Música Hodie | ISSN: 2317-6776 DOI: 10.5216/mh.v26.85233

Investigating genre similarity boundaries in Brazilian music using deep learning models

Investigando os Limites da Similaridade de Gêneros na Música Brasileira Utilizando Modelos de Aprendizado Profundo

Victória de Souza Leon Guimarães
Institute of Computing, Federal University of Amazonas (UFAM), Manaus, Amazonas, Brazil
vsg@icomp.ufam.edu.br

João Gustavo Kienen
Faculty of Arts, Federal University of Amazonas (UFAM), Manaus, Amazonas, Brazil
gustavokienen@ufam.edu.br

Rosiane de Freitas Rodrigues
Institute of Computing, Federal University of Amazonas (UFAM), Manaus, Amazonas, Brazil
rosiane@icomp.ufam.edu.br

Abstract: Understanding how musical genres are organized in learned representation spaces remains a central challenge in music information retrieval. This study investigates the latent structure induced by a Vision Transformer (ViT) model fine-tuned on Brazilian regional music. We introduce the Brazilian Regional Music Dataset (BYRM), a curated collection of 1,082 tracks distributed across ten culturally diverse genres. Our analysis focuses on the best-performing experimental configuration identified in prior experiments, in which 10-second audio segments are extracted from the 90–120 second portion of each track. Mel-spectrogram representations are used as model input, and time-local embeddings are derived from the trained ViT. To examine the organization of the learned feature space, we apply dimensionality reduction techniques, including Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and

Projection (UMAP). Cosine similarity is computed to quantify inter-genre proximity. The model achieves 81.94% classification accuracy and an F1-score of 81.84%, demonstrating strong discriminative capability. Beyond classification performance, the representation analysis reveals coherent clustering patterns, with stylistically related genres exhibiting higher proximity while structurally distinct genres remain well separated. These findings suggest that transformer-based audio embeddings effectively encode culturally grounded genre relationships within a musically diverse context.

Keywords: music information retrieval. music genre recognition. Brazilian music. vision transformers. embeddings.

Resumo: Compreender como os gêneros musicais se organizam em espaços de representação aprendidos constitui um desafio relevante na área de Recuperação de Informação Musical. Neste trabalho, investigamos a estrutura latente aprendida por um modelo Vision Transformer (ViT) ajustado para a classificação de música regional brasileira. Para isso, apresentamos o Brazilian Regional Music Dataset (BYRM), um conjunto de dados curado composto por 1.082 faixas distribuídas em dez gêneros culturalmente diversos. A análise considera a configuração experimental de melhor desempenho identificada em estudos anteriores, na qual são utilizados segmentos de 10 segundos extraídos do intervalo entre 90 e 120 segundos de cada faixa. Os mel-espectrogramas são empregados como representação de entrada do modelo, a partir dos quais são extraídos embeddings temporais do ViT treinado. Para investigar a organização do espaço latente, aplicamos técnicas de redução de dimensionalidade, incluindo Análise de Componentes Principais (PCA), Incorporação Estocástica de Vizinhos Distribuída em t (t-SNE) e Aproximação e Projeção Uniforme de Variedades (UMAP), além do cálculo da similaridade do cosseno para mensurar a proximidade entre gêneros. O modelo alcançou 81,94% de acurácia e 81,84% de F1-score, evidenciando elevada capacidade de discriminação. Para além dos resultados quantitativos, a análise

das representações indica padrões consistentes de agrupamento, nos quais gêneros estilisticamente próximos apresentam maior similaridade no espaço de embeddings, enquanto estilos mais distintos permanecem claramente separados. Esses resultados demonstram que representações baseadas em transformadores são capazes de capturar relações estilísticas culturalmente fundamentadas em contextos musicais diversos.

Palavras-chave: recuperação de informação musical. reconhecimento de gênero musical. música brasileira. vision transformers. embeddings.

Submitted on: February 18, 2026
Accepted on: March 5, 2026
Published on: April 2026

1. Introduction

In the context of Brazilian regional music, genre boundaries are often fluid and shaped by shared cultural, rhythmic, and instrumental characteristics. Genres such as samba and pagode or forró and vaneira frequently exhibit strong musical affinities, making their separation particularly challenging for automatic classification systems (Cassano; Cassano; Pechman, 2024; Fernandes, 2005). The presence of hybrid tracks, stylistic blending, and overlapping instrumentation further intensifies inter-genre similarity and complicates computational modeling (Correa, 2018; Castro, 2010). These characteristics highlight the need for representation learning approaches capable of capturing subtle stylistic nuances.

Within the broader field of Music Information Retrieval (MIR), Music Genre Recognition (MGR) has long been a central task, supporting large-scale organization and retrieval of musical content (Aucouturier; Pampalk, 2008). Although supervised machine learning methods have significantly advanced genre classification (Silla Jr.; Kaestner; Koerich, 2007), distinguishing closely related genres remains difficult due to fine-grained stylistic differences (Sturm, 2014). This difficulty becomes even more pronounced in culturally rich and regionally grounded repertoires.

Recent developments in deep learning have enabled models to learn high-level representations directly from audio signals. Transformer-based architectures (Vaswani et al., 2017), particularly the Vision Transformer (ViT) (Dosovitskiy et al., 2020), have demonstrated strong capacity to model global dependencies in spectrogram representations. By treating spectrogram patches as visual tokens, ViT encodes hierarchical and contextual information that may reveal latent genre structures within embedding spaces.

Rather than focusing exclusively on classification accuracy, this study investigates how Brazilian regional genres are organized in the latent space learned by a fine-tuned ViT model. Guided by insights from music cognition, which emphasize the structural relevance of choruses and melodic hooks as salient genre-defining segments (Byron; Rushworth; Stewart, 2025), we extract embeddings from strategically selected excerpts and analyze them using cosine similarity and dimensionality reduction techniques. To support this analysis, we introduce BYRM (Brazilian YouTube Regional Music), a publicly available dataset comprising representative regional genres. The dataset represents a central contribution of this work and aims to foster reproducibility and future research in music genre classification (GUIMARÃES; FREITAS, 2025).

2. Theoretical Background

Musical genre is not a fixed or rigidly delimited category, but rather a social construct shaped by cultural practices, aesthetic conventions, and institutional dynamics (Krogh, 2019). Genres emerge through processes of abstraction and contextual interpretation, involving symbolic, stylistic, and socio-cultural elements. From this perspective, computational models do not learn explicit genre rules but instead capture statistical regularities that reflect musical and cultural patterns present in audio data. Consequently, the embeddings produced by deep learning systems can be interpreted as computational abstractions of genre-related characteristics, encoding contextual similarities derived from diverse musical examples.

In recent years, the analysis of internal model representations has gained attention to investigate how such abstractions are structured. Embeddings, represented as fixed-length vectors in high-dimensional spaces, allow for the examination of similarity relationships between musical excerpts. Dimensionality reduction techniques such as Principal Component Analysis

(PCA) (Wold; Esbensen; Geladi, 1987), t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten; Hinton, 2008), and Uniform Manifold Approximation and Projection (UMAP) (McInnes; Healy; Melville, 2018) are commonly used to visualize these latent spaces. While PCA emphasizes global variance structure, t-SNE prioritizes local neighborhood preservation, and UMAP seeks a balance between local and global relationships with improved scalability. Quantitative similarity measures, such as cosine distance, can also be applied directly in the original embedding space to assess inter-genre proximity (Spijkervet; Burgoyne, 2021; Barnett; Garcia; Pardo, 2024).

Within the broader field of Music Information Retrieval (MIR), Music Genre Recognition (MGR) remains one of the most studied tasks (Aucouturier; Pampalk, 2008). Early computational approaches relied on handcrafted acoustic descriptors, including spectral centroid, bandwidth, zero-crossing rate, and Mel Frequency Cepstral Coefficients (MFCCs), which approximate perceptual aspects of human auditory processing (Zhang; Ras, 2007; Turab et al., 2022). With the consolidation of deep learning methods, Mel-spectrograms became a standard input representation, converting audio signals into time-frequency images aligned with the Mel scale (Schluter; Grill, 2015).

Deep neural architectures have progressively expanded the modeling capacity of MGR systems. Convolutional Neural Networks capture local time-frequency structures, Recurrent Neural Networks model sequential dependencies, and Transformer-based models leverage self-attention mechanisms to learn global contextual relationships (LeCun et al., 1998; Hopfield, 1982; Vaswani et al., 2017). The Vision Transformer (ViT) (Dosovitskiy et al., 2020), originally introduced for image classification, has been adapted for music analysis by treating spectrogram patches as input tokens. This formulation enables the modeling of long-range dependencies and

hierarchical feature representations, aligning with insights from music cognition that emphasize the structural salience of segments such as choruses, which often carry genre-defining characteristics (Byron; Rushworth; Stewart, 2025).

3. Related Work

Table 1 presents a structured overview of representative studies that investigate musical embeddings, genre classification, and similarity modeling. Rather than focusing exclusively on classification accuracy, the table highlights how different works use similarity, whether as a retrieval mechanism, a visualization strategy, or an implicit learning objective. The columns summarize the core research objective of each study, the neural architecture adopted, the use of dimensionality reduction techniques for latent space inspection, the similarity metric employed when applicable, and the datasets used for evaluation.

Table 1 – Comparison of Related Works on Similarity Use and Technique

Work	MGR	Similarity?	Model	Dim. Reduction	Similarity Technique	Dataset
CLMR (2021) (Spijkervet; Burgoyne, 2021)	✓	✓	CNN	✗	Cosine	MagnaTagATune
Music Recommender (2021) (Sheikh Fathollahi; Razzazi, 2021)	✓	✓	CNN	✗	Euclidean and Cosine	GTZAN, Emotify
Visual Embeddings (2022) (Armijo; Rivas, 2025)	✗	✓	Vision Transformer	UMAP	Visual Cluster	Craigslist and OfferUp
Audio Influence (2024) (Barnett; Garcia; Pardo, 2024)	✗	✓	Jukebox (OpenAI)	PCA, t-SNE	Euclidean and Cosine	Million Song Dataset, Spotify

EDM Benchmark (2024) (Shu et al., 2024)	✓	✓	CNN, Vision Transformer	PCA, t-SNE, UMAP	Visual Cluster	EDM Benchmark
Musicon (2024) (Luo; Hoveling; Eisemann, 2024)	✓	✓	CLMR	UMAP, PCA, t-SNE, others	Cosine	GTZAN, MagnaTagATune, others
Myna (2025) (Yonay; Hammond; Yang, 2025)	✗	✗	Vision Transformer	✗	✗	AudioSet, GTZAN, MTT, EmoMusic
This work	✓	✓	**Vision Transformer**	**PCA, t-SNE, UMAP**	**Visual Cluster and Cosine**	**BYRM (Brazilian)**

Source: prepared by the authors.
Description of the table: Comparison of related works on similarity and genre recognition.

The literature reveals heterogeneous strategies for integrating similarity into music representation learning. Some works prioritize exploratory analysis of embedding spaces without directly targeting genre recognition. For instance, Armijo and Rivas (2025) apply Vision Transformers combined with UMAP to visually inspect cluster formation, emphasizing qualitative interpretation over explicit distance-based similarity metrics. In a related direction, Barnett, Garcia, and Pardo (2024) extract embeddings from Jukebox and compute cosine and Euclidean distances to analyze stylistic proximity between artists, employing PCA and t-SNE projections for interpretability. Yonay, Hammond, and Yang (2025) propose a contrastive framework using Vision Transformers to learn general-purpose audio representations, later evaluated on benchmark datasets, although similarity itself is not the primary analytical focus.

Other studies incorporate similarity more directly within genre recognition pipelines. Sheikh Fathollahi and Razzazi (2021) employ CNN-based embeddings both for classification and recommendation, using cosine and Euclidean distance to compare musical tracks. The CLMR framework (Spijkervet; Burgoyne, 2021) models similarity implicitly through a contrastive objective optimized with cosine-based loss, demonstrating its effectiveness in downstream genre tasks. Shu

et al. (2024) benchmark CNNs and Vision Transformers for EDM sub-genre classification and complement quantitative results with dimensionality reduction visualizations to inspect class separability. The Musicon system (Luo; Hoveling; Eisemann, 2024) extends CLMR embeddings toward interactive retrieval, combining cosine similarity with UMAP-based visualization interfaces.

Together, these contributions illustrate that similarity can assume multiple roles in MIR research, functioning as a training signal, an exploratory analytical tool, or a retrieval mechanism. However, few studies explicitly center their analysis on how genre relationships are structured within the latent space itself, particularly in culturally specific repertoires.

In the Brazilian context, prior work has primarily emphasized classification performance rather than latent similarity interpretation. Traditional machine learning approaches achieved 79.7% accuracy using SVM and 86.1% on a regional dataset (Sousa; Pereira; Veloso, 2016), while Random Forest models obtained 65.3% on Brazilian genres and 64.2% when combined with GTZAN (Conceição et al., 2020). CNN-based strategies later introduced feature fusion techniques, reaching 88.5% accuracy on GTZAN (Silva et al., 2021a), and domain-adapted architectures that outperformed generalist models on the 4MuLA dataset (Silva et al., 2021b). The 4MuLA collection subsequently evolved into a large-scale multimodal benchmark comprising over 96,000 tracks across 76 genres (Silva; Silva; Marcacini, 2020).

In contrast to these approaches, the present study shifts the emphasis from pure classification performance toward the structural organization of embeddings. By leveraging the BYRM dataset, specifically curated for Brazilian regional music, we analyze inter-genre proximity using cosine similarity and multiple dimensionality reduction techniques. This

perspective enables a deeper examination of how transformer-based representations encode culturally grounded stylistic relationships.

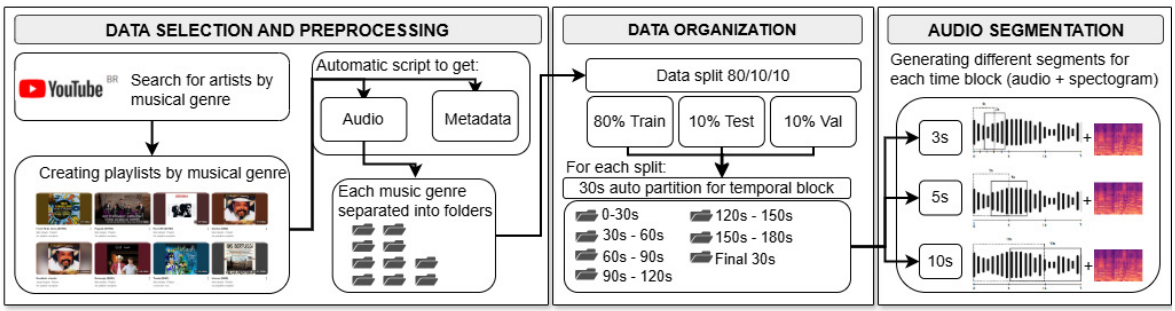
4. Methodology

This section outlines the experimental design adopted to investigate genre similarity in the latent space of a transformer-based model. The methodology encompasses dataset construction, feature representation, model training, and embedding analysis. Emphasis is placed on the extraction and examination of time-local embeddings derived from a fine-tuned Vision Transformer.

4.1. Dataset Construction and Segmentation Strategy

The Brazilian YouTube Regional Music (BYRM) dataset is a curated collection of 1,082 tracks across 10 Brazilian regional genres: Axé (101), Carimbó (103), Forró (104), Pagode (102), Brazilian Rock (107), Samba (103), Sertanejo (104), Toada (147), Vaneira (107), and Xote Gaúcho (104). Figure 1 illustrates the data preparation pipeline.

Figure 1: Overview of the BYRM dataset preparation process



Source: prepared by the authors.
Description of the image: Dataset preparation pipeline.

The tracks were collected from publicly available YouTube albums featuring representative artists of each genre. Audio files were downloaded using the Pytube library, converted to

WAV format, and associated with metadata including video title, channel ID, and file path. To avoid data leakage, the dataset was partitioned at the track level into training, validation, and test sets following an 80/10/10 split. Each track was segmented into six non-overlapping 30-second excerpts covering the intervals from 0–30 seconds to 150–180 seconds. As shown in Figure 1, each excerpt was subsequently subdivided into overlapping windows of 3, 5, and 10 seconds with 50% overlap. This segmentation strategy enables the capture of rhythmic, harmonic, and timbral patterns across multiple temporal resolutions.

4.2 Audio Representation and Feature Extraction

For each short segment, Mel-spectrograms were computed and used as input to the neural models. Spectrogram generation employed a Fast Fourier Transform window of 2048 samples, hop length of 512, 128 Mel frequency bands, and a sampling rate of 22,050 Hz. A power parameter of 2.0 was applied to enhance spectral contrast, following the default configuration of the Librosa library. These time-frequency representations preserve localized musical structures while remaining compatible with image-based neural architectures. The resulting spectrograms serve as the standardized input representation for all models evaluated in this study.

4.3 Model Architectures and Training Procedure

We adopted three classification models for the experiments: Vision Transformer (ViT-B/16), ResNet50, and a Support Vector Machine (SVM). Table 2 summarizes the configurations used for each model. All models were trained to recognize Brazilian regional music genres using the same data split. While the ViT was the primary model used to extract embeddings for the similarity analysis, ResNet50 and SVM were included for baseline comparison.

Table 2 – Summary of Model Configurations Used in the Experiments

Configuration Aspect	SVM	ResNet50	ViT
Pretrained weights	No	Yes (ImageNet)	Yes (ImageNet)
Input representation	Handcrafted features	Mel-spectrogram	Mel-spectrogram
Feature extraction	MFCC, chroma, spectral centroid, bandwidth, rolloff, ZCR	Spectrogram image	Spectrogram image
Segment durations	3s	3s	3s, 5s, 10s
Temporal excerpts	Best and worst	Best and worst	All six excerpts
Training epochs	—	200	100
Early stopping patience	—	15	10
Optimizer	—	Adam (lr=1e-4)	Adam (lr=1e-4)
Learning rate schedule	—	0.8 → 0.6 → 0.3	0.8 → 0.6 → 0.3
LR milestones (epochs)	—	15, 100, 150	10, 30, 40
Dropout rate	—	Default	30% (head), 7% (internal)

Source: prepared by the authors.
Description of the table: Model configurations used in the experiments.

The Vision Transformer was fine-tuned using Mel-spectrogram inputs derived from segments within the 90–120 second excerpt of each track. Window sizes of 3, 5, and 10 seconds were evaluated. Embeddings were extracted from the [CLS] token of the final transformer layer, which aggregates global contextual information across input patches. ResNet50 was trained on 3-second spectrogram images extracted from both the most and least informative temporal excerpts. The SVM baseline employed handcrafted acoustic descriptors, including MFCCs, chroma features, spectral centroid, bandwidth, rolloff, and zero-crossing rate. Although SVM does not provide internal embeddings comparable to transformer-based representations, its classification performance serves as a baseline reference, as shown in Table 2.

4.4 Experimental Configuration and Embedding Analysis

Previous experiments identified the 90–120 second excerpt as the most informative segment for genre classification, whereas the 0–30 second excerpt yielded comparatively lower performance (GUIMARÃES; KIENEN; FREITAS, 2025). Among the evaluated window sizes, 10-second segments consistently achieved the highest classification accuracy.

Based on these findings, the present study focuses on embeddings extracted from 10-second windows within the 90–120 second interval. To analyze how genre information is structured in the learned representation space, embeddings were extracted from the [CLS] token both before and after fine-tuning. This comparison allows qualitative examination of how supervised training reshapes the organization of genres in the latent space.

Dimensionality reduction techniques, namely Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP), were independently applied to the extracted embeddings. These projections facilitate the visualization of clustering behavior and class separability. Complementing the visual analysis, cosine similarity was computed directly in the original embedding space. For each pair of genres, the mean pairwise cosine similarity between their embeddings was calculated, resulting in a similarity matrix visualized as a heatmap in the Results section.

5. Results

This section presents the empirical findings that support the analysis of genre similarity in the learned embedding space. We first report the classification performance that motivated the selection of the experimental configuration. Subsequently, we examine how genres are organized in the latent space through visual projections and quantitative similarity measures.

5.1. Selection of the Experimental Configuration

To investigate how temporal segmentation influences genre modeling, we evaluated different segment durations across two distinct excerpts of each track: the initial 0–30 seconds and the central 90–120 seconds. Table 3 reports the performance of the Vision Transformer model for segment durations of 3, 5, and 10 seconds applied to both excerpts.

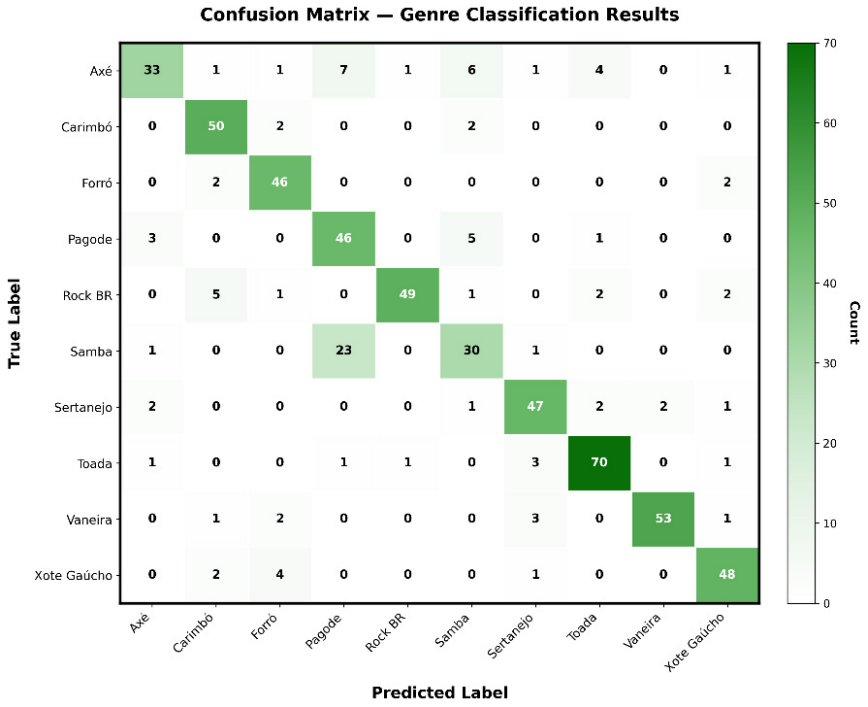
Table 3 – ViT Performance Across Segment Durations for Two Temporal Excerpts

Metric	Excerpt: 0–30s			Excerpt: 90–120s		
	3s	5s	10s	3s	5s	10s
Accuracy (%)	**74.76**	73.50	73.16	79.60	78.94	**81.94**
Precision (%)	**75.29**	76.00	75.35	80.24	79.48	**82.85**
Recall (%)	**74.76**	73.50	73.16	79.60	78.94	**81.94**
F1-score (%)	**74.91**	74.12	72.57	79.51	78.98	**81.84**

Source: prepared by the authors.
Description of the table: ViT performance across segment durations and temporal excerpts.

The highest performance was obtained using 10-second segments extracted from the 90–120 second excerpt, reaching 81.94% accuracy and 81.84% F1-score. This configuration was therefore adopted for subsequent embedding analysis. The results suggest that the central excerpt of the track provides richer genre-defining information. These sections frequently contain recurring motifs, refrains, or stable instrumental arrangements, which may enhance representational consistency. The 10-second window appears to balance contextual completeness with temporal resolution, enabling the model to capture discriminative patterns without excessive fragmentation. To further inspect classification behavior, Figure 2 presents the confusion matrix for the best-performing configuration.

Figure 2: Classification confusion matrix for 10-second segments from the 90–120s excerpt using the ViT model. Rows: actual labels; columns: predicted labels.



Source: prepared by the authors.
Description of the image: Confusion matrix of the best-performing ViT configuration.

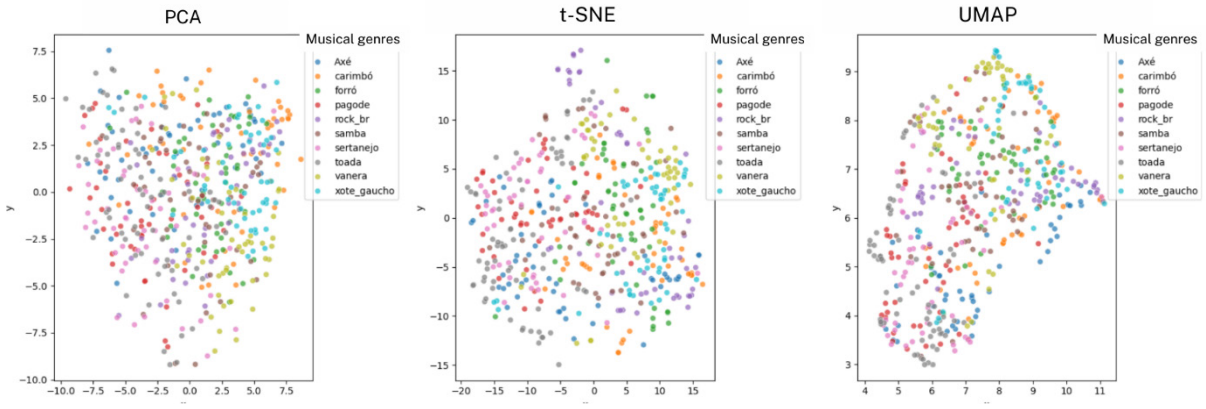
Misclassifications predominantly occur between stylistically related genres, such as samba and pagode or axé and samba. Conversely, some culturally related genres, such as vaneira and xote gaúcho, exhibit relatively low confusion rates. Notably, the matrix is asymmetric, reflecting directional prediction tendencies. These observations indicate that classification accuracy alone does not fully describe inter-genre relationships, motivating a deeper examination of the embedding space.

5.2. Latent Space Organization

To better understand how the model organizes musical information in the latent space, dimensionality reduction techniques were applied to visualize the embeddings generated by the Vision Transformer (ViT). Three methods were used for this purpose: PCA, t-SNE, and UMAP. Figure 3 illustrates the embedding distribution prior to fine-tuning, using ImageNet

pretrained weights. At this stage, genres appear largely dispersed, reflecting the absence of task specific adaptation. Nevertheless, partial proximity between certain genres can already be observed, likely driven by shared low-level acoustic features.

Figure 3: Visualization of the validation embedding space before training using PCA, t-SNE, and UMAP.



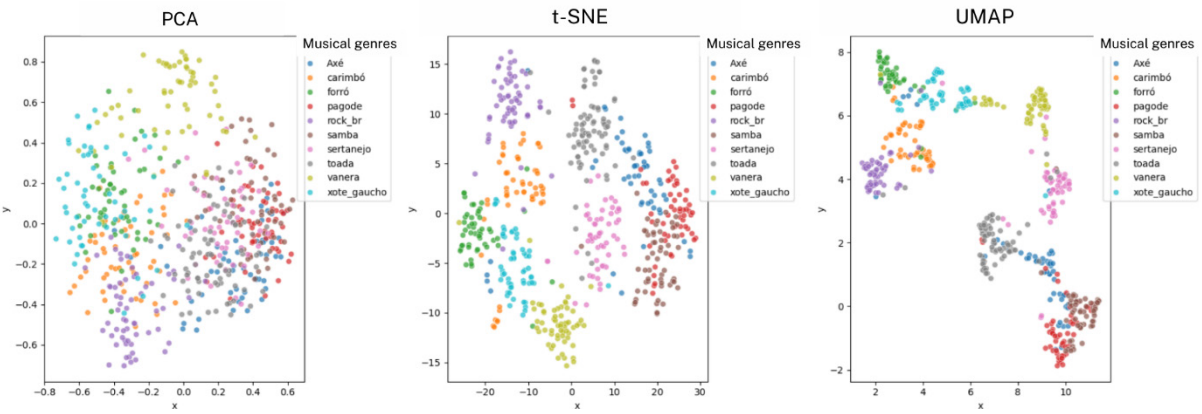
Source: prepared by the authors.
Description of the image: Embedding space visualization before fine-tuning (PCA, t-SNE, UMAP).

On the other hand, Figure 4 presents the embeddings generated from the BYRM test set after training the model. In this case, the embeddings form a much more structured space. Clear genre-specific clusters emerge, indicating that the model has learned meaningful representations that reflect stylistic and structural differences between genres. Among the three techniques, t-SNE and UMAP provided more visually informative results than PCA. These non-linear methods were more effective in revealing cluster boundaries and genre groupings.

UMAP, in particular, produced the most coherent separation among genres, with well-defined and compact clusters across the projection. This behavior can be attributed to its capacity to preserve both local neighborhood relationships and aspects of the global manifold structure, resulting in projections that maintain meaningful distances between similar and dissimilar samples. Unlike PCA, which relies on linear variance maximization

and may compress non-linear relationships, and t-SNE, which strongly emphasizes local structure at the expense of global organization, UMAP achieves a more balanced representation of the embedding geometry.

Figure 4: Visualization of the embedding space using PCA, t-SNE, and UMAP. The embeddings were extracted from the test set after training the ViT model.



Source: prepared by the authors.
Description of the image: Embedding space visualization after fine-tuning (PCA, t-SNE, UMAP).

Genres with high visual proximity: Samba and pagode form tight clusters in the lower right region, which is musically consistent since pagode is derived from samba and shares its rhythmic and harmonic foundations. Sertanejo and toada also appear grouped on the right side, indicating similar acoustic patterns, often characterized by melodic storytelling and voice-centered instrumentation. Carimbó and forró, located on the left side, exhibit rhythmic intensity and strong percussive elements, which may explain their proximity in the latent space.

Genres with partial separation and overlap: Axé appears between toada, sertanejo, and carimbó, suggesting that the model identified shared vocal and harmonic traits, even though axé tends to be more energetic. Additionally, xote gaúcho and vaneira, although positioned in different regions, present areas of intersection. This is coherent with their common regional origin and instrumentation, particularly the use of accordion-based rhythms.

Genres with clear separation: Brazilian rock is the most visually isolated genre, located distinctly in the lower area of the projection. This indicates that the model captured its contrasting characteristics, including electric instrumentation, distorted timbres, and a rhythmic structure that differs significantly from the regional genres in the dataset. This visual analysis supports the conclusion that the Vision Transformer model is capable of learning semantically meaningful representations, placing similar genres closer in the embedding space and separating those with more distinct musical traits.

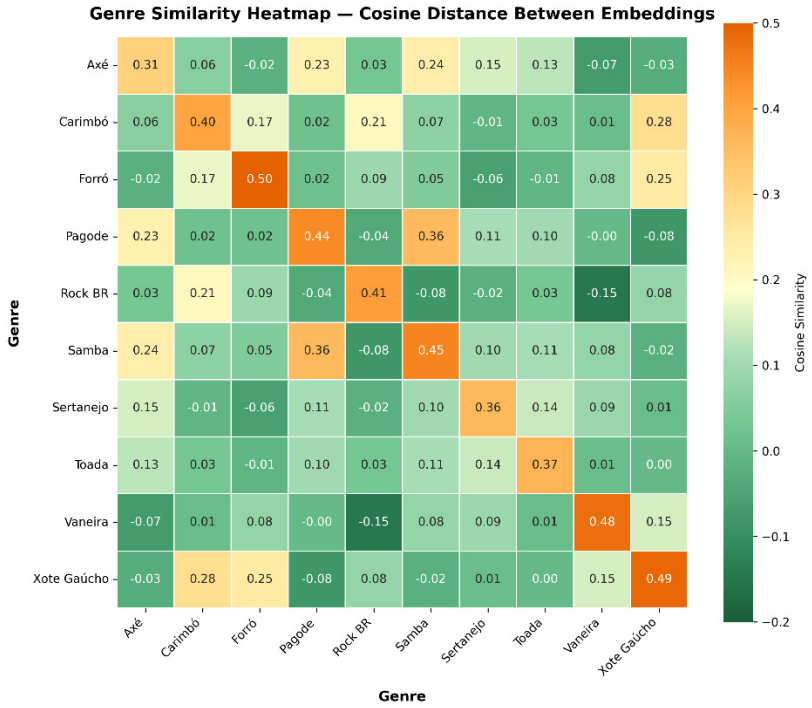
Beyond the qualitative interpretation of cluster positions, these spatial patterns suggest that the learned embedding space encodes genre relationships along multiple acoustic and stylistic dimensions simultaneously. Rather than forming rigid boundaries, the projections indicate a continuum of musical characteristics, where certain genres share transitional traits while others occupy more distinctive regions. This behavior reinforces the idea that genre, particularly in culturally rich contexts such as Brazilian regional music, is not a strictly discrete category but a structured yet fluid construct. The resulting embedding space therefore reflects both stylistic proximity and internal diversity, offering a representational perspective that complements traditional classification metrics.

5.3. Quantitative Genre Similarity

Although dimensionality reduction techniques such as PCA, t-SNE, and UMAP are useful for visualization purposes, they distort the original distances between embeddings and should not be used for quantitative similarity calculations. Therefore, to accurately measure the proximity between musical genres, the analysis was performed using the embeddings in their original vector space, with 768 dimensions, as generated by the final trained Vision Transformer model.

For each genre, the mean embedding vector was computed by averaging the embeddings of all ten-second segments from the test set that belong to that genre. The pairwise cosine similarity was then calculated between these mean vectors, resulting in a similarity matrix that reflects how the model internally relates different genres. Figure 5 presents this similarity matrix as a heatmap, revealing which genres share closer representations in the learned embedding space.

Figure 5: Cosine similarity heatmap between musical genres based on Vision Transformer embeddings.



Source: prepared by the authors.
Description of the image: Cosine similarity heatmap between musical genres.

The similarity heatmap in Figure 5 reveals how the model internally relates the musical genres based on their average embeddings. The highest cosine similarity score, 0.36, is observed between the samba and pagode genres, reflecting their similar rhythmic structures. Other strongly related pairs include carimbó and xote gaúcho with a similarity of 0.28, and carimbó and Brazilian rock with 0.21. In contrast, the lowest

similarity values occur between Brazilian rock and vaneira. These distances reflect the substantial differences in instrumentation and musical structure, particularly the presence of electric timbres and non-traditional rhythms in Brazilian rock. Unlike the confusion matrix, which reflects directional prediction tendencies, the cosine similarity heatmap is symmetric by construction, as cosine similarity between mean embeddings does not depend on prediction direction.

Finally, an intra-genre analysis of the main diagonal of the similarity matrices shows that none of the genres reach internal similarity values close to 1. The maximum observed value was approximately 0.5, indicating that even among segments of the same genre, there are considerable variations in the vector space. This behavior is expected, given the diverse nature of Brazilian musical genres, which often blend stylistic influences and exhibit substantial internal diversity. Overall, the cosine similarity analysis confirms that the Vision Transformer model learns semantically meaningful genre relationships, grouping similar genres more closely in the latent space and distancing those with divergent musical characteristics. These findings reinforce the clustering patterns observed in the UMAP projection.

6. Conclusion and Future Work

This work explored the use of a Vision Transformer (ViT) model for the classification and representation of Brazilian regional music genres, leveraging spectrogram-based representations and segment-based training. The BYRM dataset, developed specifically for this study, provided a diverse and challenging collection of audio samples, capturing the richness and complexity of Brazilian musical traditions. The results demonstrated that the ViT model is capable of learning meaningful musical representations. Visualization of the embedding space revealed genre-specific clustering, and

the cosine similarity analysis confirmed that stylistically related genres are mapped closer together in the latent space. Pairs such as samba and pagode, or vaneira and forró and carimbó, showed high proximity, while more distinct genres such as Brazilian rock were clearly separated.

These findings indicate that transformer-based audio models can structure genre relationships in a coherent way. Rather than creating strictly separated categories, the embedding space reflects gradual variations in acoustic and stylistic characteristics, which is consistent with the fluid nature of musical genres. This structured organization may be useful for applications beyond classification, including music retrieval, recommendation systems, and the analysis of stylistic relationships.

For future work, we plan to explore intra-genre variation through qualitative analysis, aiming to understand how stylistic diversity is expressed within each genre. Although previous work identified the most representative segment of the track, future studies could further examine how genre characteristics change across different musical sections. Furthermore, this approach opens opportunities for interdisciplinary collaboration with areas such as computational musicology, where genre proximity visualization may inspire new research questions about cultural, stylistic, and historical relationships between Brazilian music genres.

References

ARMIJO, Cameron; RIVAS, Pablo. Exploring visual embedding spaces induced by vision transformers for online auto parts marketplaces. *arXiv preprint arXiv:2502.05756*, 2025. Disponível em: https://arxiv.org/abs/2502.05756. Acesso em: fev. 2026.

AUCOUTURIER, Jean-Julien; PAMPALK, Elias. Introduction--from genres to tags: A little epistemology of music information retrieval

research. *Journal of New Music Research*, v. 37, n. 2, p. 87-92, 2008. DOI: 10.1080/09298210802479318. Disponível em: https://doi.org/10.1080/09298210802479318. Acesso em: fev. 2026.

BARNETT, Julia; GARCIA, Hugo Flores; PARDO, Bryan. Exploring musical roots: Applying audio embeddings to empower influence attribution for a generative music model. *In: INTERNATIONAL SOCIETY FOR MUSIC INFORMATION RETRIEVAL (ISMIR), 25., 2024, San Francisco. Anais [...]*. 2024. p. 360-368. Disponível em: https://arxiv.org/abs/2401.14542. Acesso em: fev. 2026.

BYRON, Timothy P.; RUSHWORTH, Carl T.; STEWART, Maxwell J. Popular music excerpts are rated as more memorable and salient if they involve vocals, compound hooks, and choruses. *Music Perception: An Interdisciplinary Journal*, v. 42, n. 3, p. 197-206, 2025. DOI: 10.1525/mp.2024.2322897. Disponível em: https://doi.org/10.1525/mp.2024.2322897. Acesso em: fev. 2026.

CASSANO, Lucas; CASSANO, Elisa de Oliveira Morais Nacur; PECHMAN, Robert Moses. Passado e presente da música carioca: Narrativas urbanas e evolução tecnológica. *DEBATES-Cadernos do Programa de Pós-Graduação em Música*, v. 28, p. e282408-e282408, 2024. Disponível em: https://seer.unirio.br/revistadebates/article/view/13110. Acesso em: fev. 2026.

CASTRO, Armando Alexandre. Axé music: mitos, verdades e world music. *Per Musi*, n. 22, p. 203-217, 2010. DOI: 10.1590/S1517-75992010000200017. Disponível em: https://doi.org/10.1590/S1517-75992010000200017. Acesso em: fev. 2026.

CONCEIÇÃO, Julia Luiza et al. Applying supervised learning techniques to brazilian music genre classification. In: 2020 XLVI LATIN AMERICAN COMPUTING CONFERENCE (CLEI), 2020, Montevideo. *Proceedings [...].* IEEE, 2020. p. 102-107. DOI: 10.1109/CLEI52000.2020.00019. . Acesso em: fev. 2026.

CORREA, Marcio Guedes. O conceito de gênero musical no repertório e nas áreas de antropologia, comunicação, etnomusicologia e musicologia. *ARJ--Art Research Journal: Revista de Pesquisa em Artes*, v. 5, n. 2, 2018. DOI: 10.36025/arj.v5i2.17796. Disponível em: https://periodicos.ufrn.br/artresearchjournal/article/view/17796. Acesso em: fev. 2026.

DOSOVITSKIY, Alexey et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. Disponível em: https://openreview.net/forum?id=YicbFdNTTy. Acesso em: fev. 2026.

FERNANDES, Adriana. Music, migrancy, and modernity: a study of Brazilian forró. 2005. Dissertation -- University of Illinois at Urbana-Champaign, 2005.

GUIMARÃES, Victória de Souza Leon; FREITAS, Rosiane de. BYRM: Brazilian YouTube Regional Music Dataset. Zenodo, 2025. DOI: 10.5281/zenodo.16617888. Disponível em: https://zenodo.org/records/16617888. Acesso em: 21 mar. 2026.

GUIMARÃES, Victória de Souza Leon; KIENEN, João Gustavo; FREITAS, Rosiane de. Segment-based evaluation of music genre classification models with the BYRM dataset. In: SYMPOSIUM ON KNOWLEDGE DISCOVERY, MINING AND LEARNING (KDMiLe), 2025. Anais [...]. Porto Alegre: SBC, 2025. p. 97–104. Disponível em: https://sol.sbc.org.br/index.php/kdmile/article/view/37214. Acesso em: 21 mar. 2026.

HOPFIELD, John J. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, v. 79, n. 8, p. 2554-2558, 1982. DOI: 10.1073/pnas.79.8.2554. Disponível em: https://doi.org/10.1073/pnas.79.8.2554. Acesso em: fev. 2026.

KROGH, Mads. Music-genre abstraction, dissemination and trajectories. *Dansk Musikforskning Online*, v. 9, p. 83-108, 2019. Disponível em: http://www.danishmusicologyonline.dk/arkiv/arkiv_dmo/dmo_09/dmo_09_artikel_04.pdf. Acesso em: fev. 2026.

LeCUN, Yann et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, v. 86, n. 11, p. 2278-2324, 1998. DOI: 10.1109/5.726791. Disponível em: https://ieeexplore.ieee.org/document/726791. Acesso em: fev. 2026.

LUO, Xuejiao; HOVELING, Vera; EISEMANN, Elmar. Musicon: Glyph-based design for music visualization and retrieval. In: IEEE VISUALIZATION CONFERENCE (VIS), 2024. *Proceedings [...]*. IEEE, 2024. Disponível em: https://diglib.eg.org/server/api/core/bitstreams/de693a74-b9b0-4365-bc13-abfd7544ca89/content. . Acesso em: fev. 2026.

McINNES, Leland; HEALY, John; MELVILLE, James. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. Disponível em: https://arxiv.org/abs/1802.03426. Acesso em: fev. 2026.

SCHLUTER, Jan; GRILL, Thomas. Exploring data augmentation for improved singing voice detection with neural networks. In: ISMIR, 2015. *Proceedings [...]*. 2015. p. 121-126. Disponível em: https://dav.grrrr.org/public/pub/schlueter_grill-2015-ismir.pdf. . Acesso em: fev. 2026.

SHEIKH FATHOLLAHI, Mohamadreza; RAZZAZI, Farbod. Music similarity measurement and recommendation system using convolutional neural networks. *International Journal of Multimedia Information Retrieval*, v. 10, p. 43-53, 2021. DOI: 10.1007/s13735-020-00208-1. Disponível em: https://doi.org/10.1007/s13735-020-00208-1. Acesso em: fev. 2026.

SHU, Hongzhi et al. Benchmarking sub-genre classification for mainstage dance music. *arXiv preprint arXiv:2409.06690*, 2024. Disponível em: https://arxiv.org/abs/2409.06690. Acesso em: fev. 2026.

SILLA JR., Carlos N.; KAESTNER, Celso A. A.; KOERICH, Alessandro L. Automatic music genre classification using ensemble of classifiers. In: 2007 IEEE INTERNATIONAL CONFERENCE ON SYSTEMS, MAN AND CYBERNETICS, 2007. *Proceedings [...]*. IEEE, 2007. p. 1687-1692. DOI: 10.1109/ICSMC.2007.4414136.

SILVA, Angelo Cesar Mendes da; SILVA, Diego Furtado; MARCACINI, Ricardo Marcondes. 4MuLA: A multitask, multimodal, and multilingual dataset of music lyrics and audio features. In: PROCEEDINGS OF THE BRAZILIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB, 2020. *Proceedings [...]*. 2020. p. 145-148. Disponível em: https://sol.sbc.org.br/index.php/webmedia/article/view/13696. Acesso em: fev. 2026.

SILVA, Diego Furtado et al. On generalist and domain-specific music classification models and their impacts on brazilian music genre recognition. In: SIMPÓSIO BRASILEIRO DE COMPUTAÇÃO MUSICAL (SBCM), 2021b. *Anais [...]*. SBC, 2021. p. 60-67. Disponível em: https://sol.sbc.org.br/index.php/sbcm/article/view/19427. Acesso em: fev. 2026.

SILVA, Diego Furtado et al. On the fusion of multiple audio representations for music genre classification. In: SIMPÓSIO BRASILEIRO DE COMPUTAÇÃO MUSICAL (SBCM), 2021a. *Anais [...]*. SBC, 2021. p. 37-44. Disponível em: https://sol.sbc.org.br/index.php/sbcm/article/view/19423. Acesso em: fev. 2026.

SOUSA, Jefferson Martins de; PEREIRA, Eanes Torres; VELOSO, Luciana Ribeiro. A robust music genre classification approach for global and regional music datasets evaluation. In: 2016 IEEE INTERNATIONAL CONFERENCE ON DIGITAL SIGNAL PROCESSING (DSP), 2016. *Proceedings [...]*. IEEE, 2016. p. 109-113. DOI: 10.1109/ICDSP.2016.7868526.

SPIJKERVET, Janne; BURGOYNE, John Ashley. Contrastive learning of musical representations. *arXiv preprint arXiv:2103.09410*, 2021. Disponível em: https://archives.ismir.net/ismir2021/paper/000084.pdf. Acesso em: fev. 2026.

STURM, Bob L. The state of the art ten years after a state of the art: Future research in music information retrieval. *Journal of New Music Research*, v. 43, n. 2, p. 147-172, 2014. DOI: 10.1080/09298215.2014.894533. Disponível em: https://doi.org/10.1080/09298215.2014.894533. Acesso em: fev. 2026.

TURAB, Muhammad et al. Investigating multi-feature selection and ensembling for audio classification. *arXiv preprint arXiv:2206.07511*, 2022. Disponível em: https://arxiv.org/abs/2206.07511. Acesso em: fev. 2026.

VAN DER MAATEN, Laurens; HINTON, Geoffrey. Visualizing data using t-SNE. *Journal of Machine Learning Research*, v. 9, n. 11, p. 2579-2605, 2008. Disponível em: https://jmlr.org/papers/v9/vandermaaten08a.html. Acesso em: fev. 2026.

VASWANI, Ashish et al. Attention is all you need. *Advances in Neural Information Processing Systems*, v. 30, p. 5998-6008, 2017. Disponível em: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html. Acesso em: fev. 2026.

WOLD, Svante; ESBENSEN, Kim; GELADI, Paul. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, v. 2, n. 1-3, p. 37-52, 1987. DOI: 10.1016/0169-7439(87)80084-9. Disponível em: https://doi.org/10.1016/0169-7439(87)80084-9. Acesso em: fev. 2026.

YONAY, Ori; HAMMOND, Tracy; YANG, Tianbao. Myna: Masking-based contrastive learning of musical representations. *arXiv preprint arXiv:2502.12511*, 2025. Disponível em: https://arxiv.org/abs/2502.12511. Acesso em: fev. 2026.

ZHANG, Xin; RAS, Zbigniew W. Analysis of sound features for music timbre recognition. In: 2007 INTERNATIONAL CONFERENCE ON MULTIMEDIA AND UBIQUITOUS ENGINEERING (MUE'07), 2007. *Proceedings [...]*. IEEE, 2007. p. 3-8. DOI: 10.1109/MUE.2007.85.

Acknowledgments

This work was partially supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES-PROEX) - Finance Code 001, the National/Brazilian Council for Scientific and Technological Development (CNPq), and also

supported by Amazonas State Research Support Foundation - FAPEAM - through the POSGRAD project 2024/2025.

Authorship contribution

Victória de Souza Leon Guimarães: Conceptualization, Methodology, Software, Data Curation, Formal Analysis, Investigation, Writing – Original Draft, Writing – Review & Editing.

João Gustavo Kienen: Conceptualization, Writing – Review & Editing, Supervision (Music Theory).

Rosiane de Freitas Rodrigues: Conceptualization, Writing – Review & Editing, Supervision, Project Administration, Funding Acquisition.

Financing

This work was funded by CAPES, CNPq, and FAPEAM.

Use license

This article is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

Publisher

Federal University of Goiás. School of Music and Performing Arts. Graduate Program in Music. Publication in the Portal of Periodicals UFG.

The ideas expressed in this article are the responsibility of their authors, and do not necessarily represent the opinion of the editors or the university.