Understanding genre similarity in Brazilian music through Vision Transformer embeddings

Victória Guimarães[13] **, João Gustavo Kienen**[2] **, Rosiane de Freitas** [13]

[1] Programa de Pós-Graduação em Informática – Universidade Federal do Amazonas
Manaus, AM – Brazil

[2]Faculdade de Artes – Universidade Federal do Amazonas
Manaus, AM – Brazil

[3]Instituto de Computação – Universidade Federal do Amazonas
Manaus, AM – Brazil

vsg@icomp.ufam.edu.br, gustavokienen@ufam.edu.br, rosiane@icomp.ufam.edu.br

Abstract. *This work investigates how musical genre similarity is represented in the latent space learned by a Vision Transformer (ViT) model fine-tuned on Brazilian regional music. We present BYRM, a curated dataset that contains 1,082 tracks across ten culturally diverse genres. Our analysis focuses on the best-performing configuration identified in previous experiments, which uses 10-second segments extracted from the 90 to 120 second excerpt of each track. Mel-spectrograms were used as input to train the ViT, from which time-local embeddings were extracted. Dimensionality reduction techniques (PCA, t-SNE, and UMAP) were applied to visualize the latent space, and cosine similarity was computed to assess inter-genre proximity. The model achieved 81.94% accuracy and 81.84% F1-score under the best configuration, demonstrating its ability to learn discriminative representations. The similarity analysis shows that the ViT effectively captures stylistic relationships between related genres, such as samba and pagode or vaneira and xote gaúcho, while maintaining separation from more distinct styles like Brazilian rock. These findings provide new insights into genre similarity modeling using transformer-based embeddings in a culturally rich and musically complex context.*

1. Introduction

Research in Music Information Retrieval (MIR) encompasses a wide range of tasks, including the classification of musical notes, emotions, instruments, artists, and musical genres [1]. Among these, Music Genre Recognition (MGR) plays a central role, enabling the organization and retrieval of musical content in large-scale systems. Although genre classification has evolved significantly through the adoption of supervised machine learning techniques [2], the task remains challenging due to the subtle stylistic elements that differentiate musical genres [3].

This challenge becomes particularly evident in the context of regional music, where cultural influences, shared rhythmic structures, and overlapping instrumentation contribute to high inter-genre similarity [4, 5]. In Brazilian music, for example, genres such as samba and pagode, or forró and vaneira, exhibit strong musical affinities that are difficult to separate through conventional classification techniques [6, 7]. The presence of hybrid tracks, genre mixtures, and stylistic fluidity further complicates this scenario [8, 9].

To address these complexities, recent approaches have incorporated deep learning models capable of capturing high-level representations directly from audio signals. Transformer architectures [10], and particularly the Vision Transformer (ViT) [11], have shown promising results in capturing global dependencies and modeling fine-grained patterns in spectrogram-based inputs. ViT processes spectrogram patches as visual tokens, enabling the model to organize musical information hierarchically and to encode latent features that may reflect genre characteristics in a learned embedding space.

Unlike traditional approaches that rely solely on accuracy metrics, this work shifts the focus toward understanding how Brazilian regional genres are organized within the latent space learned by a fine-tuned Vision Transformer (ViT) model. Drawing on insights from music cognition, which highlights the importance of choruses and melodic hooks as salient and genre-defining segments [12], we analyze embeddings extracted from strategically selected excerpts using cosine similarity and dimensionality reduction techniques to examine how Brazilian regional genres are distributed in the latent space and to investigate the similarity captured by the model. To support this investigation, we developed BYRM (Brazilian YouTube Regional Music), a new dataset containing tracks from representative regional genres. This dataset constitutes one of the main contributions of the study and is publicly available to foster reproducibility and future research in music genre classification [13].

2. Theoretical Background

Music Information Retrieval (MIR) is a research field dedicated to developing computational methods for accessing, analyzing, and organizing musical content [1]. One of its primary applications is Music Genre Recognition (MGR), which involves the identification of musical styles based on audio features. Early approaches rely on handcrafted descriptors such as spectral centroid, bandwidth, zero-crossing rate, and Mel Frequency Cepstral Coefficients (MFCCs), which approximate perceptual aspects of the human auditory system [14, 15]. In deep learning

pipelines, Mel-spectrograms have become the standard input representation, converting audio into time-frequency images aligned with the Mel scale [16].

Deep learning has significantly advanced MGR by enabling direct learning from spectrograms. Convolutional Neural Networks (CNNs) capture local time-frequency patterns, Recurrent Neural Networks (RNNs) model temporal dependencies, and Transformer-based models leverage self-attention to learn global relationships in the data [17, 18, 10]. The Vision Transformer (ViT) [11], originally proposed for image classification, has been adapted to music analysis by treating spectrogram patches as input tokens. This enables the model to capture long-range dependencies and build hierarchical representations of musical signals. This view aligns with findings in music cognition, which emphasize that certain segments, such as choruses, tend to be more salient and memorable, often carrying the genre-defining elements of a song [12].

The internal representations known as embeddings, learned by these models, encode latent features of audio that go beyond classification outputs. These fixed-length vectors allow for the investigation of genre similarity by analyzing the distribution and proximity of examples in the embedding space. Visualization techniques such as Principal Component Analysis (PCA) [19], t-Distributed Stochastic Neighbor Embedding (t-SNE) [20], and Uniform Manifold Approximation and Projection (UMAP) [21] are often applied to interpret these high-dimensional representations. While PCA preserves global variance, t-SNE emphasizes local neighborhoods, and UMAP balances both with better scalability. For quantitative analysis, similarity metrics such as cosine distance can be used directly in the original embedding space [22, 23].

Understanding the musical genre as a social construct shaped by cultural, aesthetic, and institutional factors is essential for interpreting these embeddings. Genres are not fixed or clearly delimited; they emerge through abstraction processes involving symbolic and contextual elements [24]. This dynamic construction shows how deep learning models represent genres, not through explicit rules, but through statistical regularities learned from data. From this perspective, genre embeddings are computational abstractions that encode musico-cultural patterns. As genre categories are performed and reconfigured socially, embeddings also reflect contextual similarities learned from diverse audio examples.

3. Related Work

Table 1 summarizes the main related works that explore musical embeddings and the analysis of similarity between audio samples. The column "MGR" indicates whether the study addresses music genre recognition. "Similarity?" shows whether similarity between tracks, genres, or artists is considered. The "Model" column describes the architecture used, such as CNNs or Vision Transformers. "Dim. Reduction" indicates whether dimensionality reduction techniques were applied to visualize or analyze embeddings. "Similarity Technique" refers to the metric used to compare embeddings, and "Dataset" identifies the collections adopted in each study.

Although not focused on genre recognition, some studies contribute valuable insights through the analysis of embeddings and musical similarity. One study applies Vision Transformers and UMAP to visualize embedding clusters, relying on qualitative inspection rather than explicit similarity metrics such as cosine or Euclidean distance [25]. In the musical domain, another work explores stylistic relationships between artists by extracting embeddings from Jukebox and comparing them using cosine and Euclidean distances, with PCA and t-SNE used to project the embeddings for visual analysis [23]. A third approach focuses on contrastive learning with spectrograms using a Vision Transformer, aiming to build general-purpose musical representations. While not dedicated to genre classification, this model is later evaluated on datasets such as GTZAN and MagnaTagATune to assess the quality of the learned embeddings [26].

Among the studies that address music genre recognition, there are notable differences in how embeddings and similarity are incorporated. The work presented in [27] proposes a CNN-based classifier and builds a music recommender system using embeddings extracted from intermediate layers, compared through cosine and Euclidean distance. In [22], the CLMR model is trained via contrastive learning and evaluated on genre classification tasks using MagnaTagATune, with similarity implicitly modeled through a cosine-based loss function. The benchmark proposed in [28] focuses on the classification of EDM sub-genres using CNNs and Vision Transformers. Although it does not compute similarity between embeddings, the study includes visualizations using PCA, t-SNE, and UMAP to examine genre separability. The Musicon system [29] builds upon CLMR embeddings to enable music retrieval through cosine similarity, combining dimensionality reduction methods with an interactive visual interface for genre-based exploration. Together, these works demonstrate diverse approaches to genre classification, where similarity may be used as a training signal, a visualization tool, or a retrieval mechanism.

Beyond the works summarized in Table 1, other studies have advanced Brazilian music genre recognition using different approaches. Classical methods achieved 79.7% accuracy with SVM and 86.1% on a regional dataset [30], while Random Forest obtained 65.3% on Brazilian genres and 64.2% when combined with GTZAN [31]. Subsequent CNN-based approaches proposed feature fusion strategies, reaching 88.5% accuracy on GTZAN [32], and domain-specific models that outperformed generalist ones on the 4MuLA dataset [33]. The 4MuLA collection later became an important multimodal benchmark with over 96,000 tracks in 76 genres [34]. In contrast to these works, this study focuses on the analysis of embeddings and the calculation of similarity between Brazilian regional music genres, leveraging the BYRM dataset specifically developed for this purpose.

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

Work	MGR	Similarity?	Model	Dim. Reduction	Similarity Technique	Dataset
Visual Embeddings (2022) [25]		✓	Vision Transformer	UMAP	Visual Cluster	Craigslist and OfferUp
Audio Influence (2024) [23]	✗	✓	Jukebox (OpenAI)	PCA, t-SNE	Euclidean and Cosine	Million Song Dataset, Spotify
Myna (2025) [26]		✗	Vision Transformer	✗	✗	AudioSet, GTZAN, MTT, EmoMusic
Music Recommender (2021) [27]		✓	CNN	✗	Euclidean and Cosine	GTZAN, Emotify.
CLMR (2021) [22]	✓	✓	CNN	✗	Cosine	MagnaTagATune
EDM Benchmark (2024) [28]		✗	CNN, Vision Transformer	PCA, t-SNE, UMAP	✗	EDM Benchmark
Musicon (2024) [29]		✓	CLMR	UMAP, PCA, t-SNE, others	Cosine	GTZAN, MagnaTagATune, others
This work	✓	✓	**Vision Transformer**	**PCA, t-SNE, UMAP**	**Cosine**	**BYRM (Brazilian)**

4. Methodology

This section describes the methodological steps adopted in this study, from dataset preparation to embedding analysis. We first present the BYRM dataset[1], a curated collection of Brazilian regional music organized into fixed temporal segments. Then, we detail the preprocessing pipeline and the extraction of Mel-spectrograms used as input to a Vision Transformer (ViT). Finally, we explain how time-local embeddings were extracted and analyzed using dimensionality reduction and cosine similarity to investigate genre proximity in the learned representation space.

4.1. BYRM Dataset

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.

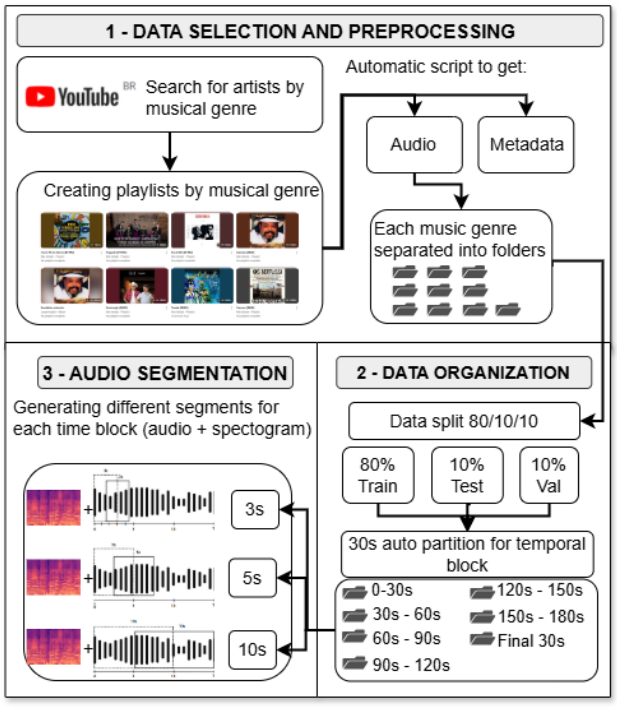


Figure 1: Overview of the BYRM dataset preparation process.

The tracks were collected from publicly available YouTube albums featuring representative artists of each genre. Then, audio was collected using the Pytube library, converted to WAV format, and accompanied by metadata such as video title, channel ID, and file path. Tracks were split into training, validation, and test sets (80/10/10) at the track level to prevent data leakage. Each track was segmented into six non-overlapping 30-second excerpts (0–30s to 150–180s). To prepare the data for model training, each 30-second excerpt was subdivided into overlapping windows of 3, 5, and 10 seconds, using a 50% overlap. This strategy enables the capture of rhythmic, harmonic, and timbral patterns at multiple time scales, supporting controlled studies of genre similarity.

4.1.1. Preprocessing and Feature Extraction

Mel-spectrograms were computed for each short segment and used as input representations to the Vision Transformer model. Spectrograms were generated using an FFT window of 2048 samples, a hop length of 512, 128 Mel frequency bands, and a sampling rate of 22,050 Hz. A power parameter of 2.0 was used to enhance the spectral contrast, which corresponds to the default configuration in the Librosa library. These spectrograms provide high-resolution time-frequency information from localized musical contexts and serve as the basis for embedding extraction.

4.1.2. Models

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.

Aspect	ViT	ResNet50	SVM
Pretrained	✓ (ImageNet)	✓ (ImageNet)	✗
Input	Mel-spectrogram	Mel-spectrogram	Handcrafted features
Segments Used	3s, 5s, 10s	3s	3s
Excerpts Used	All	Best and worst	Best and worst
Features	Spectrogram image	Spectrogram image	MFCC, chroma, spectral centroid, bandwidth, rolloff, ZCR
Milestone Epoch	10/30/40	15/100/150	✗
Dropout	30% head, 7% internal	Default	✗
Epochs	100	200	✗
Patience	10	15	✗
Optimizer	Adam (lr = 1e-4)	Adam (lr = 1e-4)	✗
LR Decay	$0.8 \rightarrow 0.6 \rightarrow 0.3$	$0.8 \rightarrow 0.6 \rightarrow 0.3$	✗

The ViT model was fine-tuned on Mel-spectrograms using segments from the 90–120s block,

[1]The BYRM dataset is available at: `https://zenodo.org/records/16617888`

with window sizes of 3s, 5s, and 10s. Embeddings were extracted from the [CLS] token output of the final transformer layer. ResNet50 was trained on 3-second spectrogram images from the best and worst-performing temporal excerpts. The SVM was trained using handcrafted features (MFCCs, chroma, spectral centroid, bandwidth, rolloff, and ZCR) extracted from the same segments. Although SVM do not produce internal embeddings comparable to those of the ViT, their performance served as a comparative reference.

4.1.3. Experimental Setup

Previous experiments conducted by the authors identified the most and least informative segments of the BYRM dataset for genre classification, corresponding to the excerpts from 90 to 120 seconds and from 0 to 30 seconds, respectively. Among these, the excerpt from 90 to 120 seconds consistently achieved the highest performance across all models and segment durations. Furthermore, experiments using segment based windows of 3, 5, and 10 seconds indicated that the 10 second segments provided the best classification results when applied to this excerpt. Based on these findings, the present analysis focuses on this configuration, which is the excerpt from 90 to 120 seconds with 10 second segments, as a case study for investigating genre similarity and the organization of the embedding space.

To investigate how genre information is structured in the embedding space, we extracted embeddings from the [CLS] token of the ViT, a special classification token that aggregates global information from all input patches. These embeddings were obtained both before fine-tuning, to qualitatively observe how genres behave prior to the model learning their distinctive characteristics, and after fine-tuning, to examine how training reshapes their organization in the latent space. To further analyze their spatial organization, we employed dimensionality reduction techniques: PCA, t-SNE, and UMAP. Each method was applied independently to the embeddings extracted from pre and post training. These projections allow to examine how genres cluster spatially and how training impacts the separability between classes. In addition to the visual analysis, we computed cosine similarity between embeddings to assess inter-genre relationships. Specifically, we calculated the mean pairwise similarity between all embeddings belonging to each pair of genres, resulting in a similarity matrix visualized as a heatmap. This analysis aims to reveal how time-local embeddings capture both the proximity and distinctiveness of Brazilian regional genres within the embedding space.

5. Results

To evaluate how the model represents genre information in the embedding space, we begin by reporting the classification results that guided the selection of the best configuration (90–120s excerpt, 10s segments). Based on this setup, we then explore the embedding space through visualizations and cosine similarity to analyze how genres are organized and related in the latent space.

5.1. Classification Performance

To further explore the temporal characteristics of genre representation, this experiment investigates how segment duration impacts classification performance across different parts of the song. ResNet50 and SVM models were used as baselines, and the Vision Transformer (ViT) consistently outperformed both. Previous experiments identified the best (90–120s) and worst (0–30s) excerpts for genre classification [35]. Table 3 summarizes the ViT performance across three segment durations applied to both excerpts.

Table 3: Performance of ViT on different segment durations for the best (90–120s) and worst (0–30s) excerpts.

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

The best performance is achieved with 10-second segments extracted from the 90 to 120 second excerpt of each track, reaching 81.94% accuracy and 81.84% F1 score. This result suggests that longer segments from more central portions of the track capture richer and more stable genre-defining features. The final minute and a half of the songs may contain recurring motifs, refrains, or instrumental arrangements that are more representative of the musical identity of the genre. The 10-second window also appears to provide a balance between temporal resolution and contextual information, allowing the Vision Transformer to learn more expressive and discriminative embeddings. This outcome reinforces the importance of carefully selecting both segment duration and temporal positioning when analyzing musical similarity through deep learning models, especially in contexts involving high inter-genre overlap. To complement the classification metrics, we analyzed the confusion matrix obtained under the best configuration (Figure 2).

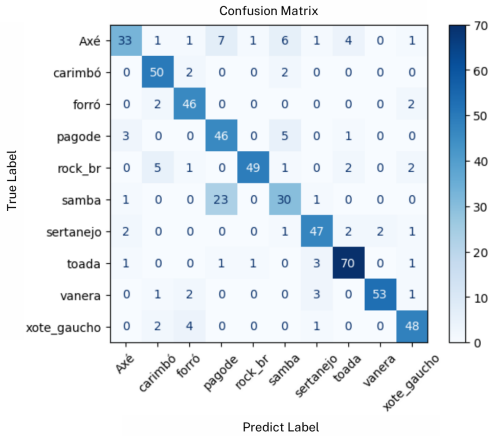


Figure 2: Confusion matrix of the ViT model using 10-second segments from the 90–120s excerpt. Rows represent true labels and columns predicted labels.

Some misclassifications occurred between stylistically related genres, such as samba and pagode or axé and samba, reflecting known rhythmic and harmonic similarities. In contrast, other closely related styles, such as vaneira and xote gaucho, were not frequently confused. This indicates that classification results alone may not fully capture inter-genre similarity. It is also worth noting that the confusion matrix is asymmetric, as it reflects directional classification errors (e.g., samba may be more frequently misclassified as pagode than the other way around). Therefore, in the following section, we investigate the embedding space learned by the ViT model to visualize and better understand how genre relationships are organized at a representational level.

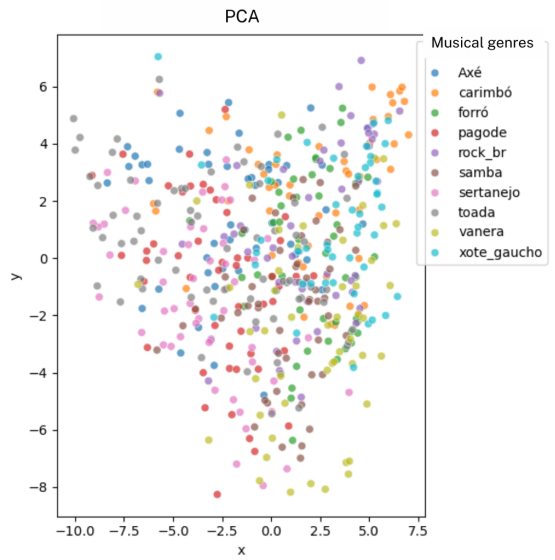
5.2. Embedding Space Visualization

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 presents the embedding space before training, using the pretrained ViT model with weights from ImageNet. At this stage, the model had not yet been exposed to the BYRM dataset, meaning it was not familiar with the musical genres, which explains why they appear more dispersed in the vector space. However, some proximity between genres can already be observed, possibly due to shared low-level acoustic features captured by the pretrained model.

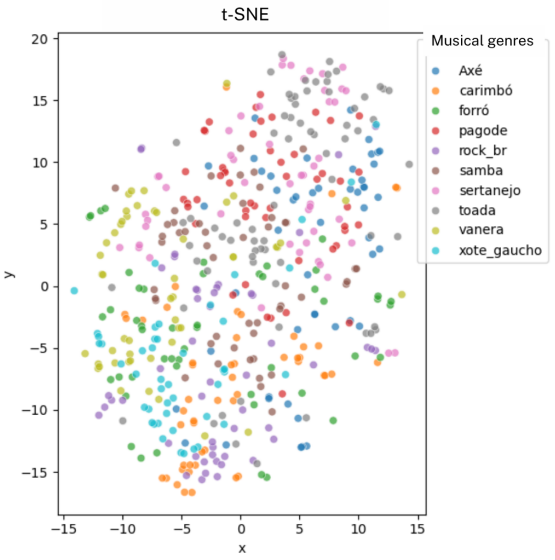
On the other hand, Figure 4 presents the embeddings generated from the BYRM test set after training the model. In this case, the embeddings form a much more structured space. Clear genre-specific clusters emerge, indicating that the model has learned meaningful representations that reflect stylistic 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, with well-defined clusters for all genres. Among the UMAP results shown in Figure 4c, the embeddings present a clear spatial organization, revealing how the model internalized the relationships between musical genres. A qualitative inspection highlights the following patterns:

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.

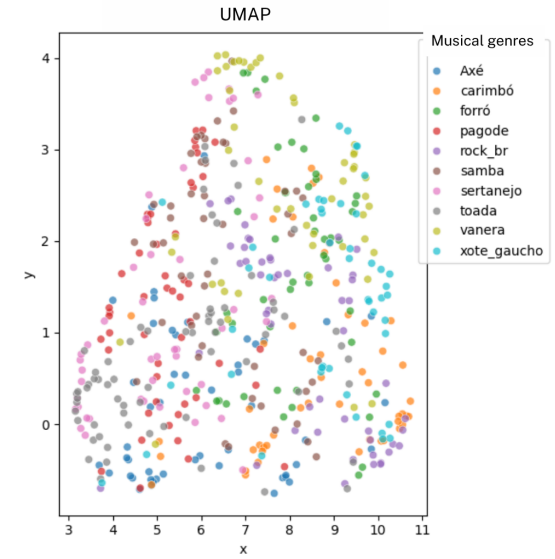
Genres with partial separation and overlap: Axé appears between toada, sertanejo, and carimbó, sug-



(a) PCA

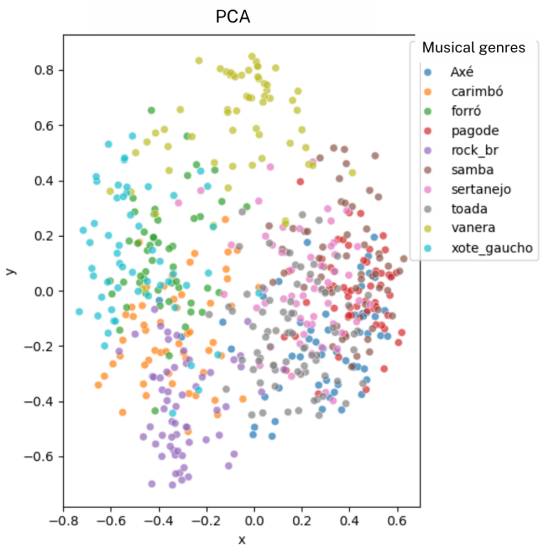


(b) t-SNE

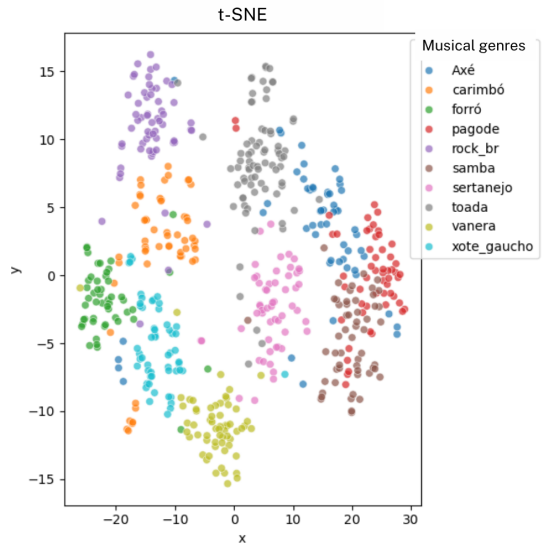


(c) UMAP

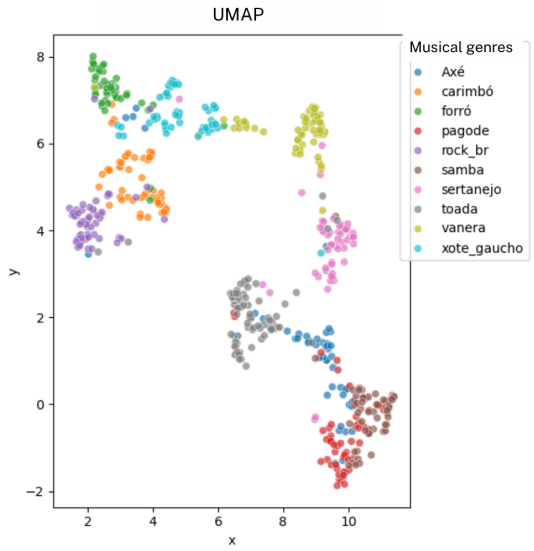
Figure 3: Visualization of the validation embedding space before training using PCA, t-SNE, and UMAP.



(a) PCA



(b) t-SNE



(c) UMAP

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.

gesting 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.

5.3. Genre Similarity Analysis

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.

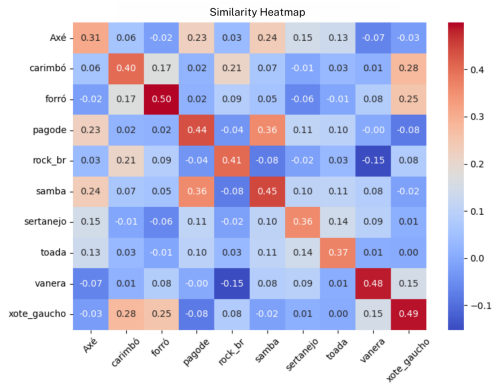


Figure 5: Cosine similarity heatmap between musical genres based on Vision Transformer embeddings.

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 vanera. These distances reflect the substantial differences in instrumentation and musical structure, particularly the presence of electric timbres and non-traditional rhythms in Brazilian rock. The high similarity between samba and pagode remains evident, as does the consistency between forró and carimbó. However, important nuances start to emerge. For instance, axé shows reduced similarity with other genres, suggesting that the model is able to capture its distinctive acoustic characteristics more precisely, making its vector representation more distinct. Unlike the confusion matrix, this similarity heatmap is symmetric by definition, since cosine similarity between two genres is identical in both directions.

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. In other words, the moderate values on the diagonal do not weaken the dataset, but rather highlight its richness and realism. 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 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.

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

[1] Jean-Julien Aucouturier and Elias Pampalk. Introduction–from genres to tags: A little epistemology of music information retrieval research. *Journal of New Music Research*, 37(2):87–92, 2008.

[2] Carlos N Silla Jr, Celso AA Kaestner, and Alessandro L Koerich. Automatic music genre classification using ensemble of classifiers. In *2007 IEEE International Conference on Systems, Man and Cybernetics*, pages 1687–1692. IEEE, 2007.

[3] Bob L Sturm. The state of the art ten years after a state of the art: Future research in music information retrieval. *Journal of new music research*, 43(2):147–172, 2014.

[4] Raul de Araújo Lima, Rômulo César Costa de Sousa, Hélio Lopes, and Simone Diniz Junqueira Barbosa. Brazilian lyrics-based music genre classification using a blstm network. In *International Conference on Artificial Intelligence and Soft Computing*, pages 525–534. Springer, 2020.

[5] Szymon Muszynski and Zbigniew Tarapata. Methods of automated music comparison based on multi-objective metrics of network similarity. *Applied Sciences*, 13(6):3567, 2023.

[6] Lucas Cassano, Elisa de Oliveira Morais Nacur Cassano, and Robert Moses Pechman. Passado e presente da música carioca: Narrativas urbanas e evolução tecnológica. *DEBATES-Cadernos do Programa de Pós-Graduação em Música*, 28:e282408–e282408, 2024.

[7] Adriana Fernandes. *Music, migrancy, and modernity: a study of Brazilian forró*. University of Illinois at Urbana-Champaign, 2005.

[8] Marcio Guedes Correa. 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*, 5(2), 2018.

[9] Armando Alexandre Castro. Axé music: mitos, verdades e world music. *Per Musi*, pages 203–217, 2010.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[12] Timothy P Byron, Carl T Rushworth, and Maxwell J Stewart. Popular music excerpts are rated as more memorable and salient if they involve vocals, compound hooks, and choruses. *Music Perception: An Interdisciplinary Journal*, 42(3):197–206, 2025.

[13] Victória de Souza Guimarães and Rosiane de Freitas. Byrm: Brazilian youtube regional music dataset, 2025.

[14] Xin Zhang and Zbigniew W Ras. Analysis of sound features for music timbre recognition. In *2007 International*

Conference on Multimedia and Ubiquitous Engineering (MUE'07), pages 3–8. IEEE, 2007.

[15] Muhammad Turab, Teerath Kumar, Malika Bendechache, and Takfarinas Saber. Investigating multi-feature selection and ensembling for audio classification. *arXiv preprint arXiv:2206.07511*, 2022.

[16] Jan Schlüter and Thomas Grill. Exploring data augmentation for improved singing voice detection with neural networks. In *ISMIR*, pages 121–126, 2015.

[17] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[18] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[19] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[20] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[21] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[22] Janne Spijkervet and John Ashley Burgoyne. Contrastive learning of musical representations. *arXiv preprint arXiv:2103.09410*, 2021.

[23] Julia Barnett, Hugo Flores Garcia, and Bryan Pardo. Exploring musical roots: Applying audio embeddings to empower influence attribution for a generative music model. *arXiv preprint arXiv:2401.14542*, 2024.

[24] Mads Krogh. Music-genre abstraction, dissemination and trajectories. *Dansk Musikforskning Online*, 9:83–108, 2019.

[25] Cameron Armijo and Pablo Rivas. Exploring visual embedding spaces induced by vision transformers for online auto parts marketplaces. *arXiv preprint arXiv:2502.05756*, 2025.

[26] Ori Yonay, Tracy Hammond, and Tianbao Yang. Myna: Masking-based contrastive learning of musical representations. *arXiv preprint arXiv:2502.12511*, 2025.

[27] Mohamadreza Sheikh Fathollahi and Farbod Razzazi. Music similarity measurement and recommendation system using convolutional neural networks. *International Journal of Multimedia Information Retrieval*, 10:43–53, 2021.

[28] Hongzhi Shu, Xinglin Li, Hongyu Jiang, Minghao Fu, and Xinyu Li. Benchmarking sub-genre classification for mainstage dance music. *arXiv preprint arXiv:2409.06690*, 2024.

[29] Xuejiao Luo, Vera Hoveling, and Elmar Eisemann. Musicon: Glyph-based design for music visualization and retrieval. 2024.

[30] Jefferson Martins De Sousa, Eanes Torres Pereira, and Luciana Ribeiro Veloso. A robust music genre classification approach for global and regional music datasets evaluation. In *2016 IEEE international conference on digital signal processing (DSP)*, pages 109–113. IEEE, 2016.

[31] Júlia Luiza Conceição, Rosiane de Freitas, Bruno Gadelha, João Gustavo Kienen, Sérgio Anders, and Brendo Cavalcante. Applying supervised learning techniques to brazilian music genre classification. In *2020 XLVI Latin American Computing Conference (CLEI)*, pages 102–107. IEEE, 2020.

[32] Diego Furtado Silva, Micael Valterlânio da Silva, Ricardo Szram Filho, and Angelo Cesar Mendes da Silva. On the fusion of multiple audio representations for music genre classification. In *Simpósio Brasileiro de Computação Musical (SBCM)*, pages 37–44. SBC, 2021.

[33] Diego Furtado Silva, Angelo Cesar Mendes da Silva, Luís Felipe Ortolan, and Ricardo Marcondes Marcacini. On generalist and domain-specific music classification models and their impacts on brazilian music genre recognition. In *Simpósio Brasileiro de Computação Musical (SBCM)*, pages 60–67. SBC, 2021.

[34] Angelo Cesar Mendes da Silva, Diego Furtado Silva, and Ricardo Marcondes Marcacini. 4mula: A multitask, multimodal, and multilingual dataset of music lyrics and audio features. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, pages 145–148, 2020.

[35] V. Guimarães, J. G. Kienen, and R. de Freitas. Segment-based evaluation of music genre classification models with the byrm dataset. In *Proceedings of the Symposium on Knowledge Discovery, Mining and Learning (KDMiLe)*, 2025. To appear.