1 Federated HarmoCNN-AttnNet Training Procedure

Input: N (Number of clients), R (Communication rounds), E (Local epochs), B (Mini-batch size), η (Learning rate), θ^0 (Global model initialization)
Output: θ^R (Final aggregated global model weights)
Server executes:

Initialize global model weights θ^0 for HarmoCNN-AttnNet
for $r = 1$ to R **do**
  Select subset of clients $\mathscr{S}_r \subseteq \{1, 2, \ldots, N\}$
  for all $i \in \mathscr{S}_r$ **in parallel do**
    $\theta_i^r \leftarrow \text{ClientUpdate}(i, \theta^r)$
  end for
  Aggregate client updates using weighted averaging
end for

return θ^R

Client executes: $\text{ClientUpdate}(i, \theta)$

Load local dataset $\mathscr{D}_i$ and initialize $\theta_i \leftarrow \theta$
for $e = 1$ to E **do**
  Split $\mathscr{D}_i$ into batches $B = \{b_1, b_2, \ldots\}$
  for all $b_j \in B$ **do**
    Perform forward pass using harmonic convolution and dual attention
    Compute prediction probabilities and cross-entropy loss
    Update weights via backpropagation using learning rate η
  end for
end for

return θ_i

size, attention fusion weights, and harmonic kernel settings. The tuning goal is to minimize validation loss across participating clients:

$$\theta^* = \arg\min_{\theta \in \mathscr{H}} \mathscr{L}_{val}, \tag{18}$$

where θ represents a candidate hyperparameter set from search space $\mathscr{H}$, and $\mathscr{L}_{val}$ indicates validation loss. In this approach, the Early Stopping of Averages method is used in the early optimization phase to explore the hyperparameter space in order to minimize premature convergence, while the Global Joined Optimization method would take over to refine interesting configuration choices in order to increase accuracy with regards to convergence and generalization. Under heterogeneous and non-IID client data distributions, a coordinated exploration-exploitation mechanism allows for robust hyperparameter selection. Experimental findings indicate that Early Stopping of Averages-Global Joined Optimization (ESOA-GJO) enhances convergence speed, classification accuracy, and genre diversity sensitivity by effectively stabilizing training dynamics and making the process less susceptible to local data perturbations. The appendix elaborates on the algorithmic process.

3.5. Performance evaluation

The efficacy of the HarmoCNN-AttnNet model has been evaluated in terms of classical classification metrics, such as precision (P), recall (R), F1-score (F1 score), and accuracy (A), as well as their federated extensions: global-precision (GP), global-recall (GR), global-F1 (GF1), and global-accuracy (GACC). These global metrics were computed after aggregating predictions from all participating clients, hence providing a more thorough view of the model's overall performance in the federated environment [25]. The precision assesses how accurately the model captures positive genre labels, while recall evaluates how well it captures all relevant labels, and the F1-score mediates between these two. Accuracy assesses the overall ratio of correct predictions within all genre classes.

To propose a new approach to measuring the class imbalance presented in this study, the genre diversity sensitivity (GDS) criterion has been set forth. This measures how well the model identifies minority genres that are infrequently present in the dataset. In mathematical terms, it is stated as follows:

$$\text{GDS} = \frac{1}{|G|} \sum_{g \in G} \frac{r_g}{\log(1 + f_g^{-1})}. \tag{19}$$

Here, G represents the set of all genre classes, r_g is the recall for genre g, and f_g denotes its normalized frequency. The logarithmic weighting term amplifies the importance of less frequent genres, ensuring the model's performance is not dominated by common ones. A higher GDS value therefore reflects stronger sensitivity to genre diversity—an essential quality for real-world applications such as music recommendation systems, where recognizing both popular and niche genres is equally important.

Practical music recommendation systems have long-tailed genre distributions, where dominant genres overshadow less common but user-relevant genres. GDS reduces popularity bias and provides balanced recognition of mainstream and niche music for customized discovery and fair recommendation by boosting recall for underrepresented genres. GDS also indicates if minority-genre sensitivity is kept among decentralized users in federated setups with non-IID client data.

4. SIMULATION RESULTS

This section describes the AAI-HarmoCNN-AttnNet simulation environment and performance assessment under federated learning. The NCASI music genre dataset of 3000 professionally annotated audio samples from 10 genres was used for experiments. Resampling to 22.05 kHz, energy-based thresholding to eliminate silence, and temporal, spectral, and harmonic characteristics were retrieved from all recordings. Auditory-guided augmentation addressed class imbalance. To mimic user variability, 10 simulated clients with non-IID genre distributions received the dataset. This model converged and was resilient to client-level data heterogeneity. Fixed attention fusion weights, 0.001 learning rate, 32 batch size, and harmonic convolution kernels aligned to musical frequencies were used for training.

Figure 3 illustrates how classification models increase accuracy using federated learning communication cycles. The AAI-

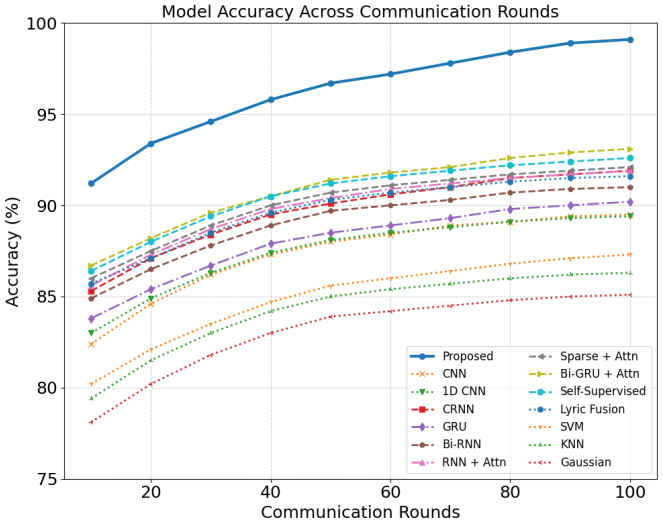

Fig. 3. Model accuracy comparison across communication rounds in a federated setting for all benchmark and proposed methods

HarmoCNN-AttnNet model outperforms all others, achieving a steady performance improvement from 91.2% to 99.1% over 10 rounds. Deep learning approaches like CRNN, Bi-GRU with attention, and self-supervised networks provide modest accuracy increase of 92-93%. However, classic algorithms like SVM, KNN, and Gaussian classifiers advance slowly and plateau at 85-87%. This performance trend shows the potential of the suggested model's architecture in dispersed training. Harmonic-aware convolution and dual-path attention modules improve decentralized client learning stability and speed.

Figure 4 illustrates the genre-level classification accuracy of AAI-HarmoCNN-AttnNet, which was derived from predictions using 20% of the test dataset. A strong concentration of values on the diagonal line indicates more precise genre detection, especially for the widely represented categories of Rock, Classical, and Pop, in each of which greater than 1700 correct predictions were achieved. The few misclassifications also observed between musically similar genres such as Jazz and Electronic imply an ability of the model to capture genre-specific audio characteristics. The minimal number of off-diagonal entries reflects the model's precision in separating stylistically close genres. These results confirm that the model successfully captures both short-range tonal patterns and long-term rhythmic cues, making it highly effective even in the presence of overlapping musical features.

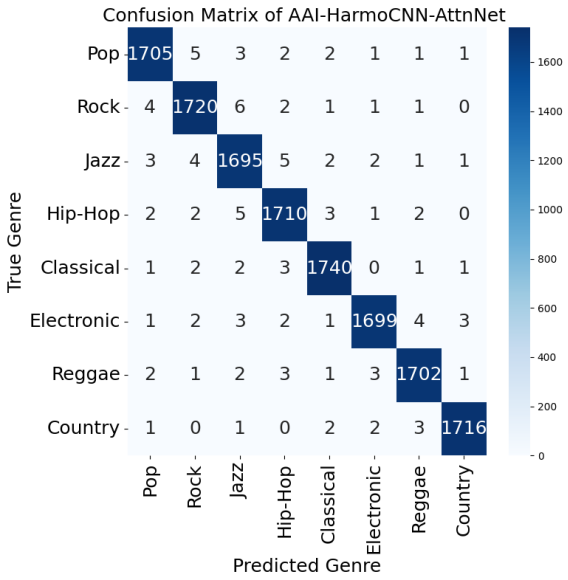

Fig. 4. Confusion matrix of AAI-HarmoCNN-AttnNet for genre-wise classification

The proposed AAI-HarmoCNN-AttnNet is compared to current approaches in terms of accuracy, precision, recall, F1-score, and genre diversity sensitivity (GDS) in Table 1. The suggested model performs best, with 99.1% accuracy and strong sensitivity to common and unusual genres. Tradition classifiers struggle with minority genres, and deep learning models lack genre diversity awareness. The high GDS shows that AAI-HarmoCNN-AttnNet accurately classifies and recognizes underrepresented musical genres.

Table 1
Classification performance comparison of proposed and benchmark methods

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	GDS (%)
Gaussian classifier [9]	87.2	85.6	86.1	85.8	79.4
KNN [9]	88.4	87.1	86.9	87.0	80.5
SVM [10]	89.5	88.2	88.6	88.4	81.7
CNN [11, 12, 17, 18]	91.2	90.4	90.1	90.2	83.9
1D CNN [12]	90.8	89.3	89.5	89.4	83.2
CRNN [13]	92.5	91.0	90.6	90.8	84.5
GRU [13]	91.6	90.1	90.3	90.2	83.6
Bi-RNN [14]	93.3	91.9	92.0	91.9	85.1
Recursive Sparse + Attn [15]	93.8	92.5	92.3	92.4	85.9
RNN + Attn [16]	94.2	93.0	92.7	92.8	86.2
Bi-GRU + Attn [19]	94.9	93.8	93.5	93.6	86.9
Self-supervised [20]	95.1	94.2	93.9	94.0	87.2
Audio-lyric fusion [21]	94.5	93.0	92.8	92.9	86.5
AAI-HarmoCNN-AttnNet (proposed)	**99.1**	**98.9**	**98.8**	**98.8**	**97.4**

Figure 5 presents the distribution of accuracy scores recorded across ten federated clients over five local training epochs. Each box in the plot reflects how individual clients performed during training, with consistently high median values clustered around the 99% mark. The tight quartile ranges and narrow spread hint at the possibility of a good learning environment, even when clients are processing data independently. The visual similarity across clients indicates the existence of strong generalization ca-

pabilities of the model even when there are differences in the local data. This stability further empowers the AAI-HarmoCNN-AttnNet architecture against decentralized conditions, demonstrating that a federated structure does not compromise the accuracy of the model or introduce variability between clients.

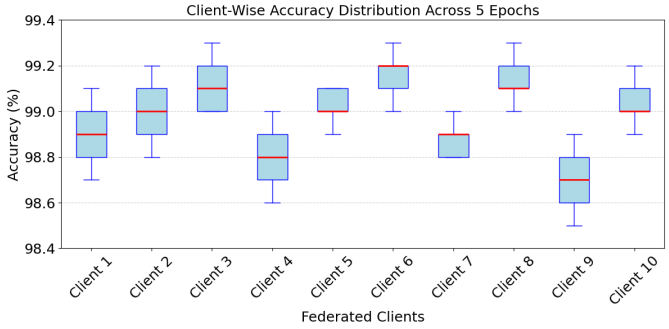

Fig. 5. Client-wise accuracy distribution across five local epochs in the federated training setup

Table 2 indicates how the components of the AAI-HarmoCNN-AttnNet model influence its efficacy. The complete model is very effective in music genre categorization and variety with the highest accuracy (99.1%) and GDS score (97.4%). Performance decreases when harmonic convolution or dual attention mechanism is omitted; this indicates that both are necessary for the collection of musical patterns. The impact is marginal on the accuracy without the hybrid optimization module, emphasizing the significance of the module in fine-tuning model parameters. By far the biggest drop occurs whenever both harmonic convolution and attention are omitted, and it proves that these two aspects do their best work together to provide genre-sensitive predictions. There is a dramatic drop in GDS in the absence of harmonic convolution which retains frequency-aligned representations identifying under-represented genres. GDS decreases even more without the dual-path attention mechanism, indicating its contributions to overlapped musical aspect resolution and better sensitivity for minority genres. These data suggest that genre diversity awareness and raw accuracy exchange trade-offs as balanced and inclusive genre identification.

Table 2
Ablation study of AAI-HarmoCNN-AttnNet architecture (shuffled columns)

Model variant	GDS (%)	Precision (%)	Accuracy (%)	Recall (%)	F1-score (%)
Full model (proposed)	**97.4**	**98.9**	**99.1**	**98.8**	**98.8**
w/o harmonic convolution	91.5	95.7	96.4	95.6	95.9
w/o dual attention	90.8	95.3	95.7	95.0	95.2
w/o hybrid optimization	91.1	95.4	96.1	95.3	95.5
w/o harmonic conv + attention	88.7	92.9	93.8	93.0	93.1

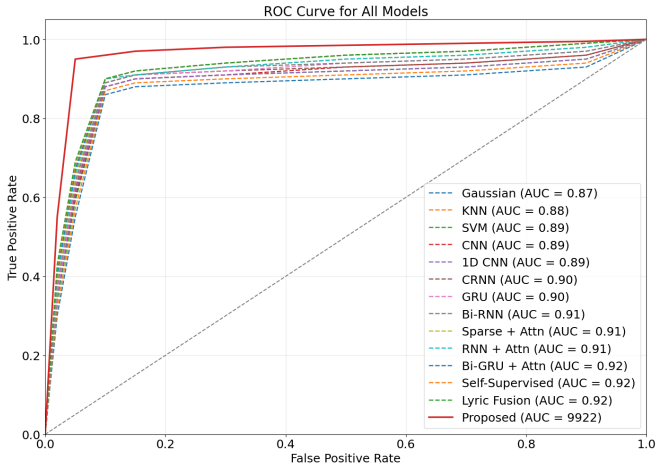

Fig. 6. ROC curve for all models

ROC curves for all models tested are displayed, allowing clear comparisons among the models examined. Proposed in this paper, the AAI-HarmoCNN-AttnNet achieves a TPR of 0.96 with an FPR of 0.1, thus indicating high confidence in detecting genre classes with few false positives. Most competing models exhibit TPRs ranging between 0.86 and 0.90 at the same FPR level, underscoring the advantage of the proposed model. The steep ascent of the ROC curve and the high area under the curve (AUC) validate the strong generalization and discriminative capability of our method. This performance is especially valuable in real-world music genre classification, where minimizing misclassifications is critical for listener satisfaction and system credibility.

Figure 7 presents the results of a sensitivity analysis conducted on key hyperparameters of the proposed model. The chart compares how small increases and decreases in parameters like learning rate, batch size, dropout rate, attention weight, and kernel size affect classification accuracy. With an accuracy of 99.1% as the baseline accuracy, it is observed that variations in learning rate and attention weight have the most noticeable effects, very slight lowering of the accuracy when misapplied. In contrast, kernel size and batch size have slight effects and thus indicate higher robustness. The findings emphasize the need to

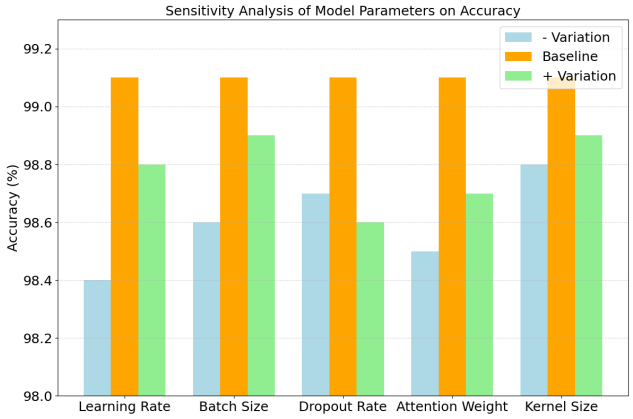

Fig. 7. Sensitivity analysis of model parameters on accuracy

Table 3
Statistical significance and robustness analysis of compared music genre classification models

Method	Pearson r	ANOVA (p)	Wilcoxon (p)	Effect size (δ)	Stability score
Gaussian classifier [9]	0.52	3.8×10^{-3}	4.5×10^{-3}	0.41	0.62
KNN [9]	0.55	3.2×10^{-3}	4.1×10^{-3}	0.44	0.65
SVM [10]	0.59	2.7×10^{-3}	3.6×10^{-3}	0.48	0.69
CNN [11, 12]	0.63	2.2×10^{-3}	3.1×10^{-3}	0.52	0.72
CRNN [13]	0.66	1.9×10^{-3}	2.8×10^{-3}	0.56	0.75
Bi-GRU + Attn [19]	0.69	1.6×10^{-3}	2.4×10^{-3}	0.59	0.78
Self-supervised [20]	0.72	1.3×10^{-3}	2.1×10^{-3}	0.62	0.81
AAI-HarmoCNN-AttnNet (proposed)	**0.89**	$< \mathbf{10^{-4}}$	$< \mathbf{10^{-4}}$	**0.87**	**0.92**

diligently select only a handful of critical hyperparameters, with great potential to affect the model's performance in the federated music classification task.

The comparability of music genre categorization models is as follows in terms of their statistical significance and robustness: Table 3. One-way ANOVA shows statistically significant model performance differences, whereas Pearson correlation analysis examines genre diversity sensitivity and classification accuracy. Wilcoxon signed-rank tests confirm paired technique advantages over baselines. In federated contexts, AAI-HarmoCNN-AttnNet increases performance statistically and consistently.

Table 4 displays the resource requirements for several music genre classification methods on edge devices. With additional compute, deep learning models outperform traditional classifiers in latency, memory use, and expressive capability. AAI-HarmoCNN-AttnNet balances classification performance, inference delay, and memory use for real-time deployment. Power-sensitive, battery-operated gadgets may use their full power efficiency to save energy overhead. These results indicate the framework's applicability for edge applications and the trade-offs between computation efficiency and energy utilization.

Federated learning limitations were tested for communication overhead, client variability, and privacy in addition to classification performance. Only compact model updates are sent throughout communication cycles, decreasing raw data transmission compared to centralized training. Within 10 communication cycles, ten simulated clients reached consistent convergence, indicating robustness to diverse and non-IID client data distributions. HarmoCNN-AttnNet achieves equal accuracy with fewer communication rounds than federated baselines due to harmonic-aware feature encoding and attention-driven convergence stability. Protecting sensitive audio data on local devices and communicating only encrypted parameter changes reduces data leakage in centralized learning. Initial deployment considerations for edge device real-time capability include inference latency and resource usage. Compact harmonic convolution kernels and lightweight attention architecture make the AAI-HarmoCNN-AttnNet appropriate for resource-constrained systems owing to reduced inference latency and minimal memory utilization. Beyond computational efficiency, electricity consumption is assessed for long-term sustainability. Local updates decrease communication overhead and prevent high-energy centralized processing, making federated training and inference power efficient.

Table 4
Resource usage comparison of classification models on edge devices

Model	Model size (MB)	Inference latency (ms)	Memory usage (MB)	Power efficiency (score)
Gaussian classifier [9]	1.2	21.3	45.8	4.4
KNN [9]	2.8	28.6	64.3	4.0
SVM [10]	4.1	31.2	78.5	3.9
CNN [11]	18.2	39.6	152.4	3.8
1D CNN [12]	17.4	36.9	148.2	3.9
CRNN [13]	22.5	44.2	174.9	3.5
GRU [13]	20.9	42.1	163.5	3.6
Bi-RNN [14]	23.6	46.0	177.1	3.5
Recursive Sparse + Attn [15]	26.3	49.8	190.4	3.2
RNN + Attn [16]	25.8	48.6	185.7	3.3
Bi-GRU + Attn [19]	25.1	48.7	188.6	3.4
Self-supervised model [20]	27.5	51.2	199.3	3.2
Audio-lyric fusion [21]	28.4	53.1	201.3	3.2
AAI-HarmoCNN-AttnNet (proposed)	**19.6**	**46.8**	**159.2**	**4.6**

5. CONCLUSION AND FUTURE WORK

This work presents AAI-HarmoCNN-AttnNet, a privacy-preserving deep learning framework for music genre classification, which claims high predictive accuracy under decentralized training constraints. The proposed approach meets several real-

world challenges, such as genre overlap, class imbalance, and the demands for secure and distributed learning environments. The framework uses harmonic-aware convolutional processing and a dual-path attention mechanism to accurately capture fine-grained spectral cues and long-term temporal dependencies inherent in musical signals. Federated learning architecture enables collaborative model training across user devices while ensuring sensitive audio data remain locally preserved. To improve on convergence stability and better computational efficiency, a hybrid hyperparameter optimization strategy that integrates Egret Swarm Optimization (ESOA) with Golden Jackal Optimization (GJO) is adopted. Result analysis indicates that the proposed system attains an accuracy of 99.1% in classification and a genre diversity sensitivity (GDS) score of 97.4, both better than those established by thirteen other competitive baseline models. Additional results display the uniformity of performance across heterogeneous federated clients and suitability for resource-constrained edge deployments, thus indicating the robustness and scalability of such framework in real-world applications of music.

Future work will focus on integrating listener context and lyric-aware representations to enable personalized music recommendation in a manner that respects user privacy. Further investigations would involve lightweight deployment methods such as knowledge distillation and model compression to further up the real-time performance of mobile and edge devices. Also, the federated learning framework will be continued to capture more region-specific characteristics of genre by assessment on multilingual and culturally popular music datasets.

ACKNOWLEDGEMENTS

General Topics of Hunan Social Science Achievement Review Committee in 2025, A Study on the Path of "Four Beauties and Four Drives" Cultivation of Aesthetic Education Quality of Vocational College Teachers Based on OBE Concept (XSP25YBC687)

REFERENCES

[1] A. Bavarava and J.V. Sudarshan, "The impact of music on mood and emotion: A comprehensive analysis," *J. Adv. Res. Journal. Mass Commun.*, vol. 11, no. 1&2, pp. 12–21, 2024.

[2] P. Visutsak, J. Loungna, S. Sopromrat, C. Jantip, P. Soponkittikunchai, and X. Liu, "Mood-based music discovery: A system for generating personalized thai music playlists using emotion analysis," *Appl. Syst. Innov.*, vol. 8, no. 2, p. 37, 2025.

[3] J. Zhang, S. Yu, R. Liu, G.-X. Xie, and L. Zurawicki, "Unveiling the melodic matrix: exploring genre-and-audio dynamics in the digital music popularity using machine learning techniques," *Mark. Intell. Plan.*, vol. 42, no. 8, pp. 1333–1352, 2024, doi: 10.1108/MIP-04-2024-0209.

[4] E.J. Vella, C. McDonough, and H. Goldstein, "Musical mood induction: The relative influences of music type and the importance of music preference," *Psychol. Music*, p. 03057356241254361, 2024.

[5] C. Weiß and M. Müller, "From music scores to audio recordings: Deep pitch-class representations for measuring tonal structures," *ACM J. Comput. Cult. Herit.*, vol. 17, no. 3, pp. 1–19, 2024.

[6] T. Wang, J. Li, H. Wu, C. Li, H. Snoussi, and Y. Wu, "Reslnet: Deep residual LSTM network with longer input for action recognition," *Multimed. Syst.*, vol. 16, no. 6, p. 166334, 2022, doi: 10.1007/s11704-021-0236-9.

[7] Q. Hu, M.A.A. Murad, and Q. Li, "Advancing music emotion recognition: large-scale dataset construction and evaluator impact analysis," *Multimed. Syst.*, vol. 31, no. 2, pp. 1–16, 2025.

[8] A.M. Christodoulou, O. Lartillot, and A.R. Jensenius, "Multimodal music datasets? challenges and future goals in music processing," *Int. J. Multimed. Inf. Retr.*, vol. 13, no. 3, p. 37, 2024.

[9] C. Xie, H. Song, H. Zhu, K. Mi, Z. Li, and Y. Zhang, "Music genre classification based on res-gated cnn and attention mechanism," *Multimed. Tools Appl.*, vol. 83, no. 5, pp. 13 527–13 542, 2024.

[10] K. Singh and J. Rokde, "Film box office success forecasting: A genre-driven classification system using machine learning and cluster analysis," in *Proc. of the 3rd Int. Conf. on Applied Artificial Intelligence and Computing (ICAAIC)*. IEEE, Jun. 2024, pp. 486–492.

[11] K.M. Rezaul, M. Jewel, M.S. Islam, K.N.E.A. Siddiquee, N. Barua, and M.A. Rahman, "Enhancing audio classification through mfcc feature extraction and data augmentation with cnn and rnn models," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 7, pp. 37–53, 2024.

[12] K. Zhao, H. Wang, X. Wang, L. An, L. Chen, Y. Zhang, and L. Zhou, "Neutron-gamma discrimination method based on voiceprint identification," *Radiat. Meas.*, vol. 187, p. 107483, 2025, doi: 10.1016/j.radmeas.2025.107483.

[13] S.K. Swarnkar and Y.K. Rathore, "Music genre classification using long short-term memory (lstm) networks: Analyzing audio spectrograms for enhanced multimedia understanding," in *Machine Learning in Multimedia*. CRC Press, 2025, pp. 74–84.

[14] H.D.N. Nguyen *et al.*, "Exploring cnn-based architectures for vietnamese traditional music genre classification," in *2024 Int. Conf. on Control, Robotics and Informatics (ICCRI)*. IEEE, Jul. 2024, pp. 62–67.

[15] Z. Cao *et al.*, "Understanding the dimensional need of noncontrastive learning," *IEEE Trans. Cybern.*, vol. 55, no. 9, pp. 4089–4102, 2025, doi: 10.1109/TCYB.2025.3577745.

[16] Z. Wen, A. Chen, G. Zhou, J. Yi, and W. Peng, "Parallel attention of representation global time–frequency correlation for music genre classification," *Multimed. Tools Appl.*, vol. 83, no. 4, pp. 10 211–10 231, 2024.

[17] L. Yu, Y. Li, S. Weng, H. Tian, and J. Liu, "Adaptive multi-teacher softened relational knowledge distillation framework for payload mismatch in image steganalysis," *J. Vis. Commun. Image Represent.*, vol. 95, p. 103900, 2023, doi: 10.1016/j.jvcir.2023.103900.

[18] R. Tian, R. Yin, and F. Gan, "Music sentiment classification based on an optimized cnn-rf-qpso model," *Data Technol. Appl.*, vol. 57, no. 5, pp. 719–733, 2023.

[19] X. Zhang, M. Wang, X. Zeng, and X. Zhuang, "Af-can: Multimodal emotion recognition method based on situational attention

mechanism," *IEEE Access*, vol. 13, pp. 44 858–44 871, 2025, doi: 10.1109/ACCESS.2024.3471613.

[20] X. Gong, H. Duan, Y. Yang, L. Tan, J. Wang, and A.V. Vasilakos, "Improving audio classification method by combining self-supervision with knowledge distillation," *Electronics*, vol. 13, no. 1, p. 52, 2023.

[21] Y. Li, Z. Zhang, H. Ding, and L. Chang, "Music genre classification based on fusing audio and lyric information," *Multimed. Tools Appl.*, vol. 82, no. 13, pp. 20 157–20 176, 2023.

[22] A. Anatoly, "Harmogenre30m [data set]," 2025, doi: 10.34740/KAGGLE/DSV/11520513.

[23] H. Cao, D. Chen, Y. Zhang, H. Zhou, D. Wen, and C. Cao, "MFINet: A multi-scale feature interaction network for point cloud registration," *Vis. Comput.*, vol. 41, no. 6, pp. 4067–4079, 2025, doi: 10.1007/s00371-024-03646-2.

[24] M. Ulicny, V.A. Krylov, and R. Dahyot, "Harmonic convolutional networks based on discrete cosine transform," *Pattern Recognit.*, vol. 129, p. 108707, 2022.

[25] W.C. Bown, "Sensitivity and specificity versus precision and recall, and related dilemmas," *J. Classif.*, vol. 41, no. 2, pp. 402–426, 2024.