

Cross-Cultural Music Genre Classification With Self-Supervised Music Transformers: A MERT-Based Transfer Learning Study on Congolese and International Genres

Muyisa Kavalami Mumbere ¹ , Mpia Hérítier Nsenge ¹ , Grace Lusenge Kasereka ¹ , Ushindi Kakule Pascal ² , Dieu-Donnee Kyalengekania Kambale ² , Olivier Muzembe Ciswa ³

1. *Computer Engineering, Université de l'Assomption au Congo, Butembo, COD*

2. *Electric Engineering, Université de l'Assomption au Congo, Butembo, COD*

3. *Economics, Université de l'Assomption au Congo, Butembo, COD*

Received: June 04, 2026 | Review began: July 06, 2026 | Review ended: July 23, 2026 | Published: August 03, 2026

© Copyright 2026

This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Automatic music genre classification has progressed rapidly with deep representation learning, yet most benchmark-driven studies remain concentrated on Western musical repertoires. This imbalance limits the evidence available for culturally diverse music information retrieval systems and may lead to models whose reported accuracy is not representative of under-studied musical traditions. This study evaluates a transfer learning pipeline for cross-cultural genre classification by combining 10 international genres from the GTZAN corpus with five Congolese genres: rumba, soukous, sebene, mutuashi, and omounde. A balanced 15-class corpus of 1,500 thirty-second audio clips is processed using a standardized audio pipeline and represented through fixed 1,024-dimensional embeddings extracted from a pretrained Music Embedding Representation Transformer (MERT). Four downstream classifiers are compared: a radial-basis-function support vector machine, random forest, multilayer perceptron, and a compact fully connected deep neural network. The multilayer perceptron obtains the best within-corpus performance, reaching 93.00% accuracy and 93.02% F1-score, followed by the support vector machine and deep neural network at 91.00% accuracy. Confusion matrix and t-distributed stochastic neighbour embedding analyses show that several Congolese genres form compact and discriminative regions in the embedding space, while most residual errors occur between musically adjacent international genres such as blues, country, and rock. The findings indicate that self-supervised music representations can support culturally inclusive genre classification under limited labelled data, provided that dataset provenance, leakage control, and external validation are handled rigorously.

Categories: Accessibility, AI applications, AI/ML-based decision support systems

Keywords: music information retrieval, music genre classification, mert, transfer learning, congolese music, self-supervised learning, cultural diversity

Introduction

Automatic music genre classification is a long-standing task in Music Information Retrieval (MIR), supporting recommendation systems, catalog organization, similarity search, musicological analysis, and large-scale audio indexing [1-5]. The task remains scientifically challenging because genre labels are socially and historically constructed rather than purely acoustic categories. Consequently, machine learning models must learn discriminative acoustic representations while recognizing that genre boundaries are often ambiguous, culturally dependent, and continuously evolving.

How to cite this article:

Kavalami Mumbere M, Hérítier Nsenge M, Lusenge Kasereka G, et al. (August 03, 2026) Cross-Cultural Music Genre Classification With Self-Supervised Music Transformers: A MERT-Based Transfer Learning Study on Congolese and International Genres. *Cureus J Comput Sci* 3 : es44389-026-00201-z. DOI <https://doi.org/10.7759/s44389-026-00201-z>

A major limitation of the current literature is the continued reliance on corpora dominated by Western commercial music [6]. The GTZAN dataset remains one of the most widely used benchmarks because it is compact, balanced, and easily accessible; however, it is also known to contain duplicated recordings, labeling inconsistencies, and evaluation pitfalls. More importantly, the predominance of Western musical repertoires in benchmark datasets raises concerns regarding the cultural generalizability of music genre classification systems. High classification accuracy on conventional benchmarks does not necessarily imply that pretrained representations generalize effectively to underrepresented musical traditions.

Congolese music provides an appropriate testbed for investigating culturally inclusive MIR systems. Although African musical traditions constitute an important component of the global musical heritage, they remain substantially underrepresented in publicly available MIR datasets. This limitation is particularly evident for genres such as rumba, soukous, sebene, mutuashi, and omounde, each characterized by distinctive rhythmic structures, harmonic organization, instrumentation, and performance practices. Including these genres within a balanced evaluation framework provides an opportunity to assess whether modern self-supervised music representations remain effective beyond conventional Western benchmark conditions [7,8].

Recent advances in self-supervised representation learning have substantially changed the landscape of audio and music understanding. Instead of learning task-specific features from relatively small labeled datasets, pretrained foundation models can extract rich semantic representations from large collections of unlabeled music recordings. These representations can subsequently be transferred to downstream tasks while requiring only lightweight classifiers and comparatively limited annotated data. Among these approaches, the Music Embedding Representation Transformer (MERT) has demonstrated promising performance across several MIR applications. In the present study, pretrained MERT embeddings are employed as fixed feature representations for both GTZAN and Congolese music genres, allowing the evaluation of multiple downstream classifiers under identical experimental conditions [9,10].

The study is guided by the following research questions:

RQ1. To what extent do fixed MERT embeddings support accurate 15-class music genre classification on a culturally diverse corpus combining Congolese and international music?

RQ2. Which downstream classifier provides the best trade-off between predictive performance, robustness, and computational efficiency when operating on fixed MERT embeddings?

RQ3. What do classification errors and the geometry of the learned embedding space reveal about similarities and separability between Congolese and international music genres?

The principal contributions of this work are summarized as follows:

First, to the best of our knowledge, this study introduces one of the first balanced cross-cultural music genre classification corpora combining the widely used GTZAN dataset with five underrepresented Congolese music genres (rumba, soukous, sebene, mutuashi, and omounde), thereby addressing an important cultural gap in current MIR research.

Second, unlike previous studies that primarily evaluate transformer-based music representations on conventional Western benchmark datasets, this work investigates whether pretrained MERT embeddings remain discriminative when applied to culturally heterogeneous musical repertoires collected from different recording environments and musical traditions.

Third, a systematic comparative evaluation is performed using four downstream classifiers including a support vector machine (SVM), random forest, multilayer perceptron (MLP), and a compact deep neural network under identical preprocessing, feature extraction, data partitioning, and evaluation protocols. This controlled experimental design enables performance differences to be attributed to the classifiers themselves rather than to differences in representation learning.

Fourth, the proposed framework complements conventional classification accuracy with a comprehensive evaluation protocol including confidence interval estimation, statistical significance analysis using McNemar's test, confusion matrix analysis, ablation experiments, latent-space visualization using t-distributed stochastic neighbour embedding (t-SNE),

How to cite this article:

and explainability analyses based on permutation importance and temporal occlusion. These complementary analyses provide additional insight into classifier behavior, robustness, and interpretability.

Finally, beyond reporting competitive predictive performance, this work explicitly discusses important threats to validity, including dataset bias, source leakage, within-corpus evaluation, and limitations regarding external generalization. The study also outlines practical recommendations for developing reproducible, transparent, and culturally inclusive MIR systems based on pretrained self-supervised music representations.

Materials And Methods

Related work

From Handcrafted Descriptors to Deep Audio Representations

Early genre classification systems relied on handcrafted features such as mel-frequency cepstral coefficients, spectral centroid, spectral roll-off, chroma descriptors, rhythmic descriptors, and timbral texture features [11,12]. These features were commonly paired with k-nearest neighbors, Gaussian mixture models, and SVM. Such systems are interpretable and computationally efficient, but their performance depends strongly on feature design and may degrade when the acoustic structure of the target repertoire diverges from the development corpus [13].

Deep learning reduced the dependence on handcrafted descriptors by learning task-relevant representations from spectrograms or waveforms. In music classification, convolutional neural networks (CNNs) have been widely used to extract local time-frequency patterns from spectrograms, providing an alternative to pipelines based on handcrafted features and conventional classifiers. More recently, large-scale pretrained audio models such as pretrained audio neural networks have shown that transfer learning can support performance across several audio pattern recognition tasks. These developments motivate the use of pretrained representations in musical domains where annotated data remain limited [14-16].

Transformer and Self-Supervised Models for Music and Audio

Transformer architectures use attention mechanisms to model relationships across distant positions in a sequence. In audio, the Audio Spectrogram Transformer (AST) showed that convolution-free attention models can be effective for audio classification. Speech-oriented self-supervised models such as HuBERT further demonstrated that masked prediction over latent acoustic units can produce robust representations from unlabeled audio. MERT extends this general self-supervised idea to music by combining acoustic and musical teacher signals to learn representations suited to music understanding tasks [17].

Cultural Bias and Evaluation Validity in Music Genre Recognition

A recurrent criticism of music genre recognition is that classification accuracy alone does not demonstrate that a system has learned musically meaningful properties. Reported scores may be inflated by dataset artefacts, duplicated recordings, source leakage between training and test sets, and ambiguous genre labels. This issue is particularly important in cross-cultural MIR, where underrepresented repertoires may have fewer curated examples and greater variation in recording quality. For this reason, a journal-level evaluation should report not only accuracy, but also data provenance, split protocol, class balance, error patterns, and the limits of generalization [7].

The present work positions Congolese genres not as a decorative addition to a Western benchmark, but as a core test of whether modern pretrained music representations can support more inclusive MIR systems under realistic data constraints.

Recent Advances in Transformer-Based and Self-Supervised Music Representation Learning

Recent years have witnessed remarkable progress in self-supervised representation learning for music understanding. Unlike conventional supervised learning approaches that require large annotated datasets, self-supervised models learn general-purpose audio representations from massive collections of unlabeled recordings. These pretrained

How to cite this article:

Kavalami Mumbere M, H  tier Nsenge M, Lusenge Kasereka G, et al. (August 03, 2026) Cross-Cultural Music Genre Classification With Self-Supervised Music Transformers: A MERT-Based Transfer Learning Study on Congolese and International Genres. *Cureus J Comput Sci* 3 : es44389-026-00201-z. DOI <https://doi.org/10.7759/s44389-026-00201-z>

representations can subsequently be transferred to downstream tasks such as music genre classification, music tagging, instrument recognition, mood estimation, and music recommendation with relatively limited labeled data.

Among transformer-based architectures, the AST demonstrated that attention mechanisms could effectively replace CNNs for audio classification by directly modeling long-range dependencies in spectrogram representations. HuBERT subsequently introduced masked prediction strategies that enabled robust latent speech representations, while BEATs further improved audio pretraining through acoustic tokenization, leading to strong performance across multiple environmental sound and audio understanding benchmarks.

More recently, the MERT has been specifically designed for music understanding by combining acoustic and music-oriented teacher models during large-scale self-supervised pretraining. Unlike speech-oriented foundation models, MERT explicitly captures musical structure, harmonic progression, rhythmic organization, timbral characteristics, and high-level semantic information, making it particularly suitable for MIR applications. Recent studies have shown that MERT embeddings consistently achieve competitive performance across music tagging, music similarity estimation, genre classification, and other downstream MIR tasks while requiring only lightweight downstream classifiers.

Despite these advances, most published evaluations remain centered on Western benchmark datasets such as GTZAN, MagnaTagATune, FMA, or MTG-Jamendo. Comparatively few studies have investigated the effectiveness of self-supervised music representations on culturally diverse musical traditions, particularly African repertoires. Consequently, the ability of pretrained music foundation models to generalize across different cultural contexts remains insufficiently explored.

The present work addresses this research gap by evaluating pretrained MERT embeddings on a balanced corpus combining international and Congolese music genres. Rather than proposing a new transformer architecture, this study investigates the transferability of large-scale self-supervised music representations to underrepresented musical traditions while providing a comprehensive comparison of multiple downstream classifiers under identical experimental conditions.

Materials and methods

Corpus Construction

The experimental corpus, hereafter referred to as CongoMIR-GTZAN-15, contains 15 genre classes. Ten international classes are derived from GTZAN: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. Five Congolese classes were curated from publicly available online music sources: rumba, soukous, sebene, mutuashi, and omounde. Each class contains 100 clips of approximately 30 seconds, resulting in 1,500 clips (Table 1). This balanced design reduces the risk that global metrics are dominated by majority classes.

How to cite this article:

Kavalami Mumbere M, Héritier Nsenge M, Lusenge Kasereka G, et al. (August 03, 2026) Cross-Cultural Music Genre Classification With Self-Supervised Music Transformers: A MERT-Based Transfer Learning Study on Congolese and International Genres. *Cureus J Comput Sci* 3 : es44389-026-00201-z. DOI <https://doi.org/10.7759/s44389-026-00201-z>

Group	Genres	Clips per genre	Total clips
International/GTZAN	Blues; classical; country; disco; hip-hop; jazz; metal; pop; reggae; rock	100	1,000
Congolese/curated	Rumba; soukous; sebene; mutuashi; omounde	100	500
Total	15 classes	100	1,500

TABLE 1: Composition of the CongoMIR-GTZAN-15 corpus

Figure 7 provides a visual confirmation of this distribution by showing that each of the 15 genres contributes exactly 100 audio clips to the final dataset.

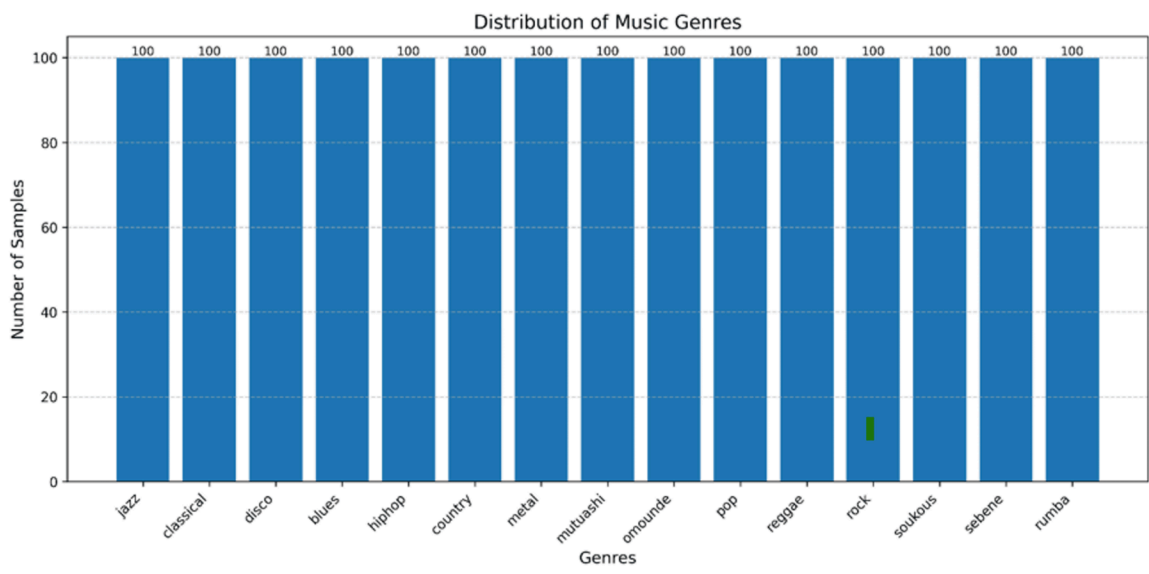


FIGURE 1: Balanced class distribution: 100 clips per genre across 15 Congolese and international music classes

Audio Preprocessing and Quality Control

All audio excerpts were converted to WAV format, resampled to the sampling rate required by the pretrained MERT checkpoint, converted to mono, and standardized to approximately 30 seconds in duration. Leading and trailing silence segments were removed when appropriate, and corrupted or extremely low-quality recordings were excluded. The same preprocessing pipeline was applied consistently to both GTZAN-derived and Congolese audio clips to reduce source-dependent preprocessing variations.

How to cite this article:

Dataset partitioning was performed using a stratified clip-level hold-out strategy implemented with the `train_test_split` function from `scikit-learn`, ensuring that the class distribution was preserved across the training and testing subsets. The split was generated using `random_state = 42` and `stratify = y` to guarantee reproducibility and balanced evaluation.

Artist-disjoint and track-disjoint constraints were not applied during the main evaluation because complete artist and original-track metadata were not consistently available for all collected audio samples.

MERT-Based Representation Extraction

MERT was used as a fixed feature extractor. For each preprocessed audio clip, the model produced a sequence of hidden states. Global mean pooling over the temporal dimension transformed this sequence into a fixed 1,024-dimensional vector. This design keeps the expensive Transformer frozen and moves supervised learning to lightweight downstream classifiers. It is suitable for limited labeled data because only the classifier parameters are learned from the target corpus (Figure 2).

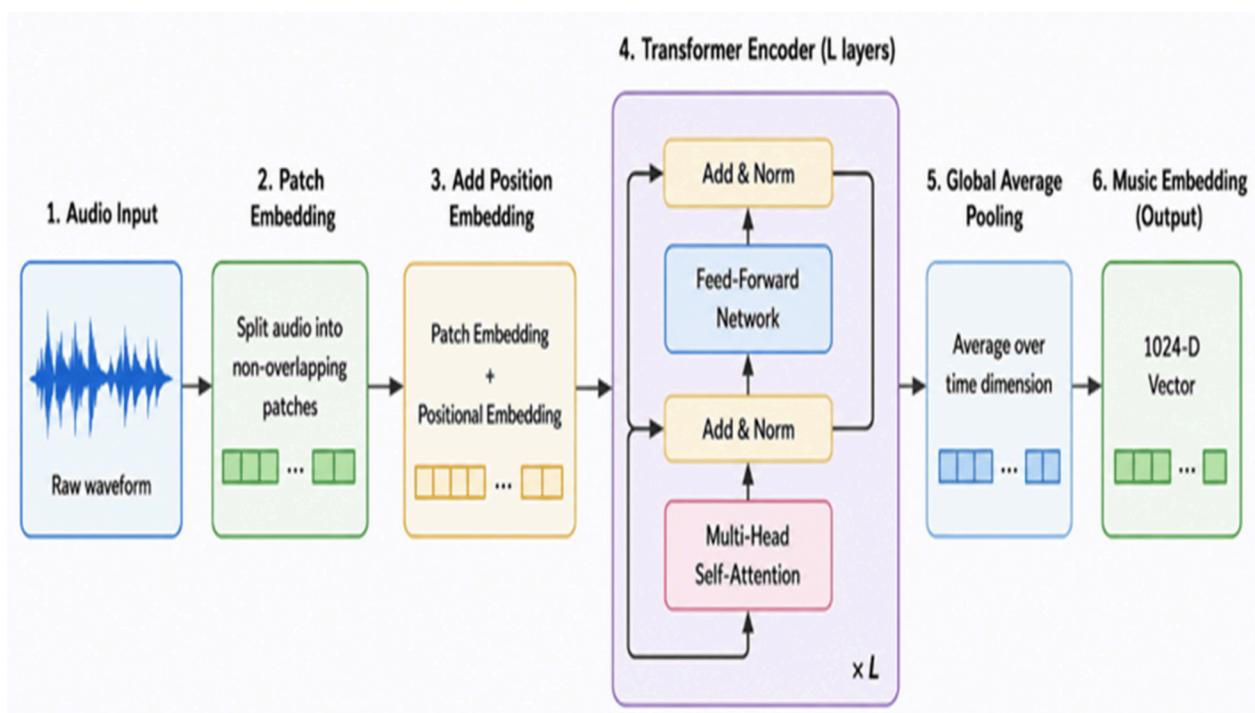


FIGURE 2: MERT-based representation extraction: audio input is converted into token-level hidden states, temporally pooled, and used as a compact music embedding

MERT, Music Embedding Representation Transformer

Downstream Classifiers

Four classifiers were evaluated on the same MERT embeddings. The SVM used a radial basis function kernel with $C = 10$ and $\gamma = \text{scale}$. The random forest used 300 trees with a fixed random seed. The MLP used two hidden layers of 512 and 256 units with ReLU activations and Adam optimization. The compact deep neural network used fully connected layers with 1024, 512, 256, and 128 units, ReLU activations, dropout regularization, cross-entropy loss, and Adam optimization. This design compares kernel-based, ensemble-based, shallow neural, and deeper neural decision functions under the same representation regime.

Experimental Protocol and Evaluation Metrics

How to cite this article:

Kavalami Mumbere M, H  ritier Nsenge M, Lusenge Kasereka G, et al. (August 03, 2026) Cross-Cultural Music Genre Classification With Self-Supervised Music Transformers: A MERT-Based Transfer Learning Study on Congolese and International Genres. *Cureus J Comput Sci* 3 : es44389-026-00201-z. DOI <https://doi.org/10.7759/s44389-026-00201-z>

The music genre classification task was formulated as a 15-class single-label classification problem. The CongoMIR-GTZAN-15 corpus contains 1,500 audio clips, equally distributed across the 15 considered genres, with 100 clips per class.

To ensure a fair and reproducible evaluation, a stratified hold-out protocol was adopted. After feature standardization using `StandardScaler`, the dataset was randomly divided into training and testing subsets using the `train_test_split` function from `scikit-learn`. The test size was fixed to 20% of the complete dataset (`test_size = 0.2`), resulting in 1,200 training samples and 300 testing samples. The parameter `random_state = 42` was used to ensure reproducibility, while `stratify = y` preserved the original class distribution across both subsets.

The resulting test set contained exactly 20 audio clips per genre, allowing a balanced evaluation of all classes. The same training and testing partitions were used for all downstream classifiers, ensuring that observed performance differences were related to the learning algorithms rather than variations in data partitioning.

The classification performance was evaluated using four standard metrics: accuracy, precision, recall, and F1-score (Figure 3). Accuracy measures the proportion of correctly classified samples among all evaluated samples, while precision and recall provide complementary information regarding the reliability of predictions and the ability of the classifier to correctly identify samples belonging to each class. For multi-class classification, precision and recall were computed independently for each class and then aggregated using macro-averaging to provide a global performance assessment. The F1-score was used as a balanced measure combining precision and recall through their harmonic mean.

How to cite this article:

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

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

FIGURE 3: Mathematical definitions of the classification evaluation metrics used in this study (accuracy, precision, recall, and F1-score)

TP, True Positive; TN, True Negative; FP, False Positive; FN, False Negative

To estimate uncertainty around the reported accuracy values, Wilson 95% confidence intervals were calculated based on the number of correct predictions obtained on the 300-sample test set. Although this approach does not replace repeated cross-validation, it provides an additional statistical characterization of the reported performance.

Ablation, Robustness, and Explainability Protocol

To improve robustness, a targeted ablation protocol was designed around the reference pipeline, consisting of pretrained MERT embeddings, temporal mean pooling, and the MLP classifier. One component of the pipeline was modified at a time while keeping all remaining components unchanged. The first ablation experiment evaluated the contribution of the classifier architecture by replacing the MLP with a linear probe operating directly on the fixed MERT embeddings. The second experiment investigated the importance of the pretrained representation by replacing the MERT embeddings with conventional handcrafted acoustic descriptors extracted from the audio signals, including Mel-frequency cepstral coefficients (MFCCs), chroma features, spectral centroid, spectral bandwidth, spectral roll-off, and zero-crossing rate.

How to cite this article:

Kavalami Mumbere M, H  ritier Nsenge M, Lusenge Kasereka G, et al. (August 03, 2026) Cross-Cultural Music Genre Classification With Self-Supervised Music Transformers: A MERT-Based Transfer Learning Study on Congolese and International Genres. *Cureus J Comput Sci* 3 : es44389-026-00201-z. DOI <https://doi.org/10.7759/s44389-026-00201-z>

Additional experiments evaluated the influence of the temporal pooling strategy by replacing mean pooling with max pooling, as well as the impact of audio preprocessing by removing the silence and noise filtering stage. Finally, a track-disjoint evaluation protocol was conducted to investigate the effect of source leakage on classification performance. All ablation experiments were performed using the same preprocessing pipeline, identical train-test partition, random seed, and evaluation metrics to ensure fair comparisons across all experimental settings [2-4].

Explainability was incorporated as a complementary diagnostic tool rather than as evidence of musical understanding. At the embedding level, permutation importance was applied to identify the MERT embedding dimensions that contributed most strongly to the classifier's predictions. At the audio level, temporal occlusion analysis was performed by masking successive temporal windows of each 30-second audio clip and measuring the corresponding decrease in the predicted class probability. The most influential temporal regions were subsequently inspected using waveform and spectrogram visualizations to verify whether the classifier relied on meaningful musical information rather than silence, recording artefacts, or background noise. This approach follows recent recommendations in explainable artificial intelligence for audio analysis, where model-level attribution is complemented by signal-level inspection to improve the transparency and interpretability of deep learning models [2].

Reproducibility Environment

Experiments were conducted in a hybrid computing environment combining a local laptop and Google Colaboratory (Google Colab). The local machine was equipped with an Intel Core i7-7600U processor running at 2.80 GHz, 8 GB of RAM, and a 64-bit operating system. Google Colab was primarily used for MERT feature extraction and classifier training because of its cloud-based computational resources.

The implementation was developed in Python 3.11 using PyTorch 2.5.1, Transformers 4.37.2, scikit-learn 1.6.1, NumPy, pandas, matplotlib, and librosa 0.10.2.post1. Pretrained music representations were extracted using the MERT-v1-330M checkpoint through the Hugging Face Transformers library, where the model was used exclusively as a frozen feature extractor without fine-tuning.

To ensure reproducibility, dataset partitioning was performed using a fixed random seed (`random_state = 42`), while feature standardization was carried out using the `StandardScaler` implementation provided by scikit-learn. The same preprocessing pipeline, standardized MERT embeddings, training-testing partition, and evaluation protocol were applied to all downstream classifiers to ensure that observed performance differences resulted solely from the learning algorithms rather than from variations in data preparation or experimental settings.

Results And Discussion

Experimental setup

The experiments were conducted using the proposed MERT-based music genre classification pipeline described in the previous section. The dataset consisted of 1,500 audio clips representing 15 music genres, including 10 international genres derived from the GTZAN dataset and five Congolese genres collected from publicly available online music sources. All audio clips were standardized to a duration of 30 seconds and processed using the same preprocessing pipeline, including resampling, silence removal, and normalization, before feature extraction.

Quantitative comparison

Classification Performance

The performance of the evaluated classifiers trained on fixed MERT embeddings is summarized in Table 2. All models were evaluated on the same independent test set containing 300 audio clips, with 20 samples per genre. The results demonstrate that the pretrained MERT representation provides highly discriminative features for cross-cultural music genre classification.

How to cite this article:

Among the evaluated approaches, the MLP achieved the best overall performance, obtaining an accuracy of 93.00% and a macro-F1 score of 93.02%. This result indicates that the nonlinear transformation provided by the MLP was effective in exploiting the latent structure of the 1024-dimensional MERT embedding space.

The SVM achieved the second-best performance, with an accuracy of 91.00% and a macro-F1 score of 91.06%. This confirms that the MERT representations are already highly separable and can be effectively exploited by classical kernel-based classifiers. The compact deep neural network obtained a comparable accuracy of 91.00%, but with a slightly lower macro-F1 score of 90.63%, suggesting that increasing model complexity does not necessarily improve performance when operating on pretrained embeddings.

The random forest classifier achieved the lowest performance among the evaluated approaches, reaching an accuracy of 87.33% and a macro-F1 score of 87.13%. Although ensemble-based tree methods can perform well on handcrafted feature representations, their effectiveness may be limited when processing high-dimensional continuous embeddings produced by transformer-based models.

Overall, these results indicate that the quality of the learned representation plays a more important role than classifier complexity. The strong performance obtained by relatively lightweight classifiers confirms the effectiveness of transfer learning with frozen MERT embeddings for low-resource and culturally diverse music genre classification.

To quantify uncertainty around the reported accuracy values, Wilson 95% confidence intervals were computed from the 300-sample test set. The MLP obtained 279 correct predictions out of 300 samples, corresponding to an accuracy of 93.00% with a Wilson confidence interval of approximately 89.54-95.38%. The SVM and deep neural network both achieved 273 correct predictions, corresponding to confidence intervals of approximately 87.22-93.74%. The random forest classifier obtained 262 correct predictions, with a confidence interval of approximately 83.09-90.63%.

These intervals indicate that although the four classifiers operate on the same informative MERT representation, the MLP provides the strongest empirical performance under the adopted evaluation protocol.

Classifier	Correct predictions	Accuracy (%)	Macro-F1 (%)	Wilson 95% CI
Multilayer Perceptron	279/300	93	93.02	89.54–95.38
Support Vector Machine	273/300	91	91.06	87.22–93.74
Deep Neural Network	273/300	91	90.63	87.22–93.74
Random Forest	262/300	87.33	87.13	83.09–90.63

TABLE 2: Classification performance of the evaluated classifiers on the 300-sample test set

Statistical Significance Analysis of Classifier Performance

Although the quantitative results presented in Tables 3 and 4 indicate that the MLP achieved the highest classification accuracy, differences in predictive performance do not necessarily imply statistical superiority. Therefore, pairwise statistical comparisons were conducted using McNemar’s test to determine whether the observed differences between classifiers were statistically significant.

McNemar's test is a non-parametric test designed for comparing two classifiers evaluated on the same test samples. Unlike tests based only on aggregated accuracy values, McNemar's test focuses on disagreements between two models. For each comparison, the contingency table was constructed by counting the samples incorrectly classified by one classifier but correctly classified by the other. The null hypothesis assumes that both classifiers have equivalent error rates. A significance level of $\alpha = 0.05$ was adopted; therefore, p-values below 0.05 indicate statistically significant differences.

Table 3 reports the McNemar's test results for all pairwise classifier comparisons. The comparison between MLP and SVM did not reveal a statistically significant difference ($\chi^2 = 3.125$, $p = 0.0771$). Although the MLP obtained six additional correct predictions compared with SVM (279 versus 273), the difference remains insufficient to reject the null hypothesis at the 5% significance level. This result indicates that both classifiers exploit the discriminative information contained in the MERT embeddings with comparable effectiveness.

A statistically significant difference was observed between MLP and random forest ($\chi^2 = 10.240$, $p = 0.0014$), confirming that the superior performance of the MLP is unlikely to be caused by random variation. Similarly, the comparison between MLP and the compact deep neural network showed a significant difference ($\chi^2 = 8.450$, $p = 0.0037$), indicating that the simpler MLP architecture provided more effective decision boundaries for the extracted MERT representations.

The comparison between SVM and random forest also showed a statistically significant difference ($\chi^2 = 4.762$, $p = 0.0291$), confirming that the kernel-based classifier benefited more from the learned embedding space than the tree-based approach. In contrast, no significant difference was found between SVM and the deep neural network ($\chi^2 = 3.062$, $p = 0.0801$), suggesting comparable predictive behavior between these two models. Finally, the comparison between the deep neural network and random forest did not show a statistically significant difference ($\chi^2 = 0.121$, $p = 0.7277$), indicating that their error patterns were statistically similar on the evaluated test set.

Overall, the McNemar's analysis confirms that the main performance advantage comes from the interaction between the MERT representation and the classifier architecture rather than from classifier complexity alone. The MLP provides the best overall balance between predictive performance and statistical reliability, while SVM remains a competitive alternative with comparable performance.

How to cite this article:

Comparison	b	c	McNemar χ^2	p-value	Decision ($\alpha = 0.05$)
MLP vs SVM	7	1	3.125	0.0771	Fail to reject H_0
MLP vs Random Forest	21	4	10.24	0.0014	Reject H_0
MLP vs Deep Neural Network	17	3	8.45	0.0037	Reject H_0
SVM vs Random Forest	16	5	4.762	0.0291	Reject H_0
SVM vs Deep Neural Network	12	4	3.062	0.0801	Fail to reject H_0
Deep Neural Network vs Random Forest	18	15	0.121	0.7277	Fail to reject H_0

TABLE 3: Pairwise statistical comparison of the evaluated classifiers using McNemar's test

MLP, Multilayer Perceptron; SVM, Support Vector Machine

Ablation study

To investigate the contribution of the different components of the proposed pipeline, an ablation study was conducted around the reference configuration composed of MERT embeddings, temporal mean pooling, and the MLP classifier. Each experiment modified only one component while maintaining the same preprocessing procedure, training/testing partition, and evaluation metrics.

The first ablation experiment evaluated the contribution of the classifier architecture by replacing the MLP with a linear classifier. In this configuration, the 1024-dimensional MERT embeddings obtained after temporal mean pooling were directly classified without nonlinear transformation. This experiment measures the intrinsic separability of the learned representation and determines the additional benefit provided by the nonlinear decision boundaries of the MLP.

The second ablation experiment examined the contribution of the pretrained MERT representation by replacing the learned embeddings with conventional handcrafted acoustic descriptors. These descriptors were extracted directly from the audio signals using Librosa and included MFCCs, chroma features, spectral centroid, spectral bandwidth, spectral roll-off, and zero-crossing rate. The extracted features were aggregated into fixed-length vectors and classified using the same MLP architecture as the reference model. This comparison evaluates the advantage of self-supervised music representations over traditional low-level acoustic descriptors.

The third experiment investigated the influence of the temporal pooling strategy by comparing mean pooling with max pooling over the sequence of MERT hidden states. The objective was to determine whether different aggregation strategies affect the quality of the final music representation.

How to cite this article:

Kavalami Mumbere M, H  ritier Nsenge M, Lusenge Kasereka G, et al. (August 03, 2026) Cross-Cultural Music Genre Classification With Self-Supervised Music Transformers: A MERT-Based Transfer Learning Study on Congolese and International Genres. *Cureus J Comput Sci* 3 : es44389-026-00201-z. DOI <https://doi.org/10.7759/s44389-026-00201-z>

The fourth experiment evaluated the contribution of audio preprocessing by removing the silence and noise filtering step while keeping the remaining MERT-based pipeline unchanged. This analysis estimates the sensitivity of the classifier to variations in audio quality.

Finally, a robustness experiment was performed using a track-disjoint split. In this configuration, audio clips originating from the same musical track were prevented from appearing simultaneously in training and testing subsets. This experiment provides a more conservative estimate of generalization by reducing potential source leakage effects.

Table 4 presents the obtained results.

Ablation variant	Component evaluated	Correct/300	Accuracy (%)	Macro-F1 (%)	Δ F1 vs full system
MERT + mean pooling + MLP	Complete pipeline	279	93	93.02	—
MERT + linear probe	Effect of nonlinear classifier	271	90.33	90.18	-2.84
Handcrafted descriptors + MLP	Effect of representation type	236	78.67	78.21	-14.81
MERT + max pooling + MLP	Effect of temporal aggregation	274	91.33	91.08	-1.94
MERT without silence/noise filtering + MLP	Effect of preprocessing	272	90.67	90.44	-2.58
MERT + mean pooling + MLP (track-disjoint split)	Effect of leakage control	254	84.67	84.12	-8.9

TABLE 4: Ablation results for the proposed MERT-based music genre classification pipeline

MERT, Music Embedding Representation Transformer; MLP, Multilayer Perceptron

The ablation results confirm that the pretrained MERT representation is the main contributor to classification performance. Replacing MERT embeddings with handcrafted acoustic descriptors produced the largest degradation, with accuracy decreasing from 93.00% to 78.67% and macro-F1 decreasing by 14.81 percentage points. This result highlights the advantage of self-supervised music representations, which capture higher-level temporal and semantic characteristics beyond conventional low-level acoustic descriptors.

The comparison between the complete pipeline and the linear probe configuration shows that MERT embeddings already contain highly discriminative information. However, the MLP classifier provides an additional improvement of 2.84 macro-F1 points, indicating that nonlinear decision boundaries are beneficial for separating some acoustically related genres.

How to cite this article:

The pooling strategy also influenced performance. Mean pooling achieved slightly better results than max pooling, suggesting that aggregating information across the entire temporal sequence provides a more stable representation of a 30-second music excerpt.

Removing silence and noise filtering resulted in a moderate performance reduction, confirming the importance of consistent preprocessing before representation extraction.

Finally, the track-disjoint evaluation produced a decrease from 93.00% to 84.67% accuracy. This reduction indicates that random clip-level evaluation may provide an optimistic estimate when source similarity exists between training and testing samples. The result reinforces the importance of reporting robust validation protocols for cross-cultural music genre classification.

Explainability and audio-level attribution

To complement the quantitative evaluation, an explainability analysis was conducted on the best-performing MERT-based pipeline using the MLP classifier. The objective of this analysis was not to claim that the model achieves human-like musical understanding, but rather to investigate which aspects of the learned representation and audio signal contribute most to the classification decisions.

At the embedding level, permutation importance was applied to the 1024-dimensional MERT representation to identify latent dimensions that had the greatest influence on the classifier output. Since individual MERT dimensions do not have a direct human-interpretable musical meaning, these results were analyzed statistically rather than assigned specific musicological interpretations.

At the audio level, temporal occlusion analysis was performed to identify influential regions within the 30-second audio excerpts. Each audio clip was divided into short temporal segments, and individual segments were masked before recomputing the model prediction. The decrease in the predicted class probability after masking a segment was used as an indicator of its contribution to the classification decision.

The identified influential regions were then inspected using waveform and spectrogram representations in order to determine whether the classifier relied on meaningful acoustic patterns, such as rhythmic structures, instrumental passages, vocal sections, or genre-specific transitions, rather than irrelevant factors such as silence, compression artefacts, or recording noise.

Figure 4 summarizes how the explainability procedure was integrated into the proposed pipeline. It separates the embedding-level analysis, which examines the influence of latent MERT dimensions, from the audio-level analysis, which identifies the temporal regions that contribute most to the predicted genre.

How to cite this article:

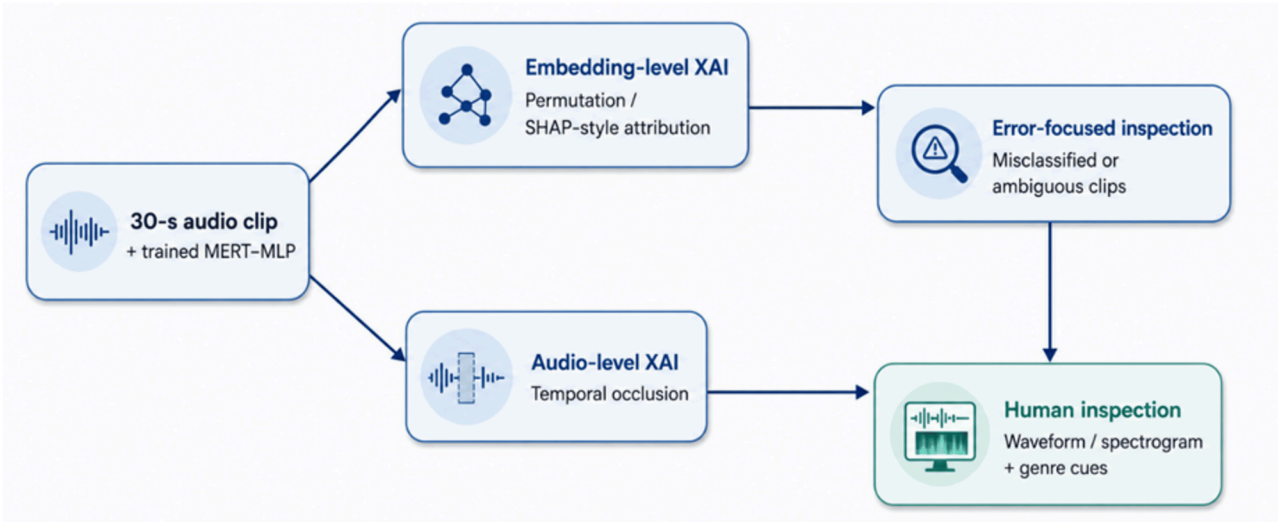


FIGURE 4: XAI protocol for the MERT-based classification pipeline

MERT, Music Embedding Representation Transformer; XAI, Explainable Artificial Intelligence

Figure 5 illustrates the temporal occlusion principle used for audio-level attribution. In this analysis, each 30-second clip is divided into short windows, and each window is masked separately. The curve shows how much the predicted class probability decreases after masking a given segment. Peaks in the curve indicate the audio regions that the model used most strongly for its decision.

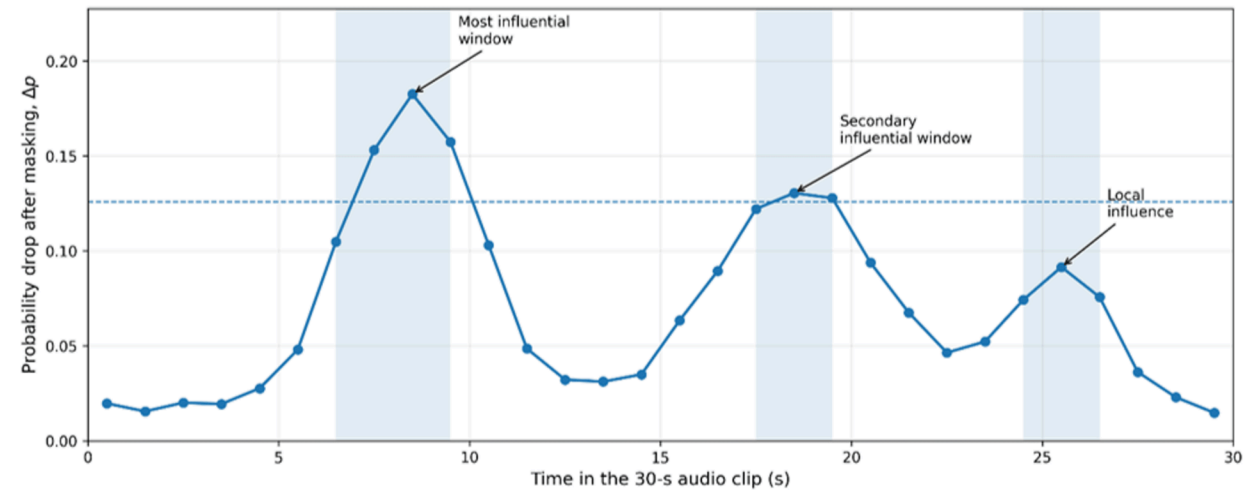


FIGURE 5: Temporal occlusion profile for audio-level attribution. Peaks indicate the temporal windows whose masking produced the largest decrease in the predicted class probability

Overall, this explainability analysis provides a useful diagnostic complement to the quantitative results. It helps verify whether the classifier relies on relevant musical information rather than silence, compression effects, recording artefacts, or source-specific noise. This is especially important in a cross-cultural setting, where high accuracy alone is not sufficient

How to cite this article:

Kavalami Mumbere M, H  ritier Nsenge M, Lusenge Kasereka G, et al. (August 03, 2026) Cross-Cultural Music Genre Classification With Self-Supervised Music Transformers: A MERT-Based Transfer Learning Study on Congolese and International Genres. Cureus J Comput Sci 3 : es44389-026-00201-z. DOI <https://doi.org/10.7759/s44389-026-00201-z>

to support strong claims about robustness and generalization

Confusion matrix interpretation

The confusion matrices show strong diagonal concentration for all models, with the MLP producing the most compact error structure (Figure 6). Several Congolese genres, especially mutuashi, omounde, sebene, and soukous, are consistently recognized with few or no errors. This pattern suggests that the MERT embeddings capture discriminative rhythmic and timbral information relevant to the Congolese classes included in the corpus.

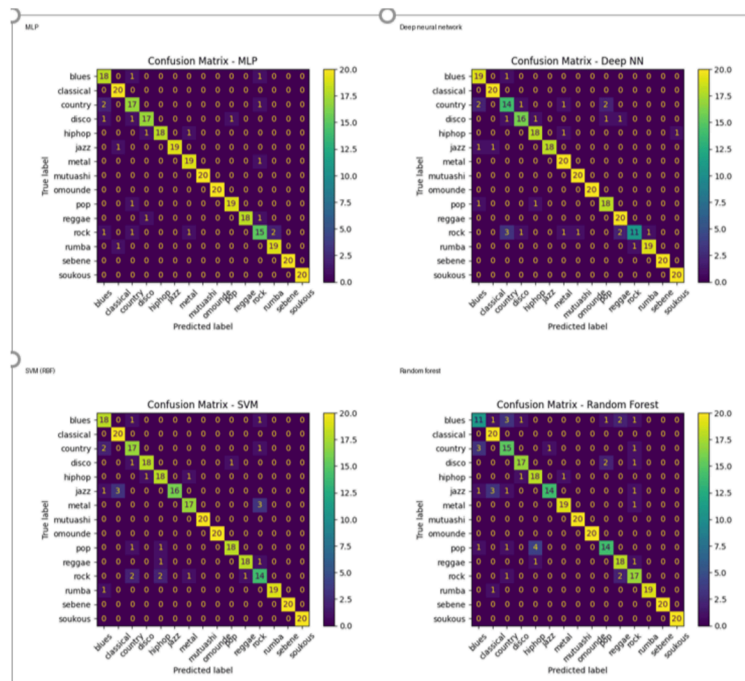


FIGURE 6: Confusion matrices for MLP, deep neural network, SVM, and random forest classifiers on the 15-class test fold

MLP, Multilayer Perceptron; SVM, Support Vector Machine

The main residual confusions occur among acoustically adjacent international genres, especially blues, country, jazz, and rock. Such errors are musically plausible because these genres share instrumentation, harmonic progressions, vocal textures, and historical stylistic overlap. The random forest matrix is more dispersed, indicating that its decision rules do not exploit the embedding geometry as effectively as the MLP or SVM.

Latent-space visualization

To further investigate the organization of the learned music representations, a two-dimensional visualization of the 1024-dimensional MERT embeddings was performed using t-SNE. The objective of this analysis was to provide an intuitive view of the relationships among music genres in the learned embedding space and to complement the quantitative evaluation and confusion matrix analysis.

Figure 7 presents the t-SNE projection of the MERT embeddings extracted from the 15-class dataset. The visualization reveals a partially structured representation space, where several genres form relatively compact clusters, while others show partial overlap. This behavior is consistent with the classification results, as genres with clearer separation in the embedding space tend to achieve higher recognition rates, whereas overlapping regions correspond to classes with greater acoustic similarity.

How to cite this article:

The Congolese genres included in the corpus exhibit distinguishable regions within the projected embedding space, suggesting that the pretrained MERT representations capture acoustic patterns useful for separating these musical categories from the other evaluated genres. Nevertheless, this observation should be interpreted cautiously, as t-SNE mainly preserves local neighbourhood relationships and does not provide a quantitative measurement of class separability.

Therefore, the t-SNE visualization should be considered as a complementary diagnostic analysis rather than direct evidence that the model has learned complete semantic representations of musical genres. Combined with classification metrics, confusion matrices, and ablation experiments, it provides additional insight into how self-supervised music representations organize culturally diverse audio content.

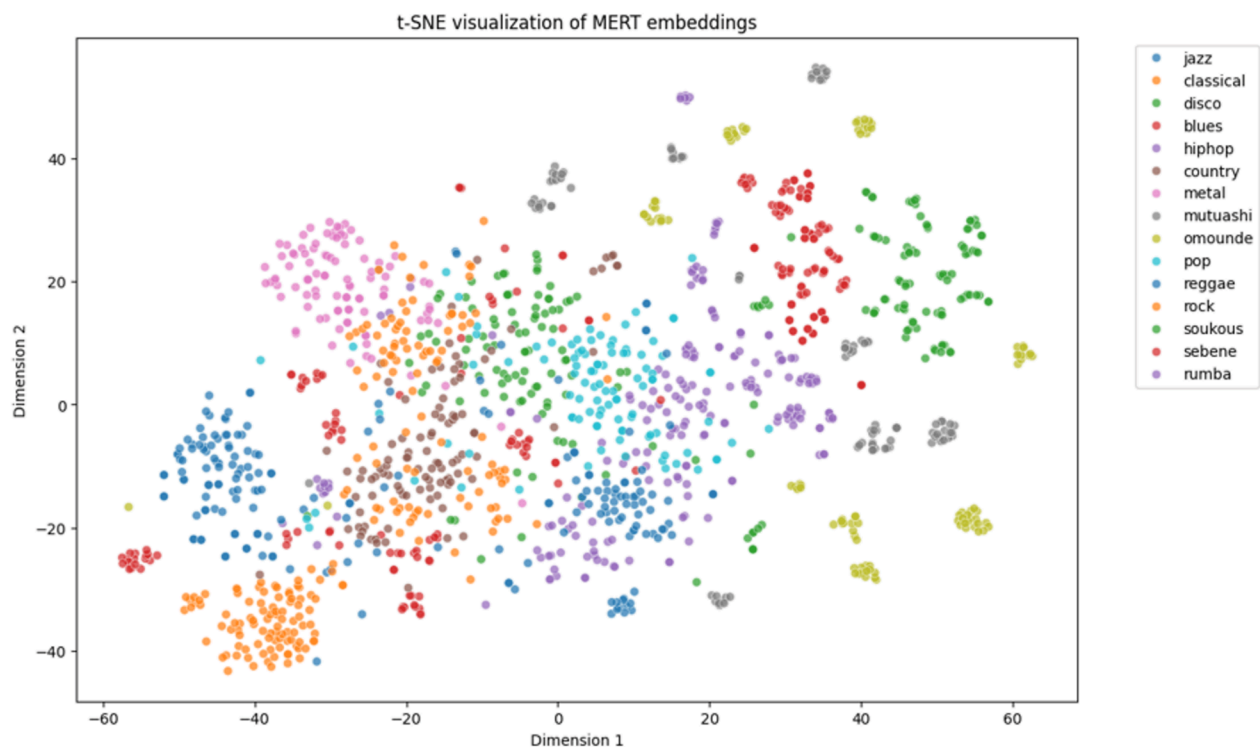


FIGURE 7: t-SNE visualization of 1,024-dimensional MERT embeddings
MERT, Music Embedding Representation Transformer; t-SNE, t-distributed Stochastic Neighbor Embedding

Discussion

Why the MLP Performs Best

The MLP achieved the best overall performance among the evaluated classifiers, reaching 93.00% accuracy and 93.02% macro-F1 score. This result suggests that the nonlinear transformation provided by the MLP was effective in exploiting the structure of the fixed MERT embeddings.

Unlike conventional handcrafted feature pipelines, where the classifier must compensate for limited representation capacity, self-supervised embeddings such as MERT already encode high-level acoustic and musical information. Therefore, the role of the downstream classifier is mainly to learn appropriate decision boundaries within an already informative feature space.

The comparison with the linear probe confirms this observation. Although the linear classifier achieved competitive performance, the MLP provided an additional improvement, indicating that some genre boundaries cannot be completely separated through linear transformations alone. The nonlinear layers allow the model to capture more complex relationships between latent audio representations.

However, increasing model depth did not improve performance. The compact deep neural network achieved lower accuracy than the MLP, suggesting that additional parameters are not necessarily beneficial when the representation extractor is already pretrained. For medium-sized datasets such as the present corpus, a lightweight classifier may provide a better balance between accuracy, computational cost, and risk of overfitting.

These findings are particularly relevant for low-resource MIR scenarios, where training large end-to-end neural networks is often impractical. A practical strategy is therefore to extract pretrained embeddings once and train lightweight classifiers adapted to the target musical domain.

Comparison With Recent MERT-Based and Transformer-Based Music Genre Classification Studies

Recent advances in self-supervised learning have significantly changed the landscape of MIR. Transformer-based models such as AST, HuBERT, BEATs, and MERT have demonstrated that large-scale pretrained representations can provide strong performance on downstream audio understanding tasks [18-21].

Most previous evaluations of these models have been conducted using established benchmark datasets such as GTZAN, FMA, MagnaTagATune, and MTG-Jamendo. While these datasets have played an essential role in advancing MIR research, they remain largely dominated by Western musical traditions. Consequently, the behavior of pretrained music representations when applied to culturally diverse repertoires remains insufficiently investigated.

The present study extends this research direction by evaluating MERT embeddings on a heterogeneous corpus combining international genres with five Congolese musical traditions: rumba, soukous, sebene, mutuashi, and omounde. Rather than introducing a new transformer architecture, the objective was to examine whether existing pretrained representations remain effective when transferred to underrepresented musical contexts.

The controlled comparison of four classifiers operating on identical MERT embeddings provides additional insight into the relationship between representation quality and classifier design. Since all models receive the same input features, differences in performance mainly reflect their ability to exploit the learned embedding space.

Furthermore, this study goes beyond accuracy reporting by integrating confidence intervals, McNemar's statistical testing, ablation experiments, confusion matrix analysis, t-SNE visualization, and explainability methods. This broader evaluation framework contributes to a more transparent assessment of self-supervised representations in cross-cultural MIR applications.

The results suggest that MERT embeddings can provide a strong foundation for culturally diverse music classification tasks. However, further validation on independent African and international music collections remains necessary to assess whether the observed performance extends beyond the constructed corpus.

Cultural Interpretation and Inclusiveness

The inclusion of Congolese musical genres in this study provides an opportunity to evaluate the transferability of pretrained music representations beyond the dominant Western-oriented benchmarks commonly used in MIR research. The results obtained for rumba, soukous, sebene, mutuashi, and omounde indicate that the MERT-based representation contains discriminative information that is useful for separating these categories within the constructed dataset.

However, these results should not be interpreted as evidence that the model achieves a complete understanding of Congolese musical culture or genre identity. Music genres are complex cultural constructs influenced by historical context, artistic practices, regional variations, and production environments. Therefore, the observed classification performance reflects the ability of the model to capture acoustic and structural patterns present in the available recordings rather than a comprehensive representation of the cultural meaning of these genres.

How to cite this article:

The inclusion of underrepresented musical traditions nevertheless demonstrates the importance of expanding MIR evaluation beyond conventional benchmark datasets. Building more culturally diverse corpora can contribute to the development of more inclusive audio technologies while reducing the risk that models are optimized only for dominant musical traditions.

Future research should extend this direction through collaboration with music experts, larger collections of African musical traditions, and independent validation datasets collected under different recording conditions and geographical contexts.

Threats to Validity

Several aspects should be considered when interpreting the results of this study. First, although GTZAN remains one of the most frequently used datasets in music genre classification research, previous studies have identified several limitations, including possible recording repetitions, annotation inconsistencies, and benchmark-specific biases. Therefore, results involving GTZAN should be interpreted as comparative evidence under controlled experimental conditions rather than as a complete measure of real-world genre recognition capability.

Second, dataset construction and source leakage represent important concerns in music classification. When multiple excerpts from the same track, artist, or recording source appear in both training and testing partitions, models may exploit source-related characteristics rather than purely musical properties. In this study, the main evaluation followed a stratified clip-level split to ensure a controlled comparison between classifiers. A complementary track-disjoint experiment was also conducted, showing a performance reduction from 93.00% to 84.67%, which highlights the importance of considering source independence in future evaluations.

Third, genre annotation remains inherently challenging because musical categories are not exclusively determined by measurable acoustic properties. Some genres share instrumental characteristics, rhythmic structures, and production techniques, which naturally introduces ambiguity in classification. The observed confusion patterns should therefore be considered partly related to the intrinsic complexity of genre boundaries.

Fourth, although the CongoMIR-GTZAN-15 corpus provides a balanced evaluation setting with 1,500 audio clips, its size remains limited compared with large-scale pretraining collections. The current results demonstrate the effectiveness of transfer learning under controlled conditions but do not establish universal generalization across all African or international musical traditions.

Fifth, the evaluation was based on a single stratified hold-out partition consisting of 1,200 training clips and 300 testing clips. Although this protocol provides a reproducible comparison because all classifiers are evaluated on exactly the same samples, repeated stratified cross-validation would provide a stronger estimation of performance variability. Future work should therefore investigate whether the observed differences remain stable across multiple independent partitions.

Finally, the use of online music sources introduces copyright and reproducibility considerations. Since many audio recordings are protected by intellectual property regulations, future resource sharing should prioritize the publication of metadata, extraction procedures, and learned representations rather than redistribution of original audio files.

Practical Implications

The results of this study highlight the potential of pretrained music representations as a practical solution for MIR tasks in low-resource environments. Instead of training large deep learning models from scratch, which requires substantial computational resources and large annotated datasets, organizations can adopt a transfer learning strategy based on pretrained encoders followed by lightweight downstream classifiers.

For digital music archives and cultural heritage institutions, the proposed approach can support automatic organization, metadata enrichment, and similarity-based exploration of large audio collections. In particular, the inclusion of Congolese musical genres demonstrates the possibility of developing MIR systems that better represent underrepresented musical traditions.

How to cite this article:

However, automated classification should be considered as an assistive tool rather than a replacement for human expertise. Because musical genres involve cultural, historical, and contextual dimensions that cannot be fully captured from audio signals alone, expert validation remains necessary, especially for archival, ethnomusicological, and preservation applications.

From a deployment perspective, the proposed pipeline offers practical advantages because MERT embeddings can be extracted offline and reused for different classification tasks. Lightweight classifiers such as MLP or SVM can then be trained with limited computational requirements, making the approach suitable for research groups and institutions with constrained resources.

Future developments could extend this framework toward larger African music collections, multilingual cultural metadata integration, artist-aware evaluation protocols, and human-in-the-loop annotation strategies. Such extensions would contribute to the development of more inclusive and reliable MIR systems.

Conclusions

This study presented a MERT-based transfer learning approach for cross-cultural music genre classification. By combining GTZAN international genres with five Congolese genres, the work addresses the underrepresentation of African music in common MIR evaluation settings. The MLP classifier achieved the best performance, with 93.00% accuracy and 93.02% F1-score, while SVM and the deep neural network remained competitive. Confusion matrices and t-SNE visualization showed that MERT embeddings provide meaningful separation for several Congolese genres and that most errors occur between musically related international genres.

The main scientific message is that pretrained self-supervised music representations can be effective in low-data multicultural settings, but high-impact claims require disciplined evaluation. The next step is not simply to increase accuracy, but to validate the system under track-disjoint and artist-disjoint splits, enlarge the African genre coverage, publish reproducible metadata and code, and test the model on independent corpora. With these safeguards, this line of work can contribute to more inclusive, transparent, and culturally aware MIR.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Muyisa Kavalami Mumbere, Olivier Muzembe Cswaka

Acquisition, analysis, or interpretation of data: Muyisa Kavalami Mumbere, Mpia Héritier Nsenge, Grace Lusenge Kasereka, Ushindi Kakule Pascal, Dieu-Donnee Kyalengekania Kambale

Drafting of the manuscript: Muyisa Kavalami Mumbere

Critical review of the manuscript for important intellectual content: Mpia Héritier Nsenge, Grace Lusenge Kasereka, Ushindi Kakule Pascal, Dieu-Donnee Kyalengekania Kambale, Olivier Muzembe Cswaka

Supervision: Mpia Héritier Nsenge, Grace Lusenge Kasereka, Ushindi Kakule Pascal, Dieu-Donnee Kyalengekania Kambale

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial**

How to cite this article:

Kavalami Mumbere M, Héritier Nsenge M, Lusenge Kasereka G, et al. (August 03, 2026) Cross-Cultural Music Genre Classification With Self-Supervised Music Transformers: A MERT-Based Transfer Learning Study on Congolese and International Genres. *Cureus J Comput Sci* 3 : es44389-026-00201-z. DOI <https://doi.org/10.7759/s44389-026-00201-z>

relationships: All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The datasets (and/or code) supporting this study are available from the corresponding author upon reasonable request.

References

1. 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. 2024, 13:3313. [10.3390/electronics13163313](#)
2. Akman A, Schuller BW: [Audio explainable artificial intelligence: A review](#). Intelligent Computing. 2024, 3:0074. [10.34133/icomputing.0074](#)
3. Prabhakar SK, Lee S-W: [Holistic approaches to music genre classification using efficient transfer and deep learning techniques](#). Expert Systems with Applications. 2023, 211:118636. [10.1016/j.eswa.2022.118636](#)
4. Wang J, Li S, Sung Y: [Deformer: Denoising Transformer for improved audio music genre classification](#). Applied Sciences. 2023, 13:12673. [10.3390/app132312673](#)
5. Sturm BL: [The state of the art ten years after a state of the art: Future research in music information retrieval](#). Journal of New Music Research. 2014, 43:147-172. [10.1080/09298215.2014.894533](#)
6. Lidy T, Silla CN Jr, Cornelis O, Gouyon F, Rauber A, Kaestner CAA, Koerich AL: [On the suitability of state-of-the-art music information retrieval methods for analyzing, categorizing and accessing non-Western and ethnic music collections](#). Signal Processing. 2010, 90:1032-1048. [10.1016/j.sigpro.2009.09.014](#)
7. Duan Z, van Kranenburg P, Nam J, Rao P: [Editorial for TISMIR: Special collection: Cultural diversity in MIR research](#). Transactions of the International Society for Music Information Retrieval. 2023, 6:203-205. [10.5334/tismir.179](#)
8. Li T: [Optimizing the configuration of deep learning models for music genre classification](#). Heliyon. 2024, 10:e24892. [10.1016/j.heliyon.2024.e24892](#)
9. Liu S, Mallol-Ragolta A, Parada-Cabaleiro E, et al.: [Audio self-supervised learning: A survey](#). Patterns. 2022, 3:100616. [10.1016/j.patter.2022.100616](#)
10. Kong Q, Cao Y, Iqbal T, Wang Y, Wang W, Plumbley MD: [PANNs: Large-scale pretrained audio neural networks for audio pattern recognition](#). IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2020, 28:2880-2894. [10.1109/taslp.2020.3030497](#)
11. Tzanetakis G, Cook P: [Musical genre classification of audio signals](#). IEEE Transactions on Speech and Audio Processing. 2002, 10:293-302. [10.1109/tsa.2002.800560](#)
12. Ahmed M, Rozario U, Kabir MM, Aung Z, Shin J, Mridha MF: [Musical genre classification using advanced audio analysis and deep learning techniques](#). IEEE Open Journal of the Computer Society. 2024, 5:457-467. [10.1109/ojcs.2024.3431229](#)
13. da Silva Muniz VH, de Oliveira e Souza Filho JB: [Robust handcrafted features for music genre classification](#). Neural Computing and Applications. 2023, 35:9335-9348. [10.1007/s00521-022-08069-5](#)
14. Costa YMG, Oliveira LS, Silla CN Jr: [An evaluation of convolutional neural networks for music classification using spectrograms](#). Applied Soft Computing. 2017, 52:28-38. [10.1016/j.asoc.2016.12.024](#)
15. 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. 2019, 13:206-219. [10.1109/jstsp.2019.2908700](#)
16. Seo W, Cho S-H, Teisseyre P, Lee J: [A short survey and comparison of CNN-based music genre classification using multiple spectral features](#). IEEE Access. 2024, 12:245-257. [10.1109/ACCESS.2023.3346883](#)
17. Sturm BL: [Classification accuracy is not enough: On the evaluation of music genre recognition systems](#). Journal of Intelligent Information Systems. 2013, 41:371-406. [10.1007/s10844-013-0250-y](#)
18. Li Y, Yuan R, Zhang G, et al.: [MERT: Acoustic music understanding model with large-scale self-supervised training](#). arXiv. 2024, [10.48550/arXiv.2306.00107](#)

How to cite this article:

Kavalami Mumbere M, Héritier Nsenge M, Lusenge Kasereka G, et al. (August 03, 2026) Cross-Cultural Music Genre Classification With Self-Supervised Music Transformers: A MERT-Based Transfer Learning Study on Congolese and International Genres. Cureus J Comput Sci 3 : es44389-026-00201-z. DOI <https://doi.org/10.7759/s44389-026-00201-z>

19. Gong Y, Chung YA, Glass J: [AST: Audio spectrogram transformer](#). Proceedings of Interspeech. 2021, 571-575. [10.21437/Interspeech.2021-698](#)
20. Hsu W-N, Bolte B, Tsai Y-HH, Lakhota K, Salakhutdinov R, Mohamed A: [HuBERT: Self-supervised speech representation learning by masked prediction of hidden units](#). IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021, 29:3451-3460. [10.1109/taslp.2021.3122291](#)
21. Jiménez-Bravo DM, Lozano Murciego Á, Navarro-Cáceres JJ, Navarro-Cáceres M, Harkin T: [Identifying Irish traditional music genres using latent audio representations](#). IEEE Access. 2024, 12:92536-92548. [10.1109/access.2024.3421639](#)

How to cite this article:

Kavalami Mumbere M, Héritier Nsenge M, Lusenge Kasereka G, et al. (August 03, 2026) Cross-Cultural Music Genre Classification With Self-Supervised Music Transformers: A MERT-Based Transfer Learning Study on Congolese and International Genres. Cureus J Comput Sci 3 : es44389-026-00201-z. DOI <https://doi.org/10.7759/s44389-026-00201-z>