Segment-based evaluation of music genre classification models with the BYRM Dataset

Victória Guimarães[1], João Gustavo Kienen[2], Rosiane de Freitas[1]

[1] Instituto de Computação, Universidade Federal do Amazonas, Brazil
{vsg,rosiane}@icomp.ufam.edu.br
[2] Faculdade de Artes, Universidade Federal do Amazonas, Brazil
gustavokienen@ufam.edu.br

Abstract. The increasing use of deep learning in music information retrieval has driven progress in music genre classification, yet many studies overlook the temporal structure of music. Most benchmark datasets provide only a single, fixed excerpt per song, limiting the investigation of how time-related factors influence classification. In this work we introduce the Brazilian YouTube Regional Music (BYRM) Dataset, a curated collection of Brazilian regional music comprising ten genres, with multiple excerpts extracted from different parts of each track. The dataset supports controlled experiments on how both excerpt position and segment duration affect model performance. BYRM includes vectorized features, preprocessed spectrograms, and metadata for reproducibility. To evaluate the dataset, we conduct experiments using a Vision Transformer (ViT), supported by SVM and ResNet50 baselines. Results show that excerpts from the middle of the song (e.g., 90 to 120 seconds) yield better performance, and that optimal segment duration varies by genre. BYRM enables fine-grained analysis of genre-specific temporal patterns and supports future research on temporal modeling and genre similarity.

CCS Concepts: • **Computing methodologies** → **Machine learning algorithms**.

Keywords: Brazilian music, music genre classification, music segmentation, temporal structure, Vision Transformer

1. INTRODUCTION

Research in Music Information Retrieval (MIR) encompasses various fields of study, including tasks such as the classification of musical notes, emotions, artists, and musical genres [Aucouturier and Pampalk 2008]. Music Genre Recognition (MGR), in particular, has been widely studied using supervised learning algorithms [Silla Jr et al. 2007]. Recent advances include deep models like CNNs [Meng 2024], RNNs [Dai et al. 2016], and Transformers [Vaswani et al. 2017], which model long-range dependencies via self-attention. The Vision Transformer (ViT) [Dosovitskiy et al. 2020], originally developed for image classification, treats spectrogram patches as tokens and shows potential for music genre recognition.

Many studies on genre recognition rely on datasets that ignore the temporal structure of music, often using a fixed 30 second excerpt per track. Datasets like GTZAN [Sturm 2013] and ISMIR [ISMIR 2004] provide only one excerpt per song, and even larger collections such as FMA [Defferrard et al. 2016] do not specify the excerpt position. Most recent works follow this approach, using 30 second segments to extract features like MFCCs or spectrograms. In contrast, research in music cognition shows that segments such as choruses or melodic hooks are more salient and representative [Byron et al. 2025], reinforcing the need to examine how excerpt position and duration affect classification performance.

In this work, we introduce the Brazilian YouTube Regional Music (BYRM) Dataset, a benchmark focused on Brazilian regional genres with high internal similarity, which makes the classification task more challenging. It includes multiple 30-second excerpts per track, with vectorized features and spectrograms suitable for deep learning. The Dataset was designed to test two hypotheses: (1) whether the position of the excerpt within the song influences classification performance, and (2) whether the segment duration (3s, 5s, 10s) impacts model accuracy. Experiments were conducted using a Vision Transformer (ViT), a Support Vector Machine (SVM), and a CNN (ResNet50) to compare architectures and evaluate result consistency.

2. THEORETICAL BACKGROUND

Music Information Retrieval (MIR) is a research field focused on developing methods to access and organize musical content [Aucouturier and Pampalk 2008]. One of its most explored applications is Music Genre Recognition (MGR), which involves extracting audio features that reflect spectral and temporal characteristics. Handcrafted features like spectral centroid and bandwidth capture timbral properties [Zhang and Ras 2007], while zero-crossing rate reflects rhythmic content [Turab et al. 2022]. MFCCs approximate human hearing, and Mel-spectrograms [Cheng et al. 2020] enable time-frequency representations suited for computer vision models.

Deep learning has become standard in audio classification. CNNs extract local time-frequency patterns from spectrograms [LeCun et al. 1998], while Transformer-based models [Vaswani et al. 2017] capture global relationships via self-attention. The Vision Transformer (ViT) [Dosovitskiy et al. 2020], originally designed for images, has been applied to spectrograms by treating patches as tokens to learn long-range dependencies in music.

Musical genre is a socially constructed label influenced by cultural and stylistic patterns [Fabbri et al. 1982]. Classifying genres is particularly challenging due to overlapping acoustic traits, especially in regional music where rhythmic and instrumental similarities are common [Cerati 2021]. Studies in music cognition suggest that specific segments, such as choruses or melodic motifs, are more salient for genre recognition [Byron et al. 2025]. However, most datasets, including GTZAN [Sturm 2013], ISMIR [ISMIR 2004], and FMA [Defferrard et al. 2016], rely on a fixed excerpt per track, often ignoring how segment duration and position influence classification. This has led to the adoption of segment-based strategies that explore these temporal factors more systematically.

3. RELATED WORK

Most related works rely on a single excerpt per track, usually from the beginning, limiting the analysis of temporal variation. Table I highlights how our approach differs by using multiple excerpts and explicitly considering temporal positioning to better capture representative genre characteristics.

Table I: Comparison of related works in music genre classification.

Work	Segment Strategy	Segment Length	Technique	Track Duration	Temporal Pos.	Dataset
[Barbosa et al. 2015]	Fixed	30s	SVM, KNN, others.	30s	✗	Brazilian
[De Sousa et al. 2016]			SVM		✗	GTZAN + Brazilian
[Cheng et al. 2020]			CNN		✗	GTZAN
[Medhat et al. 2020]			MCNN		✗	GTZAN, ISMIR
[Wijaya et al. 2024]			BiLSTM		✗	GTZAN, ISMIR
[Zhuang et al. 2020]	Short	3s	Transformer	30s	✗	GTZAN
[Xie et al. 2024]		5s	CNN + Transformer		✗	GTZAN
[da Conceição et al. 2020]	Variable	30s, 40s, 60s, 120s	KNN, SVM, others.	30s, 40s, 60s, 120s	✓	Brazilian
[Dai et al. 2016]	Unspecified	Not mentioned	LSTM	Full track	✗	ISMIR
[Silva et al. 2021]			KNN, SVM, others.	Full track	✗	Brazilian
[Zhao et al. 2022]			Swin Transformer	30s	✗	GTZAN, FMA
This work (BYRM)	**Variable**	**3s, 5s, 10s**	**ViT, ResNet, SVM**	**Full track (30s per excerpt)**	✓	**BYRM (Brazilian)**

Some works use full tracks or unspecified segments, such as [Dai et al. 2016], [Silva et al. 2021], and [Zhao et al. 2022], without exploring temporal positioning. Others adopt a fixed 30-second segment, like [Barbosa et al. 2015], [De Sousa et al. 2016], [Cheng et al. 2020], [Medhat et al. 2020], and [Wijaya et al. 2024], typically taken from the beginning of the track. Short-segment strategies (3 to 5 seconds) are used by [Zhuang et al. 2020] and [Xie et al. 2024], but still without temporal variation. Only [da Conceição et al. 2020] explores different durations and positions of the excerpts, combining Brazilian genres with GTZAN. In contrast, our segment-based approach systematically evaluates multiple segment durations (3s, 5s, and 10s) across different positions in the track using ViT, SVM, and ResNet, revealing where genre-specific patterns are most effectively captured.

4. DATASET CONSTRUCTION METHODOLOGY

The Brazilian YouTube Regional Music (BYRM[1]) dataset was built by adapting the four phases of the Knowledge Discovery in Databases (KDD) process [Fayyad 1997], given its focus on extracting structured knowledge from raw data. The adapted stages include: (1) selection and preprocessing, (2) organization, (3) transformation, and (4) presentation. Figure 1 illustrates the complete pipeline. BYRM comprises 10 genres representative of different regions of Brazil: toada, carimbó, axé, forró, sertanejo, rock brasileiro, samba, pagode, vaneira, and xote gaúcho. The dataset is publicly available to support reproducibility and foster future research in music genre classification [Guimarães and Freitas 2025]. The following items detail each stage, describing the methods and procedures used to ensure a robust and well-structured dataset.

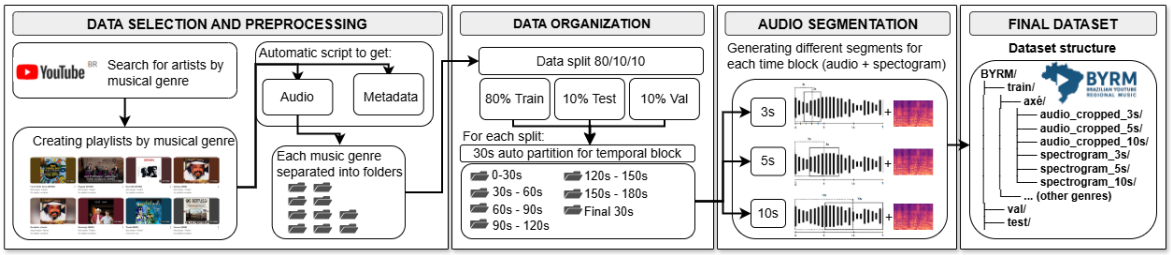

Fig. 1: BYRM dataset creation pipeline, from YouTube audio collection to temporal segmentation.

(1) **Data Selection and Preprocessing:** A total of 1,082 tracks were collected from YouTube albums using Pytube. An automated script saved the audio in WAV format and metadata (video name, ID, channel, file path). Public playlists per genre streamlined batch downloads of audio and video.

(2) **Data Organization:** The dataset was split into training, validation, and test sets (80/10/10) at the track level to prevent data leakage. Each track was segmented into non-overlapping 30-second excerpts (e.g., 0–30s, 30–60s, ..., 150–180s) to analyze genre characteristics over time. A final 30-second excerpt was also extracted to examine features near the end of songs.

(3) **Audio Segmentation:** Each 30-second excerpt was segmented into overlapping windows of 3, 5, and 10 seconds (50% overlap), capturing rhythmic, harmonic, and timbral patterns at multiple temporal resolutions. This yields 19, 11, and 5 segments per track, respectively.

(4) **Final Dataset:** All versions preserve the original train/validation/test split to ensure fair comparison and reproducibility. Each excerpt is organized by genre and segment duration, and includes both vectorized feature clips and corresponding spectrograms across three temporal positions. Metadata is also provided, including information about the source of each track, such as the original YouTube video used for audio extraction.

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

The final stage involved evaluating BYRM through genre classification experiments. The results support its robustness for capturing temporal and regional characteristics, positioning it as a valuable resource for MIR research.

5. DATASET CHARACTERISTICS

The BYRM dataset comprises 1.082 tracks distributed across 10 Brazilian genres: Axé (101), Carimbó (103), Forró (104), Pagode (102), rock brasileiro (107), Samba (103), Sertanejo (104), Toada (147), Vanera (107), and Xote Gaúcho (104). Track durations vary, with an average of 207.77 seconds. For consistency, only the first 180 seconds of each track were used, allowing the extraction of multiple 30-second excerpts for temporal analysis. Table II presents the number of segments generated per excerpt using 3, 5, and 10 seconds segments with 50% overlap. This approach captures rhythmic, harmonic, and timbral variations while ensuring balanced coverage across temporal positions.

Table II: Number of 3s, 5s, and 10s audio segments generated per temporal block.

Segment Length	**0–30**	**30–60**	**60–90**	**90–120**	**120–150**	**150–180**	**Final 30s**
3s	20.558	20.558	20.511	20.358	19.508	16.715	20.558
5s	11.902	11.902	11.872	11.779	11.264	9.615	11.902
10s	5.410	5.410	5.395	5.349	5.084	4.288	5.410

For feature extraction, Mel-spectrograms were generated for each audio instance, as they are widely used in training deep models like CNNs and Transformers. Spectrograms were computed with parameters balancing temporal and spectral resolution: FFT size 2048, hop length 512, 128 Mel bands, sampling rate 22,050 Hz, and power set to 2.0. These representations effectively capture timbral and rhythmic patterns, especially in short temporal windows crucial for genre recognition.

6. EXPERIMENTAL SETUP

The evaluation was structured to assess how both segment position and duration affect genre classification. First, all 30 second excerpts from the BYRM dataset were divided into overlapping 3 second segment-based units, and the Vision Transformer (ViT), used as the principal model, was trained separately on each temporal block to identify the best and worst performing excerpts. Then, the ViT was retrained on these selected excerpts using 3 second, 5 second, and 10 second segment-based units to evaluate the impact of segment duration and determine whether longer segments offer greater discriminative power. Finally, SVM and ResNet50 models were trained on 3 second segment-based units to enable comparative analysis under identical temporal conditions. Table III summarizes the training strategies and architectural choices adopted for each model.

Table III: 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	Acoustic features
Milestone Epoch	10/30/40	15/100/150	✗
Dropout	30% head, 7% internal	Default	✗
Epochs	100	200	✗
Patience	10	15	✗

To ensure reproducibility, we fixed the random seed (42) in all deep learning experiments. Models were implemented in Python using PyTorch (2.5.1), scikit-learn (1.5.1), and Librosa (0.10.2). ViT and ResNet50 were trained on an NVIDIA RTX 4090 GPU, while the SVM ran on an Intel Core

i9-14900HX CPU. Both ViT and ResNet50 were initialized with ImageNet-pretrained weights and trained using Mel-spectrogram images as input. The ViT model was applied across all segment durations (3s, 5s, and 10s) and temporal excerpts, while ResNet50 and SVM were evaluated only on the best and worst performing excerpts using 3-second segments. Training followed a progressive unfreezing strategy, with specific milestone epochs, and included regularization techniques such as dropout (30% head, 7% internal for ViT) and learning rate decay ($0.8 \rightarrow 0.6 \rightarrow 0.3$). Both deep learning models were trained using Adam optimizer (lr = 1e-4) with early stopping (patience of 10 for ViT and 15 for ResNet50).

The SVM model operated on a 74-dimensional acoustic feature vector extracted with Librosa, including 20 MFCC means and standard deviations, 12 chroma means and standard deviations, and scalar descriptors for spectral centroid, bandwidth, rolloff, and zero-crossing rate. Hyperparameters were optimized via grid search with cross-validation. SVM was included as a classical baseline due to its widespread use in music genre classification studies.

7. RESULTS

This section presents the results obtained from experiments conducted with the BYRM dataset. To investigate how the temporal position of the audio excerpt influences classification performance, the ViT model was trained separately on each 30-second temporal block using 3 second segment-based units with 50% overlap. Performance was evaluated using four standard metrics: accuracy, precision, recall, and F1-score. As shown in Table IV, the 90-120 second excerpt achieved the highest performance, while the initial excerpt of 0-30 seconds resulted in the lowest scores.

Table IV: Classification performance of ViT across different 30-second excerpts (3s segments).

Excerpt	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
0–30s	**74.76**	**75.29**	**74.76**	**74.91 ✗**
30–60s	77.24	78.67	77.24	77.49
60–90s	78.84	79.40	78.84	78.98
90–120s	**79.60**	**80.24**	**79.60**	**79.51 ✓**
120–150s	79.49	80.61	79.49	79.38
150–180s	74.41	75.84	74.41	74.22
Final 30s	76.16	76.94	76.16	75.96

Although the 150-180 second excerpt had a lower F1 score, the 0-30 second segment was chosen as the worst performing excerpt due to the consistency of the dataset. Since not all tracks reach 180 seconds, the final segment suffers from class imbalance. F1-score was adopted as the primary evaluation metric due to the multiclass nature of the task. Unlike accuracy, which may be affected by class imbalance, F1-score harmonizes precision and recall, providing a more balanced measure of model performance across all classes. Thus, this metric was used to identify the best and worst performing excerpts for subsequent comparative experiments. However, since aggregate metrics can mask the individual behavior of specific genres, Figure 2 presents the F1 score for each genre in each 30-second excerpts.

These experimental results support the first hypothesis, which proposed that the position of the excerpt within the song influences the classification performance. The excerpt from 90-120 second yielded the highest overall accuracy, while the initial 0-30 second showed the worst performance. This suggests that central portions of a song tend to carry more genre-defining information, aligning with insights from music cognition. However, the analysis also revealed that some genres performed better in other segments. For example, Samba and Axé achieved their highest F1-scores in the 30-60 second excerpt, and Vaneira peaked in the 60-90 second. These variations indicate that each genre presents its most distinctive features at different points in the song. To further explore the temporal characteristics of genre representation, it was investigated whether segment duration also affects classification performance. This experiment investigates how segment duration impacts classification performance

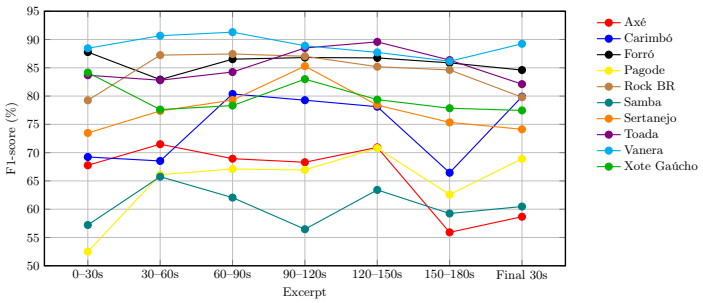

Fig. 2: F1-score for each genre across different 30-second segments using the ViT model.

across different parts of the song. Table V summarizes the ViT model performance for three segment durations applied to the best and worst performing excerpts.

Table V: 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

As shown in Table V, the best performance is achieved with 10 second segments in the excerpt from 90 to 120 seconds, reaching 81.84% F1 score. In contrast, in the excerpt from 0 to 30 seconds, shorter segments of 3 seconds yield slightly better results than longer ones, although the overall performance remains considerably lower than in the later excerpt. To investigate the impact of segment duration, Figure 3 presents the F1-score per genre for different segment lengths in the best-performing excerpt (90–120s).

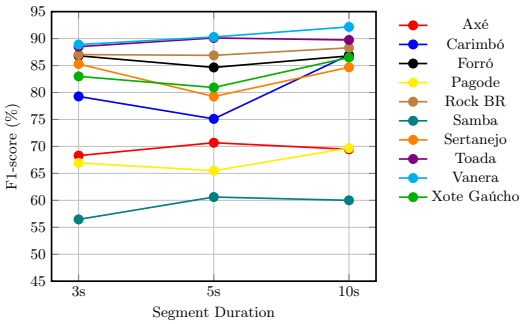

Fig. 3: F1-score per genre on the 90–120s excerpt using different segment durations (ViT).

Segment duration also played a significant role, confirming the second hypothesis. Although longer segments generally led to better results, some genres were more accurately classified with shorter windows. These findings indicate that each genre responds differently to variations in segment duration, and that there is no single segment length that is optimal for all genres. This suggests that the temporal characteristics of each genre influence how effectively it can be classified, reinforcing the need for flexible segmentation strategies that consider musical structure and genre-specific traits. Finally, to compare the effectiveness of different architectures, we evaluated ViT, SVM, and ResNet50 on the best and worst performing excerpts identified in the previous analysis. Table VI summarizes the performance of each model across four evaluation metrics.

Table VI: Performance comparison of ViT, SVM, and ResNet50 on the best and worst excerpts.

Metric	90–120s			0–30s		
	ViT	**SVM**	**ResNet50**	**ViT**	**SVM**	**ResNet50**
Accuracy (%)	**79.60**	70.06	56.41	**74.76**	63.02	51.37
Precision (%)	**80.24**	69.41	56.49	**75.29**	63.09	50.98
Recall (%)	**79.60**	70.06	56.41	**74.76**	63.02	51.37
F1-score (%)	**79.51**	69.45	56.09	**74.91**	62.58	50.94

Among the evaluated models, ViT consistently outperformed both SVM and ResNet50 across all metrics and excerpts. Its ability to capture spectral-temporal patterns in Mel-spectrograms likely contributed to this superior performance. Despite being a classical machine learning method, the SVM achieved competitive results, particularly in the best-performing excerpt, possibly due to certain genres with highly distinguishable acoustic patterns that boosted its overall performance. ResNet50, in contrast, demonstrated the lowest scores in both scenarios. This result may be related to training dynamics that require further adjustment, such as fine-tuning the learning rate or modifying the regularization strategy. Future experiments could explore these parameters to improve performance.

While external comparisons are not applicable due to the uniqueness of the BYRM dataset, the internal results clearly demonstrate the superior performance of the ViT model. Its self-attention mechanism enables it to capture long-range spectral-temporal dependencies more effectively than convolution-based or handcrafted approaches, making it particularly well-suited for regional music classification tasks.

8. CONCLUDING REMARKS

The BYRM dataset comprises 10 Brazilian regional genres and offers multiple excerpts from different parts of each song, allowing controlled investigations into the temporal aspects of genre classification. Despite focusing on a limited number of genres and relying on a curated YouTube selection, the dataset brings valuable contributions to the field. Its design enables a detailed evaluation of how excerpt position and segment duration affect classification, a factor often overlooked in existing datasets that rely on fixed segments. Initial experiments using a Vision Transformer (ViT), along with SVM and ResNet50 baselines, demonstrated consistent and interpretable results. Central excerpts, particularly between 90 and 120 seconds, led to higher classification accuracy, while segment duration influenced performance differently across genres. These results reinforce the importance of temporal structure in music classification. Future work includes expanding the dataset with more tracks per genre and incorporating additional genres. Further directions involve measuring genre similarity in the embedding space and testing prediction aggregation across entire songs. Overall, BYRM stands out as a regional dataset that reflects the temporal and stylistic richness of Brazilian music, fostering research on genre recognition and temporal dynamics in musical audio.

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.

REFERENCES

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

Barbosa, J., McKay, C., and Fujinaga, I. Evaluating automated classification techniques for folk music genres from the brazilian northeast. *Computer Music: Beyond the frontiers of signal processing and computational models*, 2015.

Byron, T. P., Rushworth, C. T., and Stewart, M. J. 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.

Cerati, G. Difficult to define, easy to understand: the use of genre categories while talking about music. *SN Social Sciences* 1 (12): 288, 2021.

Cheng, Y.-H., Chang, P.-C., and Kuo, C.-N. Convolutional neural networks approach for music genre classification. In *2020 International Symposium on Computer, Consumer and Control (IS3C)*. IEEE, pp. 399–403, 2020.

da Conceição, J. L., de Freitas, R., Gadelha, B., Kienen, J. G., Anders, S., and Cavalcante, B. Reconhecendo gêneros musicais brasileiros com técnicas de aprendizagem de máquina supervisionada. In *Anais do XLVII Seminário Integrado de Software e Hardware*. SBC, pp. 186–197, 2020.

Dai, J., Liang, S., Xue, W., Ni, C., and Liu, W. Long short-term memory recurrent neural network based segment features for music genre classification. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, pp. 1–5, 2016.

De Sousa, J. M., Pereira, E. T., and Veloso, L. R. A robust music genre classification approach for global and regional music datasets evaluation. In *2016 IEEE international conference on digital signal processing (DSP)*. IEEE, pp. 109–113, 2016.

Defferrard, M., Benzi, K., Vandergheynst, P., and Bresson, X. Fma: A dataset for music analysis. *arXiv preprint arXiv:1612.01840*, 2016.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Fabbri, F. et al. A theory of musical genres. two applications. In *Popular music perspectives*. Vol. 1. Iaspm, pp. 52–81, 1982.

Fayyad, U. Knowledge discovery in databases: An overview. In *International Conference on Inductive Logic Programming*. Springer, pp. 1–16, 1997.

Guimarães, V. d. S. and Freitas, R. d. Byrm: Brazilian youtube regional music dataset, 2025.

ISMIR. Ismir 2004 genre classification contest. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*. Barcelona, Spain, 2004.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86 (11): 2278–2324, 1998.

Medhat, F., Chesmore, D., and Robinson, J. Masked conditional neural networks for sound classification. *Applied Soft Computing* vol. 90, pp. 106073, 2020.

Meng, Y. Music genre classification: A comparative analysis of cnn and xgboost approaches with mel-frequency cepstral coefficients and mel spectrograms. *arXiv preprint arXiv:2401.04737*, 2024.

Silla Jr, C. N., Kaestner, C. A., and Koerich, A. L. Automatic music genre classification using ensemble of classifiers. In *2007 IEEE International Conference on Systems, Man and Cybernetics*. IEEE, pp. 1687–1692, 2007.

Silva, D., Zampar, L., Rodrigues, F., and Gomes, C. Modelo automático de classificação de gêneros musicais amazônicos. In *Anais do XVIII Simpósio Brasileiro de Computação Musical*. SBC, pp. 225–228, 2021.

Sturm, B. L. The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use. *arXiv preprint arXiv:1306.1461*, 2013.

Turab, M., Kumar, T., Bendechache, M., and Saber, T. Investigating multi-feature selection and ensembling for audio classification. *arXiv preprint arXiv:2206.07511*, 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* vol. 30, 2017.

Wijaya, N. N., Muslikh, A. R., et al. Music-genre classification using bidirectional long short-term memory and mel-frequency cepstral coefficients. *Journal of Computing Theories and Applications* 1 (3): 243–256, 2024.

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

Zhang, X. and Ras, Z. W. Analysis of sound features for music timbre recognition. In *2007 International Conference on Multimedia and Ubiquitous Engineering (MUE'07)*. IEEE, pp. 3–8, 2007.

Zhao, H., Zhang, C., Zhu, B., Ma, Z., and Zhang, K. S3t: Self-supervised pre-training with swin transformer for music classification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 606–610, 2022.

Zhuang, Y., Chen, Y., and Zheng, J. Music genre classification with transformer classifier. In *Proceedings of the 2020 4th international conference on digital signal processing*. pp. 155–159, 2020.