



**UNIVERSIDADE FEDERAL DO AMAZONAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA**

**RECONHECENDO GÊNEROS MUSICAIS BRASILEIROS SIMILARES POR
ABORDAGEM BASEADA EM SEGMENTOS E APRENDIZADO PROFUNDO
USANDO TRANSFORMADORES VISUAIS**

Victória de Souza Leon Guimarães

Orientador(a): Rosiane de Freitas Rodrigues, Ph.D.

Co-orientador(a): João Gustavo Kienen, Dr.

Setembro de 2025

Manaus - AM

Victória de Souza Leon Guimarães

RECONHECENDO GÊNEROS MUSICAIS BRASILEIROS SIMILARES POR
ABORDAGEM BASEADA EM SEGMENTOS E APRENDIZADO PROFUNDO
USANDO TRANSFORMADORES VISUAIS

Dissertação de Mestrado apresentado ao Programa
de Pós-Graduação em Informática do Instituto de
Computação Universidade Federal do Amazonas,
como parte dos requisitos necessários à obtenção
do título de Mestre em Informática.

Orientador(a): Rosiane de Freitas Rodrigues, Ph.D.

Co-orientador(a): João Gustavo Kienen, Dr.

Setembro de 2025

Manaus - AM

Ficha Catalográfica

Elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

G963r Guimarães, Victória de Souza Leon
Reconhecendo gêneros musicais brasileiros similares por abordagem baseada em segmentos e aprendizado profundo usando transformadores visuais / Victória de Souza Leon Guimarães. - 2025.
98 f. : il., color. ; 31 cm.

Orientador(a): Rosiane de Freitas Rodrigues.
Coorientador(a): João Gustavo Kienen.
Dissertação (mestrado) - Universidade Federal do Amazonas, Programa de Pós-Graduação em Informática, Manaus, 2025.

1. Recuperação de informação musical. 2. Aprendizado profundo. 3. Redes de transformadores. 4. Similaridade entre gêneros musicais. 5. Música brasileira. I. Rodrigues, Rosiane de Freitas. II. Kienen, João Gustavo. III. Universidade Federal do Amazonas. Programa de Pós-Graduação em Informática. IV. Título



Ministério da Educação
Universidade Federal do Amazonas
Coordenação do Programa de Pós-Graduação em Informática

FOLHA DE APROVAÇÃO

"RECONHECENDO GÊNEROS MUSICAIS BRASILEIROS SIMILARES POR ABORDAGEM BASEADA EM SEGMENTOS E APRENDIZADO PROFUNDO USANDO TRANSFORMADORES VISUAIS"

VICTÓRIA DE SOUZA GUIMARÃES

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Profa. Dra. Rosiane de Freitas-Rodrigues - **Presidente**

Profa. Dra. Eulanda Miranda dos Santos - **Membro Interno**

Prof. Dr. Waldir Sabino da Silva Jr - **Membro Externo**

Prof. Dr. Diego Furtado da Silva - **Membro Externo**

Manaus, 10 de setembro de 2025.



Documento assinado eletronicamente por **Waldir Sabino da Silva Júnior**, Professor do Magistério Superior, em 17/09/2025, às 08:58, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Rosiane de Freitas Rodrigues**, Professor do Magistério Superior, em 17/09/2025, às 11:28, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Eulanda Miranda dos Santos, Professor do Magistério Superior**, em 17/09/2025, às 17:37, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Diego Furtado Silva, Usuário Externo**, em 02/12/2025, às 09:01, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufam.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2788878** e o código CRC **C7AC648B**.

Avenida General Rodrigo Octávio, 6200 - Bairro Coroado I Campus Universitário Senador Arthur Virgílio Filho, Setor Norte - Telefone: (92) 3305-1181 / Ramal 1193
CEP 69080-900, Manaus/AM, coordenadorppgi@icomp.ufam.edu.br

Referência: Processo nº 23105.040331/2025-76

SEI nº 2788878

Dedico este trabalho para minha grande amiga Kimberli Lopes (*in memoriam*),
que, assim como este trabalho, foi música por onde passou.

Agradecimentos

Em primeiro lugar, agradeço à minha mãe, Danielle de Souza, por ter investido na minha educação e por nunca duvidar da minha capacidade. Obrigada por acreditar em mim, e por sempre abrir uma janela quando uma porta se fechava.

Agradeço ao meu esposo, Anthony Leon, que foi um grande companheiro e amigo nessa jornada, ouvindo minhas queixas, teorias e sendo sempre minha fortaleza. Obrigada por me lembrar constantemente da minha capacidade e por toda a sua compreensão.

Aos meus filhos de quatro patas, Chewie e Aurora, que me acompanharam dia e noite enquanto eu desenvolvía esta dissertação, oferecendo amor e apoio silencioso.

Agradeço também a mim mesma, por não desistir diante dos obstáculos, por seguir em frente e por toda a minha perseverança.

À minha família e aos amigos fiéis, que entenderam os momentos em que precisei me ausentar para estudar, por ouvirem meus desabafos e por serem minha válvula de escape. Obrigada pela amizade.

À minha orientadora, professora Rosiane de Freitas, por todo o aprendizado e pela parceria ao longo desta pesquisa.

A todo o grupo de pesquisa AlgoX, em especial aos meus queridos amigos Larissa Pessoa, Pedro Matias e Elton Alencar, que levo com carinho desde a graduação e que foram grandes companheiros de estudos e conversas.

À Universidade Federal do Amazonas, instituição essencial na minha formação profissional, por me oferecer uma estrutura de qualidade para realizar meus estudos, incentivo à pesquisa e um dos melhores corpos docentes do Estado do Amazonas.

“Seu melhor e mais sábio refúgio contra todos os problemas está na sua ciência.”

Ada Lovelace

RECONHECENDO GÊNEROS MUSICAIS BRASILEIROS SIMILARES POR
ABORDAGEM BASEADA EM SEGMENTOS E APRENDIZADO PROFUNDO
USANDO TRANSFORMADORES VISUAIS

Victória de Souza Leon Guimarães

Setembro/2025

Orientador(a): Rosiane de Freitas Rodrigues

João Gustavo Kienen

Resumo:

Esta pesquisa investiga o uso de mecanismos de autoatenção em arquiteturas de transformadores visuais, com foco no modelo Vision Transformer (ViT), para o reconhecimento e análise de similaridade entre gêneros musicais. O estudo envolve a construção do BYRM Dataset, uma base curada de músicas regionais brasileiras composta por 1.082 faixas distribuídas em dez gêneros culturalmente diversos, extraídas automaticamente a partir de álbuns no YouTube. A base permite experimentos controlados com múltiplos trechos por faixa, variando tanto a posição temporal quanto a duração dos segmentos. O modelo ViT foi treinado com espectrogramas de Mel como entrada e avaliado por meio de métricas como acurácia, precisão, revocação e f1-score. A melhor configuração alcançou 81,94% de acurácia e 81,84% de f1-score. Além da classificação, esta dissertação também propõe uma análise da representação vetorial aprendida pelo modelo, aplicando técnicas de redução de dimensionalidade (PCA, t-SNE e UMAP) e medidas quantitativas de similaridade, como a distância cosseno, para investigar a proximidade entre gêneros no espaço latente. Os resultados indicam que o ViT é capaz de capturar relações estilísticas relevantes, agrupando gêneros semelhantes como samba e pagode, ou vaneira e xote gaúcho, ao mesmo tempo em que separa estilos mais distintos, como o rock brasileiro. Este trabalho contribui para a área de Recuperação de Informação Musical (MIR), demonstrando o potencial de modelos baseados em atenção na classificação e interpretação de gêneros musicais em contextos ricos e diversos culturalmente.

Palavras-chave: recuperação de informação musical, aprendizado profundo, redes de transformadores, espectrograma, análise temporal, similaridade entre gêneros musicais, redução de dimensionalidade, música brasileira.

RECONHECENDO GÊNEROS MUSICAIS BRASILEIROS SIMILARES POR
ABORDAGEM BASEADA EM SEGMENTOS E APRENDIZADO PROFUNDO
USANDO TRANSFORMADORES VISUAIS

Victória de Souza Leon Guimarães

September/2025

Advisor: Rosiane de Freitas Rodrigues

João Gustavo Kienen

Abstract:

This research investigates the use of self-attention mechanisms in visual transformer architectures, focusing on the Vision Transformer (ViT) model for musical genre recognition and similarity analysis. The study involves the construction of the BYRM Dataset, a curated collection of Brazilian regional music comprising 1.082 tracks across ten culturally diverse genres, automatically extracted from YouTube album videos. The dataset supports controlled experiments using multiple excerpts per track, varying both the temporal position and the segment duration. The ViT model was trained using Mel-spectrograms as input and evaluated with metrics such as accuracy, precision, recall, and F1-score. The best configuration achieved 81,94% accuracy and 81,84% F1-score. Beyond classification, this dissertation also proposes an analysis of the vector representations learned by the model, applying dimensionality reduction techniques (PCA, t-SNE, and UMAP) and quantitative similarity measures, such as cosine distance, to investigate the proximity between genres in the latent space. Results indicate that ViT successfully captures meaningful stylistic relationships, grouping similar genres such as samba and pagode or vaneira and xote gaúcho. This work contributes to the field of Music Information Retrieval (MIR), demonstrating the potential of attention-based models in the classification and interpretation of musical genres in culturally rich and diverse contexts.

Keywords: music information retrieval, deep learning, transformer networks, spectrogram, temporal analysis, genre similarity, dimensionality reduction, Brazilian music.

Sumário

Lista de Figuras	1
Lista de Tabelas	4
1 Introdução	7
1.1 Motivação	9
1.2 Objetivos	10
1.3 Organização da Dissertação	11
2 Fundamentação Teórica	12
2.1 Recuperação de Informação e Gêneros Musicais	12
2.2 Conceitos Fundamentais de Áudio Digital	14
2.2.1 Representação Discreta de Sinais Musicais	14
2.2.2 Propriedades Acústicas e Perceptuais do Áudio	15
2.2.3 Características Musicais: Harmonia, Melodia e Ritmo	15
2.2.4 Características Clássicas de Áudio	16
2.2.5 Visualizações no Domínio Tempo–Frequência	17
2.3 Redução de Dimensionalidade	18
2.4 Similaridade	19
2.4.1 Similaridade Musical	19
2.4.2 Similaridade Vetorial	20

2.5	Aprendizado Profundo	22
2.5.1	Técnicas de Aprendizado Profundo	22
2.5.2	Mecanismos de Atenção e Transformadores	25
2.5.3	Aprendizado por Transferência	26
2.5.4	Métricas de Avaliação	27
3	Reconhecimento de Gêneros Musicais	29
3.1	Definição Geral do Problema de Reconhecimento de Gêneros Musicais	29
3.2	Trabalhos Relacionados	30
3.2.1	Trabalhos em MGR	31
3.2.2	Bases de Dados	34
3.2.3	Trabalhos com Estratégias de Segmentação Temporal e Redução	
	de Dimensionalidade	37
3.2.4	MGR Utilizando Arquitetura de Transformadores	40
4	Metodologia	42
4.1	Visão Geral	42
4.2	Experimento Inicial	43
4.3	Base de Dados BYRM	44
4.4	Preparação dos Dados	47
4.5	Divisão dos Dados	49
4.6	Configuração dos Experimentos	51
4.7	Modelos de Classificação Utilizados	52
4.7.1	Procedimento de Treinamento SVM	53
4.7.2	Procedimento de Treinamento ResNet50	54
4.7.3	Procedimento de Treinamento ViT	54
4.8	Arquitetura do Vision Transformer Aplicada a Espectrogramas	56

4.9	Análise dos Embeddings e Similaridade entre Gêneros Musicais	58
5	Resultados	60
5.1	Revisão dos Experimentos da Qualificação	60
5.2	Análise da Influência Temporal na Classificação de Gêneros Musicais . .	62
5.2.1	Comparação da Classificação entre Trechos: Melhor e Pior Posição	64
5.3	Comparação de Desempenho entre Diferentes Modelos de Classificação .	67
5.4	Avaliação Baseada na Duração dos Segmentos	72
5.5	Visualização e Análise dos Embeddings no Espaço Vetorial	74
5.6	Análise de Similaridade entre Gêneros Musicais	80
5.7	Discussão dos Resultados	84
5.7.1	Temporalidade	84
5.7.2	Espaço Vetorial e Similaridade	86
6	Considerações Finais	88

Lista de Figuras

2.1	Forma de onda do sinal de áudio (acima) e seu respectivo mel-espectrograma (abaixo), evidenciando a distribuição de energia em bandas perceptuais ao longo do tempo.	17
2.2	Exemplos em planos cartesianos distintos ilustrando os valores da similaridade cosseno: (a) máxima similaridade, (b) nenhuma similaridade, (c) máxima dissimilaridade.	21
2.3	Evolução das arquiteturas de redes neurais profundas com destaque para o surgimento dos transformadores e sua aplicação em visão computacional.	24
2.4	Relação <i>encoder</i> e <i>decoder</i> presentes em uma arquitetura de redes de transformadores.	26
4.1	Gêneros musicais regionais brasileiros que integram a base de dados BYRM por região característica.	44
4.2	Processo de criação da base de dados BYRM.	45
4.3	Representação da segmentação e sobreposição de clipes de áudio com diferentes durações.	48
4.4	Exemplos visuais dos gêneros musicais presentes na base BYRM.	50
4.5	Arquitetura ViT.	56
5.1	Funções de perda e acurácia ao longo do treinamento com a base GTZAN e BYRM.	61
5.2	Matriz de confusão após o ajuste fino.	62

5.3	Acurácia por gênero em cada trecho de 30 segundos com o modelo ViT.	64
5.4	F1-score por gênero em cada trecho de 30 segundos com o modelo ViT.	64
5.5	Funções de perda e acurácia ao longo do treinamento para o melhor (90–120 segundos) e pior trecho (0–30 segundos).	65
5.6	Matrizes de confusão do melhor e pior trecho classificado com o modelo ViT.	66
5.7	Métricas de desempenho por classe para o classificador SVM no trecho de 90–120 segundos.	68
5.8	Matriz de confusão obtida com o SVM aplicado ao trecho de 90–120 segundos da base BYRM.	69
5.9	Funções de perda e acurácia ao longo do treinamento para o melhor trecho (90–120 segundos)	70
5.10	Matriz de confusão obtida com a ResNet50 aplicada ao trecho de 90–120 segundos da base BYRM.	71
5.11	Métricas de desempenho por classe para o classificador ResNet50 no segmento de 90–120 segundos.	71
5.12	F1-score por gênero no trecho de 90–120s com diferentes durações de segmento (ViT).	73
5.13	Projeções dos embeddings do trecho 90–120s com segmentos de 3s, antes e após o treinamento do ViT. Cada subfigura apresenta as representações via PCA, t-SNE e UMAP.	75
5.14	Projeções dos embeddings do trecho 90–120s com segmentos de 5s, antes e após o treinamento do ViT. Cada subfigura apresenta as representações via PCA, t-SNE e UMAP.	76
5.15	Projeções dos embeddings do trecho 90–120s com segmentos de 10s, antes e após o treinamento do ViT. Cada subfigura apresenta as representações via PCA, t-SNE e UMAP.	77

5.16	Projeções dos embeddings do pior trecho 0–30s com segmentos de 3s, antes e após o treinamento do ViT. Cada subfigura apresenta as represen- tações via PCA, t-SNE e UMAP.	78
5.17	Projeções dos embeddings do pior trecho 0–30s com segmentos de 5s, antes e após o treinamento do ViT. Cada subfigura apresenta as represen- tações via PCA, t-SNE e UMAP.	79
5.18	Projeções dos embeddings do pior trecho 0–30s com segmentos de 10s, antes e após o treinamento do ViT. Cada subfigura apresenta as represen- tações via PCA, t-SNE e UMAP.	79
5.19	Mapas de calor da similaridade de cosseno entre os gêneros musicais com embeddings do trecho 90–120s, considerando diferentes durações de seg- mentos.	81

Lista de Tabelas

2.1	Definições e Fórmulas das Principais Métricas de Avaliação.	27
3.1	Tabela comparativa dos principais trabalhos relacionados da literatura para a tarefa de classificação de gêneros musicais.	31
3.2	Tabela comparativa das principais bases de dados utilizadas em problemas de reconhecimento de gênero musical.	35
3.3	Comparação entre os trabalhos da literatura com foco em estratégias de segmentação, duração dos trechos analisados e posicionamento temporal.	38
3.4	Comparação entre os trabalhos relacionados quanto ao uso de similaridade e técnica adotada.	39
4.1	Principais artistas representativos por gênero musical no BYRM.	46
4.2	Quantidade de segmentos gerados por duração e por bloco temporal.	47
4.3	Divisão dos dados por bloco temporal e duração dos segmentos.	51
4.4	Resumo das configurações dos modelos utilizadas nos experimentos.	52
5.1	Métricas gerais obtidas na qualificação após ajuste fino.	61
5.2	Desempenho do ViT nos diferentes blocos de 30 segundos (segmentos de 3s).	63
5.3	Comparação de desempenho entre ViT, SVM e ResNet50 nos melhores (90–120s) e piores (0–30s) trechos da base.	68

5.4	Desempenho do ViT com diferentes durações de segmento nos trechos de	
	melhor (90–120s) e pior (0–30s) desempenho.	72
5.5	Melhor e pior trecho por gênero, com F1-score máximo e mínimo.	84

Lista de Acrônimos e Abreviaturas

MIR	Music Information Retrieval
CNN	Convolutional Neural Network
ViT	Vision Transformer
RNN	Recurrent Neural Network
NLP	Neural Language Processing
MGR	Music Genre Recognition
MFCC	Mel-Frequency Cepstral Coefficients
ANN	Artificial Neural Network
ViT	Vision Transformer
ResNet	Residual Network
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
XGB	Extreme Gradient Boosting
MLP	Multilayer Perceptron
MCNN	Masked Convolutional Neural Network
LSTM	Long Short-Term Memory
GTZAN	GTZAN Genre Collection Dataset
ISMIR	International Society for Music Information Retrieval
FMA	Free Music Archive
STFT	Short-Time Fourier Transform
BYRM	Brazilian Youtube Regional Music

Capítulo 1

Introdução

A área de recuperação de informação musical, ou MIR (do inglês, *Music Information Retrieval*), abrange vários campos de estudos, onde inclui atividades como a classificação de notas musicais, de emoções, de artistas, e de gêneros musicais [Aucouturier and Pampalk 2008]. A classificação ou reconhecimento de gêneros musicais (do inglês, *Music Genre Recognition*) é uma área que vem sendo estudada ao longo dos anos, e é uma tarefa significativa no campo da pesquisa MIR.

Para realizar o reconhecimento automático do gênero musical, existem pesquisas que utilizam algoritmos de aprendizagem de máquina supervisionada para resolver este problema [Silla Jr et al. 2007], e recentemente, propostas mais complexas foram apresentadas, utilizando redes neurais convolucionais ou CNNs (do inglês, *Convolutional Neural Network*) [Meng 2024] e redes neurais recorrentes ou RNNs (do inglês, *Recurrent Neural Network*) [Dai et al. 2016]. O reconhecimento de gêneros musicais é uma atividade desafiadora devido às similaridades e nuances que caracterizam cada gênero [Sturm 2014], visto que características sutis em músicas são difíceis de serem capturadas por estratégias computacionais, especialmente quando a análise se baseia apenas em fragmentos das músicas.

Além disso, muitos gêneros compartilham estruturas musicais semelhantes, o que torna ainda mais desafiador o reconhecimento automático de estilos com alto grau de similaridade, como é o caso dos gêneros regionais. Nesses contextos, há uma combinação de constantes variações ao longo da música, como mudanças rítmicas e instrumentais, e uma forte dependência de elementos culturais para a correta identificação do gênero [Correa

[2018]. Esse desafio se intensifica diante de influências culturais sobrepostas, estruturas rítmicas compartilhadas e instrumentações semelhantes, que elevam os níveis de similaridade entre estilos musicais [de Araújo Lima et al. 2020, Muszynski and Tarapata 2023]. Além disso, a presença de faixas híbridas, misturas de estilos e a fluidez estética típica da música popular tornam a tarefa de classificação ainda mais complexa, exigindo abordagens computacionais mais sensíveis às nuances contextuais [Cassano et al. 2024, Correa 2018, Castro 2010].

As redes convolucionais são úteis para extrair informações relevantes de dados de áudio, e são eficientes para detectar padrões locais e invariantes devido ao compartilhamento de parâmetros e à natureza das operações convolucionais. No entanto, elas não conseguem integrar bem o contexto das informações sequenciais de forma global. As RNNs podem manter informações sobre o contexto de dados anteriores e usá-los para prever o próximo item na sequência [Goodfellow et al. 2016]. Ainda assim, a memória de curto prazo empregada nessas arquiteturas ainda é uma limitação, visto que os modelos recorrentes processam a entrada sequencialmente, o que significa que para calcular o estado atual, é necessário ter a saída anterior, o que impossibilita paralelizar o processo [Colliot 2023].

De outra forma, redes de transformadores (do inglês, *transformer*) [Vaswani et al. 2017] utilizam mecanismo de autoatenção, onde simulam a técnica que o cérebro humano utiliza para se concentrar nas partes relevantes de uma atividade, para entender o contexto global. Esse mecanismo permite a compreensão do contexto a longo prazo, onde cada entrada é processada de forma independente e possui em sua arquitetura a possibilidade de realizar múltiplos cálculos de atenção de uma só vez, o que torna os transformadores paralelizáveis. A capacidade de capturar dependências de longo alcance nas sequências torna essa abordagem eficaz em diversas aplicações, incluindo tradução automática, processamento de linguagem natural, e também a classificação de áudio, como o reconhecimento de gênero musical. Uma adaptação relevante é o Vision Transformer (ViT) [Dosovitskiy et al. 2020], que trata espectrogramas de áudios como imagens e divide o sinal em *patches* processados como *tokens* visuais. Essa estratégia permite organizar a informação musical de forma hierárquica e aprender representações latentes que refletem padrões culturais e estilísticos.

Diferentemente das abordagens tradicionais que se concentram apenas na acurá-

cia, este trabalho propõe uma análise de como os gêneros musicais regionais se organizam no espaço latente aprendido por um modelo Vision Transformer ajustado. Com base em estudos de cognição musical que apontam refrões e ganchos melódicos como segmentos representativos e salientes [Byron et al. 2025], este estudo analisa os *embeddings* extraídos de trechos estrategicamente selecionados. Utilizando métricas de similaridade vetoriais e técnicas de redução de dimensionalidade, examina-se se a disposição espacial dos gêneros nesse espaço vetorial reflete afinidades musicais reconhecidas. Desta forma, esta dissertação busca aplicar arquiteturas de redes profundas de transformadores para explorar seu comportamento na atividade de reconhecimento e caracterização de gêneros musicais regionais.

1.1. Motivação

As redes de transformadores foram introduzidas para resolver problemas que necessitam da interpretação de contexto, especialmente em tarefas de processamento de linguagem natural ou NLP (do inglês, *Natural Language Processing*) [Colliot 2023]. Entretanto, existem outros problemas em que a interpretação do contexto global é igualmente importante, semelhante à análise humana, é necessário analisar os *frames* para entender um vídeo completo, assim como a interpretação de notas musicais para compreender uma melodia, as palavras para entender um texto, e como neste trabalho, entender pequenas partes de um áudio de uma música para compreender o gênero musical.

A essência do uso das redes de transformadores na classificação de gêneros musicais se encontra na sua capacidade de capturar padrões complicados, dependências e relações temporais nos dados de música devido seu mecanismo de autoatenção. Ao contrário de outros métodos, que muitas vezes enfrentam dificuldades para representar as estruturas complexas da música, os transformadores têm um desempenho superior na modelagem de sequências de dados [Vaswani et al. 2017]. Dessa forma, explorar essa arquitetura é relevante por sua capacidade de superar limitações dos modelos tradicionais, oferecendo uma abordagem mais eficaz e precisa para o reconhecimento e a categorização de gêneros musicais. A proposta tem o potencial de melhorar significativamente a precisão e a eficiência na classificação automática de gêneros musicais, considerando as nuances e variações entre estilos com alto grau de similaridade.

Estudos em cognição musical indicam que determinados segmentos da música,

como refrões, possuem maior relevância perceptiva e contribuem fortemente para a definição de estilo [Byron et al. 2025]. Apesar disso, muitos trabalhos na literatura utilizam apenas trechos fixos de 30 segundos por faixa, frequentemente ignorando a posição e a duração desses segmentos [Sturm 2013, ISMIR 2004, Defferrard et al. 2016]. Essa limitação impede uma análise mais rica sobre o papel que diferentes partes da música exercem na percepção e classificação do gênero. Dessa forma, com base na análise do contexto técnico-científico atual, propõe-se a seguinte questão de pesquisa principal:

- (1) É possível contribuir com o aprimoramento do processo de reconhecimento de gêneros musicais com alto grau de similaridade utilizando mecanismos de autoatenção?

E as seguintes questões complementares:

- (2) A posição temporal do trecho analisado dentro da música influencia o reconhecimento de gênero musical?
- (3) A duração dos segmentos de áudio impacta significativamente no reconhecimento de gênero musical?

Para responder à questão de pesquisa deste trabalho, a metodologia científica adotada é de **natureza aplicada**, focando na implementação e treinamento de uma arquitetura de rede de transformadores e na realização de experimentos para identificar e compreender os resultados obtidos, adotando uma abordagem **quali-quantitativa** com uma **análise descritiva** para descrever e interpretar os experimentos realizados. A abordagem metodológica combina métodos **qualitativos**, incluindo a revisão da literatura para entender o cenário atual e examinar a distribuição dos gêneros no espaço de *embeddings*, com base em técnicas de visualização e similaridade vetorial. O método **quantitativo** envolve a criação de uma base de dados, extração de características de áudio e a implementação e avaliação da arquitetura, utilizando métricas como acurácia, precisão, revocação e F1-score para avaliar o desempenho dos modelos.

1.2. Objetivos

Esta seção apresenta o objetivo geral e os objetivos específicos definidos para este trabalho para responder à questão de pesquisa.

Objetivo Geral

Este trabalho tem como objetivo geral propor uma estratégia para realizar o reconhecimento de gêneros musicais utilizando uma arquitetura de rede de transformadores.

Objetivos Específicos

Para atingir o objetivo geral proposto, serão considerados os seguintes objetivos específicos:

- Gerar uma base de dados com gêneros musicais regionais representativos;
- Identificar representações de áudio para modelos de aprendizado profundo conforme estado da arte;
- Explorar arquitetura de transformadores para extração de características em áudios, de modo a avaliar o reconhecimento de gêneros musicais;
- Investigar a organização interna dos gêneros musicais no espaço latente aprendido pelo modelo.

1.3. Organização da Dissertação

O trabalho está organizado da seguinte maneira: o Capítulo **1** apresenta a introdução do trabalho, com a motivação para o estudo, a questão de pesquisa, e os objetivos gerais e específicos da dissertação. Seguido do Capítulo **2** que fornece a base teórica necessária para a compreensão do estudo, abordando temas como Recuperação de Informação Musical e Aprendizado Profundo. O Capítulo **3** destaca o problema da pesquisa, e os trabalhos relacionados da literatura sobre reconhecimento de gêneros musicais e bases de dados. Esse capítulo apresenta a proposta da dissertação. Após isso, o Capítulo **4** detalha os métodos e procedimentos utilizados na pesquisa, incluindo a visão geral da metodologia, a descrição das bases de dados, o pré-processamento dos dados, e a configuração do modelo. Esta seção é crucial para entender como os experimentos foram conduzidos.

O Capítulo **5** apresenta os resultados obtidos nessa dissertação, incluindo uma discussão sobre os resultados e identificando melhorias para o processo. Finalizando com o Capítulo **6** com uma discussão sobre as conclusões da pesquisa, as limitações encontradas e as possíveis direções para trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Este capítulo discorre sobre pontos fundamentais para o entendimento deste trabalho.

2.1. Recuperação de Informação e Gêneros Musicais

De acordo com o autor Aucouturier [\[Aucouturier and Pampalk 2008\]](#), MIR é uma área de pesquisa que se concentra no desenvolvimento de métodos para acessar e organizar informações musicais. Esse campo de pesquisa inclui diversas atividades como a classificação automática de gêneros musicais, análise de áudio, processamento de sinais, reconhecimento de padrões e inteligência artificial. O estudo de MIR impulsionou o desenvolvimento de algoritmos e tecnologias para análise e classificação de música com base no conteúdo de áudio, ao invés de informações editoriais, facilitando a recuperação de informações musicais de forma automatizada, o que vêm sendo estudado até hoje. Uma dessas atividades, foi o reconhecimento de gêneros musicais ou MGR (do inglês, *Music Genre Recognition*), que é considerada uma das principais aplicações desse campo de pesquisa.

A compreensão do gênero musical como uma construção social moldada por fatores culturais é essencial para interpretar representações musicais. Gêneros não são categorias fixas ou claramente delimitadas, ao contrário, emergem por meio de processos de abstração que envolvem elementos simbólicos e contextuais [\[Krogh 2019\]](#). Na literatura, não se encontra uma definição formal do que são gêneros musicais. Entretanto, alguns autores buscam encontrar uma forma de rotular ou ao menos identificar o que é um gênero musical. Desta forma, o gênero musical pode ser definido como um conjunto de eventos

musicais (reais ou possíveis) cujo curso é governado por um conjunto definido de regras socialmente aceitas [Fabbri et al. 1982]. O termo eventos musicais refere-se a qualquer tipo de atividade realizada em torno de eventos que envolvem som, como apresentações ao vivo e gravações. O que destaca também a relevância sociológica para esta definição, onde as regras são moldadas e aceitas pelas comunidades envolvidas, como, por exemplo, por questões culturais.

Além disso, o gênero musical é considerado um dos principais descritores utilizados na indústria fonográfica, sendo amplamente empregado tanto por profissionais do setor quanto pelos próprios consumidores para organizar, buscar e consumir música [Pachet et al. 2000]. Desta forma, é definido que os gêneros musicais são categorias que ajudam a organizar e classificar músicas, facilitando a navegação em catálogos de música. Também é abordado que a rotulação de gêneros musicais considera uma combinação de elementos como distorção, altura e intensidade vocal, harmonia, melodia e ritmo, variando em importância entre os gêneros musicais [Junior 2006].

Desta forma, pode-se concluir que os gêneros musicais são utilizados para comunicar preferências musicais e descrever experiências musicais de maneira eficiente, embora sua definição exata possa variar dependendo do contexto e dos participantes da interação [Cerati 2021]. Estes pontos refletem a dificuldade em encontrar uma definição universal para gêneros musicais e deixa em evidência a complexidade do problema estudado. Este desafio reflete a natureza dos gêneros musicais, que incorporam aspectos técnicos, formais, comportamentais, sociais, ideológicos, econômicos, além de elementos musicais como intensidade, harmonia, melodia e ritmo. A evolução constante da música e a diversidade cultural contribuem para essa dificuldade, tornando o conceito de gênero musical dinâmico e contextual.

A compreensão do gênero musical também se beneficia de uma perspectiva cognitiva e musicológica. Estudos em cognição musical sugerem que certos trechos da música, como refrões e motivos melódicos recorrentes, apresentam maior saliência perceptiva e memorabilidade, sendo mais eficazes para o reconhecimento de gêneros [Byron et al. 2025]. Esses segmentos são mais facilmente lembrados pelos ouvintes e servem como âncoras cognitivas que facilitam a identificação estilística. Trechos vocais retirados da linha melódica principal tendem a ser julgados como mais marcantes do que trechos instrumentais ou de acompanhamento, o que reforça a importância estrutural dos refrões na

percepção e categorização musical.

No campo da teoria musical, especialmente nas canções populares da indústria cultural, observa-se que essas obras geralmente seguem uma estrutura que inclui introdução, estrofes e refrões. A introdução, por sua vez, pode desempenhar funções variadas, como preparar o ouvinte emocionalmente, criar contraste ou antecipar temas melódicos, mas nem sempre apresenta os elementos mais característicos do gênero. Em gêneros como a toada de boi-bumbá, por exemplo, é comum que as introduções sejam líricas e melódicas, enquanto a parte seguinte revela uma identidade mais percussiva e exultante, trazendo características marcantes do estilo. Isso demonstra que a representatividade temporal dentro de uma faixa musical não é uniforme, sendo necessário selecionar com critério os segmentos utilizados em análises computacionais de gênero.

Considerando esses aspectos, entender o gênero musical como um construtor social e perceptivo, fortemente ligado à cultura e à cognição, é fundamental para a interpretação dos dados e resultados em tarefas de classificação automática. Essa abordagem permite que a análise computacional seja mais sensível aos elementos que realmente evocam a identidade dos gêneros musicais.

2.2. Conceitos Fundamentais de Áudio Digital

A representação digital de sinais sonoros é a base para o desenvolvimento de sistemas automáticos de análise musical, como é o caso da classificação de gêneros. Neste contexto, é essencial compreender como os sinais acústicos são discretizados e quais propriedades perceptuais e físicas são extraídas. Esta seção se apoia principalmente no livro de [Müller 2015] para apresentar os conceitos fundamentais do processamento de áudio.

2.2.1. Representação Discreta de Sinais Musicais

O som, no domínio analógico, é uma variação contínua da pressão do ar ao longo do tempo. Para que esse sinal possa ser manipulado computacionalmente, é necessário convertê-lo em uma sequência de valores discretos, processo realizado por meio da amostragem e da quantização. A amostragem consiste em registrar o valor do sinal em intervalos regulares de tempo, definidos pela taxa de amostragem, geralmente expressa em Hz. A quantização, por sua vez, transforma os valores contínuos das amostras em números finitos, geralmente inteiros, representáveis digitalmente [Müller 2015, p. 60–62].

O Teorema de Nyquist–Shannon estabelece que a taxa de amostragem deve ser pelo menos o dobro da maior frequência presente no sinal, sob pena de ocorrência de aliasing, um artefato que distorce a representação do conteúdo espectral [Müller 2015, p. 50–51]. Um exemplo prático é a taxa de 44 100 Hz utilizada em áudio digital (como em CDs), capaz de representar sinais com frequências de até 22.050 Hz.

2.2.2. Propriedades Acústicas e Perceptuais do Áudio

As propriedades de um sinal sonoro podem ser descritas por grandezas físicas, mas também estão associadas a percepções humanas fundamentais [Müller 2015, p. 19–56]:

- **Amplitude:** valor máximo do deslocamento da onda sonora em torno do ponto de equilíbrio, diretamente relacionada à intensidade percebida. Sons mais altos apresentam maior amplitude.
- **Frequência:** número de oscilações por segundo (Hz), associado à percepção de altura tonal.
- **Pitch (altura):** percepção subjetiva da frequência fundamental de um som. Embora associada à frequência física, a percepção de pitch depende de fatores como contexto harmônico e conteúdo espectral.
- **Tom:** termo muitas vezes usado como sinônimo de pitch, mas que pode incluir também a componente harmônica percebida de uma nota (como em timbres instrumentais).
- **Intensidade e Dinâmica:** variação de energia ao longo do tempo, capturada por medidas como Short-Time Energy (STE) ou Root Mean Square Energy (RMSE).
- **Timbre:** qualidade que distingue sons com mesma frequência e intensidade, relacionada à distribuição espectral e harmônicos de um instrumento.
- **Distorção:** refere-se à introdução de harmônicos não presentes no sinal original ou à modificação não linear da forma de onda.

2.2.3. Características Musicais: Harmonia, Melodia e Ritmo

Do ponto de vista perceptivo-musical, os sinais de áudio carregam propriedades estruturais que auxiliam na identificação de estilos musicais:

- **Melodia:** sequência linear de notas musicais com variação de pitch, que forma a linha principal de uma música. Pode ser extraída a partir da estimativa de pitch ao longo do tempo [Müller 2015, p. 49–50].

- **Harmonia:** combinação simultânea de notas ou frequências, normalmente representada no espectro de frequências como a presença de múltiplos harmônicos. A análise espectral e o uso de filtros harmônicos são estratégias para modelar tais relações [Müller 2015, p. 54–55].
- **Ritmo:** organização temporal dos eventos sonoros. Pode ser inferido por meio da detecção de onsets, análise de periodicidade e padrões de batida (beat tracking). [Müller 2015, p. 305–333].

2.2.4. Características Clássicas de Áudio

Com base em [Müller 2015, p. 24–26, 77–80], [Schlüter and Grill 2015] e [Tzanetakis and Cook 2002], apresentam-se a seguir as principais características clássicas extraídas diretamente do sinal de áudio, amplamente utilizadas em tarefas de análise e classificação musical [Guimarães et al. 2025]:

- **Zero-Crossing Rate (ZCR):** taxa de cruzamentos de zero por quadro, frequentemente associada à presença de componentes de alta frequência.
- **Short-Time Energy (STE) e RMS Energy:** medidas de energia local que refletem variações de intensidade e dinâmica do sinal.
- **Envelope Temporal:** curva suavizada que evidencia ataques e decaimentos, contribuindo para a caracterização da articulação sonora.
- **Mel Frequency Cepstral Coefficients (MFCCs):** coeficientes extraídos a partir de um banco de filtros na escala mel, que sintetizam o conteúdo espectral de forma compacta e perceptualmente relevante.
- **Spectral Centroid:** valor médio ponderado das frequências presentes, relacionado à percepção de brilho do som.
- **Spectral Bandwidth:** medida da dispersão das frequências em torno do centróide, associada à riqueza harmônica.
- **Spectral Rolloff:** frequência abaixo da qual se concentra uma porcentagem definida da energia espectral (por exemplo, 85%), útil na distinção entre sons harmônicos e percussivos.
- **Chroma Features:** representação da energia distribuída entre as 12 classes de altura tonal (notas da escala temperada), permitindo capturar aspectos harmônicos e de progressões de acordes.

2.2.5. Visualizações no Domínio Tempo–Frequência

Uma das formas mais eficazes de representar sinais sonoros em tarefas de reconhecimento é o espectrograma, obtido por meio da Short-Time Fourier Transform (STFT). Essa técnica segmenta o sinal em janelas sobrepostas e aplica a transformada de Fourier em cada segmento, gerando uma matriz tempo–frequência que revela o conteúdo espectral ao longo do tempo, onde a apresentação espectral resultante é comumente chamada de espectrograma [Müller 2015, p. 94–98]. Complementarmente, o mel-espectrograma aplica um banco de filtros na escala mel, uma escala perceptual baseada em como o ouvido humano percebe variações de frequência. Essa representação comprime as regiões de alta frequência e expande as de baixa frequência, tornando-a mais adequada para tarefas perceptuais e modelos de aprendizado profundo [Müller 2015, p. 98–102].

A Figura 2.1 exemplifica a transição entre as representações no domínio do tempo e no domínio tempo–frequência, destacando como a estrutura espectral do sinal pode ser visualmente analisada por meio do mel-espectrograma.

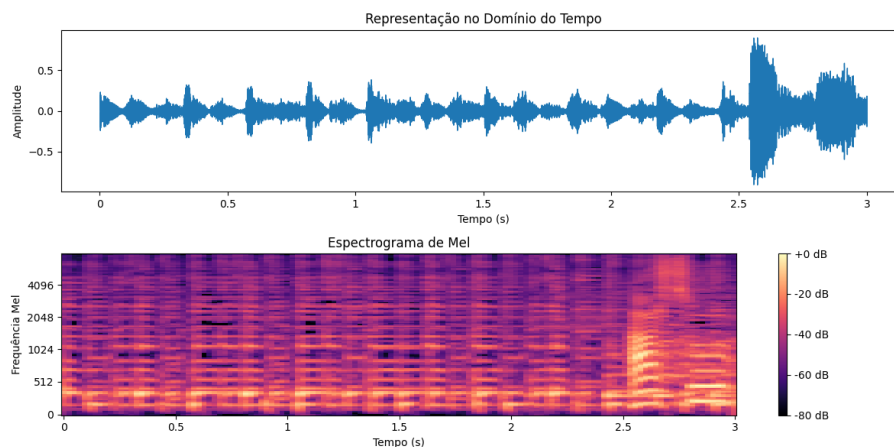


Figura 2.1. Forma de onda do sinal de áudio (acima) e seu respectivo mel-espectrograma (abaixo), evidenciando a distribuição de energia em bandas perceptuais ao longo do tempo.

Além das características espectrais e temporais tradicionais, como MFCCs (Mel Frequency Cepstral Coefficients) e a taxa de cruzamento zero, o uso de mel-espectrogramas tem se mostrado particularmente poderoso em aplicações de classificação musical e reconhecimento de padrões [Schlüter and Grill 2015]. Essa representação permite capturar nuances tonais de forma mais alinhada à percepção humana [Dieleman and Schrauwen 2014, Cheng et al. 2020]. Alguns parâmetros importantes na construção de mel-

espectrogramas incluem [Müller 2015, Gold et al. 2011]:

- **Tamanho da Janela (window size):** número de amostras processadas em cada transformação. Janelas maiores oferecem melhor resolução de frequência, enquanto janelas menores fornecem maior resolução temporal;
- **Deslocamento da Janela (hop size):** número de amostras de avanço entre janelas consecutivas. Um deslocamento de 50% é comum para garantir sobreposição adequada;
- **Número de Bandas Mel:** quantidade de filtros na escala mel aplicados ao espectrograma. Mais bandas permitem uma análise tonal mais detalhada;
- **Taxa de Amostragem:** número de amostras de áudio capturadas por segundo.

2.3. Redução de Dimensionalidade

Modelos de aprendizado profundo aprendem representações internas dos dados chamadas de *embeddings*. Esses vetores de comprimento fixo codificam características latentes extraídas de entradas complexas, como sinais de áudio ou espectrogramas, e são capazes de capturar relações semânticas ou estilísticas entre os dados. Diferentemente de representações tradicionais, como vetores esparsos ou *one-hot*, os *embeddings* aprendidos posicionam entidades similares mais próximas entre si no espaço vetorial, permitindo generalizações mais eficazes [Goodfellow et al. 2016].

No contexto da análise de gêneros musicais, os *embeddings* permitem representar cada faixa como um vetor em um espaço de alta dimensionalidade, no qual é possível investigar similaridades e relações estilísticas. Essa representação vetorial possibilita o estudo da separabilidade entre os gêneros, bem como a identificação de sobreposições e agrupamentos naturais que refletem características musicais compartilhadas [Van den Oord et al. 2013]. Como esses vetores habitam espaços com dezenas ou centenas de dimensões, sua visualização direta não é viável. Para facilitar a interpretação, são aplicadas técnicas de redução de dimensionalidade, como PCA, t-SNE e UMAP, que projetam os *embeddings* em duas ou três dimensões, preservando suas relações de proximidade.

A Análise de Componentes Principais (PCA) [Wold et al. 1987] é uma técnica linear de redução de dimensionalidade que projeta os dados em novas direções chamadas de componentes principais, buscando maximizar a variância explicada. Essa abordagem é eficaz para capturar a estrutura global dos dados e identificar padrões de maior dispersão,

mas apresenta limitações na preservação de relações locais quando os dados possuem distribuições não lineares.

A técnica t-Distributed Stochastic Neighbor Embedding (t-SNE) [Van der Maaten and Hinton 2008] é uma abordagem não linear que prioriza a preservação das vizinhanças locais. Ela é especialmente eficaz para revelar agrupamentos e transições entre categorias, como os diferentes gêneros musicais, embora possa distorcer as distâncias globais, dificultando a interpretação da estrutura completa do espaço.

O Uniform Manifold Approximation and Projection (UMAP) [McInnes et al. 2018] combina os pontos fortes do t-SNE com maior escalabilidade e melhor fidelidade às estruturas locais e globais. UMAP tem sido amplamente adotado em contextos de aprendizado profundo por proporcionar visualizações mais estáveis e interpretáveis, o que facilita a análise da organização dos *embeddings* em espaços de alta dimensionalidade.

Essas técnicas, em conjunto, formam a base para a interpretação dos espaços latentes gerados por modelos de aprendizado profundo, possibilitando uma análise mais rica da estrutura interna dos dados musicais e das relações entre os gêneros [Guimarães et al. 2025b].

2.4. Similaridade

A similaridade entre faixas musicais é um aspecto relevante tanto do ponto de vista perceptivo quanto computacional. No reconhecimento de gêneros, é comum que diferentes estilos compartilhem características como ritmo, instrumentação, melodia e timbre, o que dificulta sua distinção. Essa semelhança musical pode gerar confusão tanto para ouvintes quanto para sistemas automatizados. Para além da análise acústica, abordagens baseadas em aprendizado profundo permitem representar músicas por meio de vetores numéricos, chamados *embeddings*.

2.4.1. Similaridade Musical

A questão da similaridade entre gêneros musicais também é um fator importante utilizado para distinguir entre os diferentes gêneros musicais existentes. Por exemplo, a instrumentação de uma música é um fator determinante para distinguir gêneros musicais, especialmente no reconhecimento automatizado [Percino et al. 2014]. Gêneros com uma baixa uniformidade instrumental, como música experimental ou jazz, tendem a ser mais

facilmente identificados, pois utilizam instrumentos específicos que não são comuns em outros gêneros. Por outro lado, gêneros mais populares, como o pop, que possuem uma alta uniformidade instrumental, ou seja, usam os mesmos instrumentos que muitos outros estilos, podem ser mais difíceis de diferenciar.

Além da instrumentação, outras características como timbre, ritmo e melodia também podem ser consideradas para o cálculo da similaridade. Técnicas de aprendizado de máquina, como os modelos de mistura gaussiana ou GMM (do inglês, *Gaussian Mixture Models*) e k-vizinhos mais próximos, ou KNN (do inglês, *k-nearest neighbors*), são amplamente utilizados para reconhecer padrões acústicos que diferenciam gêneros [Muszynski and Tarapata 2023]. Entretanto, a similaridade entre gêneros ainda é um desafio para esses classificadores. Por exemplo, gêneros com padrões rítmicos similares, como samba e pagode, podem gerar confusões, a menos que o sistema capture nuances mais detalhadas da instrumentação ou da harmonia.

De forma geral, identificar gêneros musicais se torna ainda mais difícil quando estes são considerados similares a outro, especialmente em contextos onde há uma sobreposição de elementos musicais, como instrumentação, ritmo e estilo. Os gêneros regionais são um exemplo dessa dificuldade, pois muitas vezes compartilham características similares com gêneros vizinhos, o que torna a classificação desses gêneros difícil até para o ouvido humano.

2.4.2. Similaridade Vetorial

Com o avanço dos modelos de aprendizado profundo, tornou-se possível representar faixas musicais por meio de vetores em espaços de alta dimensionalidade, denominados *embeddings*. Essas representações vetoriais capturam características latentes complexas extraídas dos dados de entrada, como espectrogramas de áudio, e codificam informações semânticas e estilísticas relevantes. Essa representação permite que diferentes faixas musicais sejam comparadas quantitativamente com base em sua proximidade no espaço vetorial, o que é fundamental para tarefas como classificação e análise de similaridade entre gêneros.

Uma das métricas mais utilizadas para quantificar a similaridade entre *embeddings* é a similaridade cosseno. Essa medida calcula o cosseno do ângulo entre dois vetores, desconsiderando suas magnitudes e focando apenas em sua orientação no espaço. Essa

propriedade é especialmente útil para compensar diferenças de escala entre vetores, o que ocorre, por exemplo, quando diferentes faixas possuem intensidades sonoras distintas mas distribuições de características semelhantes [Manning 2009]. Matematicamente, dada duas representações vetoriais x e y em $\mathbb{R}^n$, a similaridade cosseno é definida na Equação 2.1:

$$\text{sim}_{\cos}(x, y) = \frac{x \cdot y}{\|x\| \|y\|} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \cdot \sqrt{\sum_{i=1}^n y_i^2}} \quad (2.1)$$

Para que essa métrica reflita de forma mais real as relações estilísticas aprendidas pelo modelo, é fundamental que o cálculo da similaridade seja realizado diretamente no espaço original dos *embeddings*, sem a aplicação de técnicas de redução de dimensionalidade. Isso garante que as propriedades semânticas codificadas durante o treinamento do modelo sejam preservadas, permitindo análises mais precisas sobre a distribuição e a separação entre os gêneros musicais [Spijkervet and Burgoyne 2021, Barnett et al. 2024].

A Figura 2.2 ilustra três situações distintas da similaridade cosseno entre dois vetores $\vec{x}$ e $\vec{y}$, representados em planos cartesianos separados:

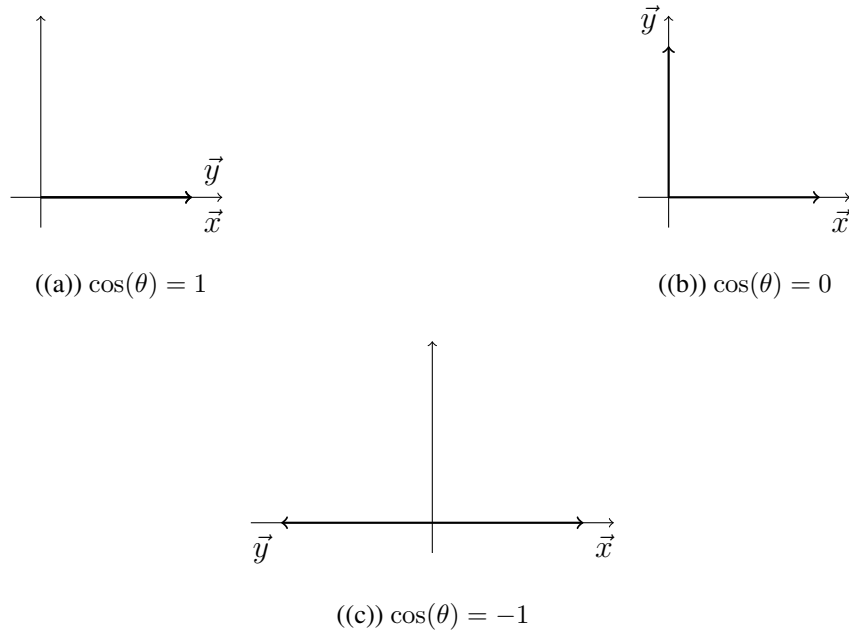


Figura 2.2. Exemplos em planos cartesianos distintos ilustrando os valores da similaridade cosseno: (a) máxima similaridade, (b) nenhuma similaridade, (c) máxima dissimilaridade.

As três situações ilustram a variação do valor de $\cos(\theta)$ no intervalo $[-1, 1]$, que

é o intervalo natural da função cosseno, e demonstram como a orientação angular entre vetores pode ser usada para inferir o grau de similaridade entre representações vetoriais.

Na Figura 2.2(a), os vetores $\vec{x}$ e $\vec{y}$ apontam exatamente na mesma direção. Isso indica máxima similaridade, ou seja, os *embeddings* representados são praticamente idênticos em termos de direção. É o cenário ideal em tarefas de recuperação de informação ou classificação baseada em similaridade. Em 2.2(b), os vetores são ortogonais (ângulo de 90°), o que significa que não há correlação entre eles. Na prática, isso representa ausência de similaridade entre os *embeddings*. Esse caso é importante para indicar classes ou itens sem nenhuma relação vetorial. Por fim, na Figura 2.2(c), os vetores apontam em direções opostas, formando um ângulo de 180°. Isso representa máxima dissimilaridade, ou seja, os vetores são tão diferentes quanto possível no espaço vetorial.

A análise vetorial baseada em embeddings complementa a avaliação qualitativa ao revelar como o modelo organiza internamente os diferentes estilos musicais. Por meio da observação da densidade, dispersão e agrupamento dos vetores, é possível explorar padrões de similaridade e investigar quais gêneros apresentam maior sobreposição no espaço latente aprendido.

2.5. Aprendizado Profundo

Esta seção apresenta as arquiteturas e técnicas em aprendizado profundo utilizadas na literatura para reconhecimento de áudio.

2.5.1. Técnicas de Aprendizado Profundo

De acordo com Haykin, redes neurais artificiais ou ANN (do inglês, *Artificial Neural Network*), referenciada apenas como redes neurais, foram motivadas desde a sua concepção pelo reconhecimento de que o cérebro humano computa de uma forma totalmente diferente da forma como é realizada a computação por computadores digitais [Haykin 1994]. Uma ANN pode ser definida por ser uma estrutura caracterizada por elementos de processamento simples (neurônio ou nó), que são capazes de realizar operações de processamento de dados e representação do conhecimento [Basheer and Hajmeer 2000]. Sua arquitetura é formada por uma camada de entrada, na qual cada neurônio artificial recebe um valor de entrada. Possui também camadas intermediárias onde é feito a maioria do processamento, e por fim, a camada de saída na qual o resultado é concluído.

Antes da popularização do aprendizado profundo, técnicas de classificação baseadas na maximização da margem entre as classes ganharam destaque, como é o caso do trabalho seminal de Boser, Guyon e Vapnik, que apresentou um algoritmo de treinamento para classificadores de margem ótima [Boser et al. 1992]. A solução proposta, mais tarde consolidada como as Máquinas de Vetores de Suporte (SVMs), fundamenta-se na ideia de encontrar um hiperplano que maximize a separação entre os exemplos de diferentes classes, utilizando apenas os vetores mais próximos à fronteira de decisão, conhecidos como vetores de suporte. Essa abordagem é especialmente eficaz em cenários com dados limitados e contribuiu significativamente para a compreensão dos princípios de generalização em modelos supervisionados.

Com o avanço das redes neurais, o aprendizado profundo se consolidou como uma subárea do campo de aprendizado de máquina. Diferente de uma ANN simples, o aprendizado profundo utiliza uma maior quantidade de camadas intermediárias onde cada camada recebe como entrada o resultado da camada anterior [Russell and Norvig 2020]. A primeira arquitetura desenvolvida foi com a apresentação dos algoritmos perceptron [Rumelhart et al. 1986], que ainda são utilizados na literatura em problemas de reconhecimento de áudios musicais [Silva et al. 2021a].

As redes neurais convolucionais, chamadas de CNNs [LeCun et al. 1998] têm sido amplamente utilizadas no campo de reconhecimento e classificação de áudio, uma vez que muitas vezes fornecem resultados positivos para esta tarefa [Esposito et al. 2021, Cheng et al. 2020]. Para processar áudio utilizando CNNs, o áudio deve ser transformado em uma representação visual ou numérica que preserve a informação temporal e espectral. As representações mais comuns incluem espectrogramas, espectrogramas mel e coeficientes cepstrais de frequência mel, ou MFCCs [Purwins et al. 2019].

As CNNs são uma categoria específica de redes neurais profundas que empregam a operação de convolução para processar os dados de entrada [Colliot 2023], essa rede neural possui conexões locais, especialmente nas camadas iniciais, o que significa que cada neurônio está conectado apenas na região da camada anterior. Esses padrões de pesos são replicados em várias unidades da mesma camada, o que é conhecido como kernel. O processo de aplicação do kernel aos píxeis da imagem é chamado de convolução [Russell and Norvig 2020]. Os kernels e convoluções são mais fáceis de ilustrar em uma dimensão visual, onde um kernel move-se sobre a imagem, aplicando a mesma operação

a diferentes partes da imagem para extrair características como bordas, texturas e outros padrões locais.

Entre os avanços importantes em arquiteturas profundas está a introdução das Redes Neurais Residualizadas, sendo a ResNet50 uma das variantes mais influentes [He et al. 2016]. A arquitetura ResNet introduziu o conceito de conexões residuais entre camadas, permitindo o treinamento eficaz de redes muito profundas sem os problemas de degradação de desempenho. Essas conexões facilitam a propagação do gradiente, melhorando a estabilidade e a capacidade de generalização do modelo, o que torna a ResNet50 amplamente adotada em tarefas envolvendo classificações de imagens e, por extensão, representações espectrais de áudio.

Redes neurais recorrentes são redes de aprendizado profundo voltadas para processamento de dados sequenciais [Hopfield 1982]. De acordo com Russel, RNNs permitem *loops* em neurônios intermediários, onde cada ciclo tem um atraso para que as unidades possam tomar como entrada um valor calculado a partir de sua própria saída em uma etapa anterior do cálculo [Russell and Norvig 2020]. Essa estrutura de rede oferece possibilidade para computar sequências de comprimento variável, no qual os dados têm relação um com o outro ao decorrer do tempo. Desta forma, este tipo de rede se torna objeto de estudo em problemas de detecção de gêneros musicais, visto que reconhecem padrões em sequências de dados. A evolução das arquiteturas de redes neurais profundas, desde os primeiros modelos como perceptrons e redes recorrentes até estruturas mais recentes baseadas em mecanismos de atenção, pode ser visualizada de forma resumida na Figura 2.3.

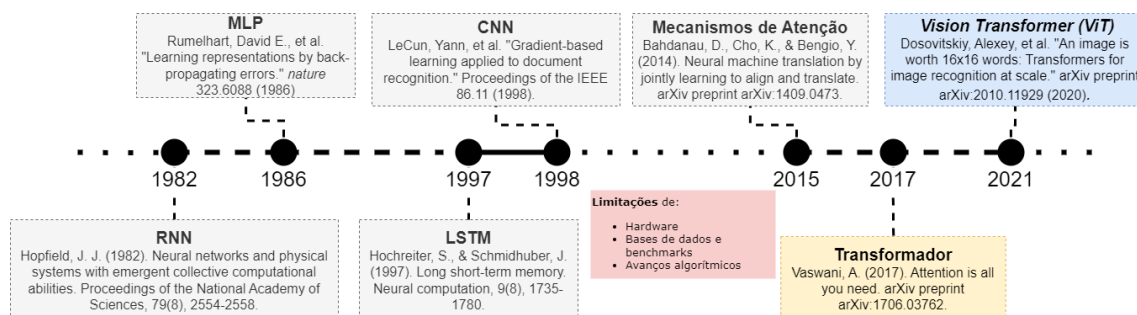


Figura 2.3. Evolução das arquiteturas de redes neurais profundas com destaque para o surgimento dos transformadores e sua aplicação em visão computacional.

Essa linha do tempo destaca os principais marcos teóricos e práticos que influenciaram o desenvolvimento de técnicas em aprendizado profundo, com impacto direto em áreas como análise de áudio e visão computacional. O surgimento de arquiteturas baseadas em transformadores marca uma mudança significativa nesse panorama, tema que será aprofundado na próxima subseção.

2.5.2. Mecanismos de Atenção e Transformadores

O mecanismo de atenção permite que o modelo considere as partes da frase relevantes para prever a próxima palavra e facilita a compreensão das relações entre palavras que estão mais distantes [Bahdanau et al. 2014]. A autoatenção, proposta por Vaswani [Vaswani et al. 2017] formaliza o conceito de contexto, onde aprende os padrões fundamentais de como partes da entrada correspondem umas às outras. Ao reunir informações do mesmo conjunto, dada uma sequência de *tokens*, um *token* pode focar em seus *tokens* vizinhos para calcular sua saída. Isso significa que cada *token* pode consultar *tokens* vizinhos para entender melhor seu contexto, ajudando o modelo a capturar relações entre *tokens* próximos e distantes na mesma entrada. De forma geral, enquanto a atenção pode se concentrar em diferentes entradas (por exemplo, de uma sentença original para sua tradução), a autoatenção foca em diferentes partes da mesma entrada, ajudando a capturar o contexto dentro de uma única sequência de dados.

O Transformador é uma arquitetura de rede neural profunda introduzida no artigo "Attention Is All You Need" [Vaswani et al. 2017]. Esta arquitetura foi desenvolvida para lidar com tarefas de tradução de sequência, como tradução de idiomas, e substitui a necessidade de redes neurais recorrentes (RNNs) ou convolucionais (CNNs) por um mecanismo totalmente baseado em atenção [Vaswani et al. 2017]. A arquitetura do transformador revolucionou como as redes neurais processam sequências, introduzindo o conceito de autoatenção, o qual permite que cada palavra de uma sequência de entrada, considere todo o contexto da frase, em vez de apenas as palavras anteriores. O que é semelhante a como uma pessoa presta atenção em diferentes partes de uma conversa ou texto, e tenta entender o contexto do que foi passado.

A arquitetura é composta por duas camadas principais, chamadas de *decoder* e *encoder* que pode ser vista na Figura 2.4. A camada *encoder* é responsável por processar a entrada original e transformá-la em uma representação intermediária, que captura as

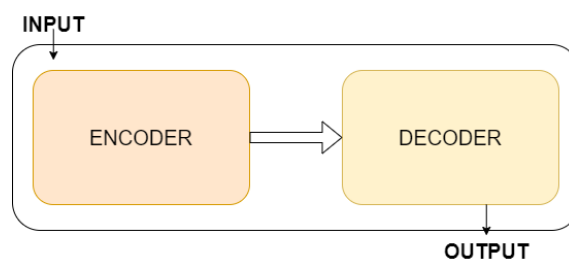


Figura 2.4. Relação *encoder* e *decoder* presentes em uma arquitetura de redes de transformadores.

relações contextuais entre as palavras da sequência. A camada *decoder* é responsável por gerar a saída final a partir da representação intermediária fornecida pelo *encoder*. Assim como o *encoder*, o *decoder* é composto por múltiplos blocos empilhados, mas com uma arquitetura ligeiramente diferente. Combinando essas duas camadas principais, o transformador consegue transformar uma sequência de entrada em uma sequência de saída de maneira altamente eficiente e precisa.

Nos últimos anos, a arquitetura de transformadores tem sido amplamente adaptada e modificada para diferentes tarefas, indo além do processamento de linguagem natural, seu campo de origem, e se expandindo para áreas como visão computacional. Uma variação notável é o transformador de visão computacional ou ViT (do inglês, *Vision Transformer*) [Dosovitskiy et al. 2020], que utiliza sequências de imagens, onde as divide em *patches* de tamanho fixo e trata cada *patch* como um *token* na sequência. O ViT se destacou por alcançar resultados de ponta em *baselines* de reconhecimento de imagens, como o ImageNet, especialmente quando pré-treinado em grandes volumes de dados. Uma das principais inovações do ViT é a capacidade de superar as redes neurais convolucionais tradicionais, como a ResNet [He et al. 2016], em tarefas de classificação de imagens, mesmo com menos recursos computacionais durante o treinamento. A simplicidade e a escalabilidade do ViT o tornam uma alternativa poderosa para tarefas que anteriormente eram dominadas por CNNs.

2.5.3. Aprendizado por Transferência

Aprendizado por transferência é uma técnica no campo do aprendizado de máquina onde um modelo treinado em uma tarefa, geralmente com uma grande quantidade de dados, é reutilizado como ponto de partida para uma outra tarefa relacionada [Yosinski et al. 2014]. Em vez de treinar um modelo completo a partir de sua inicialização padrão para

uma nova tarefa, o aprendizado por transferência permite que o conhecimento adquirido anteriormente seja transferido para o novo modelo, acelerando o processo de treinamento e possivelmente resultando em melhor desempenho [Goodfellow et al. 2016]. Esse processo envolve geralmente o ajuste fino de um modelo pré-treinado em uma tarefa específica, como a classificação de imagens em larga escala, e aplicá-lo a uma nova tarefa que pode ter um conjunto de dados menor ou mais específico.

2.5.4. Métricas de Avaliação

Para problemas de classificação, uma maneira de verificar o desempenho de um algoritmo é através da **matriz de confusão**, que mostra a correspondência entre as classes observadas e aquelas previstas pelo modelo. A matriz é dividida em quatro partes: verdadeiros positivos (VP), falsos positivos (FP), verdadeiros negativos (VN) e falsos negativos (FN) [Sammut and Webb 2011]. Esses valores também são usados para calcular métricas importantes como **acurácia**, **precisão**, **revocação** e **F1-score** [Powers 2011, Sokolova and Lapalme 2009]. A Tabela 2.1 apresenta as fórmulas para calcular as métricas, juntamente com suas definições.

Tabela 2.1. Definições e Fórmulas das Principais Métricas de Avaliação.

Métrica	Fórmula	Definição
Acurácia	$\frac{VP+VN}{VP+VN+FP+FN}$	A proporção de instâncias corretamente previstas (tanto positivas quanto negativas) em relação ao total de instâncias.
Recall	$\frac{VP}{VP+FN}$	Proporção de verdadeiros positivos em relação ao total de exemplos que realmente são positivos.
Precisão	$\frac{VP}{VP+FP}$	Proporção de verdadeiros positivos em relação ao total de exemplos classificados como positivos.
F1-Score	$\frac{2 \times \text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}}$	A média harmônica entre precisão e recall, equilibrando ambas as métricas.

Outro conceito importante de métrica, é durante a avaliação de modelos de aprendizado de máquina onde é necessário observar a **função de perda**, responsável por quantificar a diferença entre as previsões do modelo e os valores reais [Goodfellow et al. 2016]. A redução da função de perda leva frequentemente a um aumento na acurácia do modelo, pois indica que as previsões estão mais próximas dos valores reais. Entretanto, uma diminuição na função de perda nem sempre garante um aumento na precisão, especialmente se o modelo sofrer de **sobreajuste** (do inglês, *overfitting*), que ocorre quando o modelo

aprende muito bem os detalhes dos dados de treinamento, e não consegue generalizar bem para novos dados.

De outra forma, o **subajuste** (do inglês, *underfitting*) acontece quando o modelo falha em capturar a estrutura dos dados de treinamento, levando a uma baixa de desempenho tanto no conjunto de treinamento quanto nos conjuntos de teste, sendo em geral, consequência de um modelo excessivamente simples, de treinamento insuficiente ou de uma representação inadequada dos dados [Goodfellow et al. 2016].

Capítulo 3

Reconhecimento de Gêneros Musicais

Este capítulo apresenta a definição geral do problema e os detalhes envolvidos, além dos trabalhos relacionados a esta pesquisa.

3.1. Definição Geral do Problema de Reconhecimento de Gêneros Musicais

Com o avanço das tecnologias digitais, houve um crescimento significativo na disponibilização de recursos musicais digitais, resultando em vastas bibliotecas musicais online. Esse grande volume de dados gera a necessidade de mecanismos eficazes para a organização e recuperação de conteúdos musicais, sendo a classificação automática de gêneros uma das abordagens mais utilizadas [Downie 2003].

Apesar dos avanços recentes em métodos computacionais aplicados ao reconhecimento automático de gêneros musicais, essa tarefa continua sendo desafiadora, especialmente quando envolve gêneros com elevado grau de similaridade. Um exemplo claro dessa dificuldade está na música regional brasileira, onde diferentes gêneros frequentemente compartilham estruturas rítmicas, padrões harmônicos, instrumentações e contextos culturais semelhantes [Correa 2018, Castro 2010, Sturm 2014, de Araújo Lima et al. 2020]. Além disso, faixas híbridas, fusões estilísticas e a fluidez característica da música popular tornam os limites entre gêneros especialmente pouco definidos.

Entre as diversas técnicas já exploradas, destacam-se abordagens que analisam a melodia, realizam separação de fontes (como voz ou instrumentos específicos) ou extraem diferentes aspectos acústicos das faixas musicais [Salamon et al. 2012]. Contudo, esses métodos enfrentam limitações decorrentes da grande variedade de instrumentos, timbres e

melodias existentes nas músicas populares, bem como da coexistência frequente de múltiplos gêneros em uma mesma faixa [Medhat et al. 2020]. Ainda que exista uma vasta literatura sobre classificação musical, uma lacuna importante permanece aberta quanto ao reconhecimento de gêneros altamente similares e pouco explorados, como aqueles encontrados na música regional brasileira, marcada por particularidades culturais e estilísticas únicas [de Araújo Lima et al. 2020, Muszynski and Tarapata 2023, Percino et al. 2014].

Cabe enfatizar que os gêneros musicais são construções sociais moldadas por fatores culturais, estéticos e contextuais, não sendo categorias rígidas [Krogh 2019]. Modelos de aprendizado profundo aprendem essas categorias por meio de padrões estatísticos extraídos dos dados, gerando representações internas (*embeddings*) que refletem tanto as características sonoras quanto as relações socioculturais subjacentes.

Dessa forma, o presente trabalho visa não apenas a classificação automática de gêneros musicais regionais brasileiros com alta similaridade, mas também busca entender como esses gêneros se organizam no espaço vetorial gerado pelos modelos. Além disso, são investigados diferentes trechos musicais quanto às suas posições e durações, considerando que certas partes, como refrões, podem ser mais representativas e informativas para a percepção do gênero musical [Byron et al. 2025].

Com base nessas observações, é possível identificar a relevância de se investigar metodologias capazes de explorar representações musicais mais ricas e precisas. Nesse sentido, a próxima seção apresenta trabalhos relacionados que fundamentam e motivam a proposta metodológica adotada nesta pesquisa.

3.2. Trabalhos Relacionados

Esta seção será dividida em três partes. Na primeira, serão abordados os principais trabalhos relacionados na área de MGR, com ênfase nas técnicas e arquiteturas utilizadas para o reconhecimento de gêneros musicais, além das bases de dados empregadas nesses estudos. A segunda será dedicada a uma análise detalhada das bases de dados encontradas na literatura. Por fim, a terceira parte apresenta uma comparação dos trabalhos mais alinhados a esta dissertação, com foco em estudos que investigam estratégias de segmentação temporal, o uso de diferentes trechos das músicas, ou a aplicação de técnicas de redução de dimensionalidade para análise e visualização de *embeddings* musicais.

3.2.1. Trabalhos em MGR

Na Tabela 3.1 pode-se verificar os principais trabalhos relacionados, com ênfase na arquitetura utilizada para abordar o problema de reconhecimento de gêneros musicais.

Tabela 3.1. Tabela comparativa dos principais trabalhos relacionados da literatura para a tarefa de classificação de gêneros musicais.

Aprendizado	Referência	Título	Técnica	Base de Dados	Característica do Áudio
Clássico	[Barbosa et al. 2015]	Evaluating Automated Classification Techniques for Folk Music Genres from the Brazilian Northeast	SVM, Naive Bayes, Árvore de Decisão, KNN, e outros	Base Própria	Espectral, temporal e ritmo
	[De Sousa et al. 2016]	A Robust Music Genre Classification Approach for Global and Regional Music Datasets Evaluation	SVM	GTZAN, Base Própria	Espectral, temporal, e ritmo
	[Conceição et al. 2020]	Reconhecendo Gêneros Musicais Brasileiros com Técnicas de Aprendizagem de Máquina Supervisionada	KNN, SVM, Floresta Aleatória	GTZAN, Base Própria	Espectral e temporal
	[Silva et al. 2021a]	Modelo Automático de Classificação de Gêneros Musicais Amazônicos	KNN, XGB, SVM e MLP	Base Própria	Espectral e temporal
Profundo	[Dai et al. 2016]	Long Short-term Memory Recurrent Neural Network-based Segment Features for Music Genre Classification	LSTM	ISMIR	MFCCs
	[Cheng et al. 2020]	Convolutional Neural Networks Approach for Music Genre Classification	CNN	GTZAN	Mel-espectrograma
	[Medhat et al. 2020]	Masked Conditional Neural Networks for Sound Classification	MCNN	Ballroom, GTZAN, ISMIR2004, Urbandound8k, ESC-10 e ESC-50	Mel-espectrograma
	[Zhuang et al. 2020]	Music Genre Classification with Transformer Classifier	Transformer	GTZAN	Mel-espectrograma
	[Silva et al. 2021c]	Fusion of Multiple Audio Representations for Music Genre Classification	CNN	GTZAN, Homburg, Extended Ballroom, FMA	Mel-espectrograma, MFCC, CQT, Chroma, Tempogram
	[Silva et al. 2021b]	On Generalist and Domain-Specific Music Classification Models and Their Impacts on Brazilian Music Genre Recognition	CNN	4MuLA	Mel-espectrograma
	[Zhao et al. 2022]	S3T: Self-Supervised Pre-Training with Swin Transformer for Music Classification	Transformer	GTZAN, FMA, MagnaTagATune	CQT-espectrograma
	[Wijaya et al. 2024]	Music-Genre Classification using Bidirectional Long ShortTerm Memory and Mel-Frequency Cepstral Coefficients	LSTM	GTZAN, ISMIR	MFCCs
	[Xie et al. 2024]	Music Genre classification Based on res-gated CNN and Attention Mechanism	CNN e Transformer	GTZAN, Ballroom	Mel-espectrograma

No problema de reconhecimento de gêneros musicais, muitos estudos na literatura abordam essa questão utilizando algoritmos de aprendizado de máquina supervisionado sem a aplicação de arquiteturas profundas, como para a classificação de música regional, utilizando SVM (Support Vector Machine), Naive Bayes, Árvore de Decisão, kNN,

Multilayer Perceptron (MLP), AdaBoost e Floresta Aleatória [Barbosa et al. 2015]. A base de dados utilizada neste trabalho consistiu em 75 amostras de cinco gêneros de músicas características do nordeste brasileiro, onde foram extraídas características de áudio espectrais, temporais e rítmicas. Entre os algoritmos utilizados, o Naive Bayes e o SVM apresentaram o melhor desempenho, ambos com uma acurácia de aproximadamente 60%, refletindo que a similaridade entre os gêneros regionais representa um desafio significativo para uma classificação mais precisa.

Da mesma maneira, outro trabalho para reconhecer gêneros musicais regionais utilizou algoritmos clássicos, como KNN, SVM e Floresta Aleatória, para realizar a classificação, chegando a uma acurácia de 79,7% com o modelo SVM, além disso, os autores desenvolveram uma base regional brasileira que obteve acurácia de 86.11% [De Sousa et al. 2016]. Semelhante a essa abordagem, outro autor conseguiu com o algoritmo de floresta aleatória uma precisão de 65,3% para os gêneros brasileiros, e para o experimento com a base GTZAN mais músicas brasileiras, alcançou a precisão de 64,2% [Conceição et al. 2020].

Seguindo esta linha, outro trabalho abordou a falta de pesquisas que contemplem gêneros regionais, focando em criar uma base de dados e utilizar modelos de aprendizado de máquina para essa tarefa, onde o estudo inclui a criação de uma base de dados contendo músicas de diversos gêneros amazônicos, como: andino, brega, carimbó, cumbia, marabaixo, merengue, pasillo, salsa e vaqueirada [Silva et al. 2021a]. Diferente dos anteriores, o modelo de redes neurais MLP apresentou os melhores resultados entre os modelos testados, chegando a uma acurácia de 76,13%.

Trabalhando com abordagem de arquiteturas profundas, existem trabalhos que abordam o uso de redes neurais recorrentes para melhorar a classificação de gêneros musicais [Dai et al. 2016]. A abordagem se concentra em capturar as dependências temporais dos dados de áudio, as quais os autores identificam que são frequentemente ignoradas em métodos baseados em características extraídas de pequenos segmentos temporais (quadros) de um sinal de áudio. Utilizando a abordagem com características de quadro, o modelo LSTM alcançou uma acurácia de 83,78%, e com a abordagem com características de segmento, alcançou a acurácia de 86,42% [Dai et al. 2016].

Outro trabalho apresenta uma abordagem para classificação de gêneros musicais

usando uma combinação de redes LSTM bidirecionais e coeficientes de frequência mel-cepstrais (MFCCs) [Wijaya et al. 2024]. O problema abordado é como capturar as características temporais das faixas de áudio para melhorar a precisão da classificação. O modelo foi aplicado nas bases de dados GTZAN e ISMIR2004, que contêm músicas de diferentes gêneros. A arquitetura BiLSTM foi usada para explorar as dependências temporais de ambas as direções nas faixas de áudio. O modelo conseguiu alcançar uma acurácia de 92,9% na base GTZAN e 95,2% na base ISMIR2004.

Utilizando redes convolucionais, têm-se trabalhos utilizando a base de dados GTZAN para realizar a tarefa de classificação à qual alcançou 77% de acurácia [Cheng et al. 2020]. Outro autor utilizou uma MCLNN (do inglês, *Masked Conditional Neural Networks*) para a classificação de sons, tanto musicais quanto ambientais [Medhat et al. 2020]. A MCLNN é projetada para melhorar a eficiência e a precisão da classificação de áudio, capturando melhor as características de tempo-frequência dos sinais de áudio, onde o autor conseguiu uma acurácia de 85,1% para a classificação de gêneros musicais com a base de dados GTZAN.

De maneira complementar, [Silva et al. 2021c] propõem uma abordagem baseada em CNN com fusão de múltiplas representações de áudio, como mel-espectrograma, MFCC, CQT, chroma e tempogram, combinadas por meio de estratégias de early e late fusion. O modelo alcançou até 88,5% de acurácia na base GTZAN, evidenciando que a combinação de diferentes características pode aumentar o desempenho na tarefa de classificação de gêneros musicais. Outro trabalho utilizou CNN, especificamente o modelo musicnn, para investigar o impacto de modelos generalistas e específicos ao domínio na tarefa MGR utilizando a base 4MuLA, onde os autores demonstraram que modelos especializados em músicas brasileiras, mesmo com menos dados, obtiveram melhor desempenho, com destaque para o modelo em português, que alcançou ROC-AUC de 0.89 e PR-AUC de 0,79 na classificação de músicas nacionais [Silva et al. 2021b].

Já para classificação de gêneros musicais com transformadores, um trabalho apresentou uma arquitetura baseada na arquitetura de transformadores, que incluem camadas de autoatenção *multi-head* e camadas totalmente conectadas [Zhuang et al. 2020]. Para o treinamento, o modelo utilizou 6 blocos de codificação e 2 cabeças na camada de autoatenção. A acurácia foi a principal métrica de avaliação utilizada no estudo. Após 10 épocas de treinamento, o modelo alcançou uma acurácia de 76,0% com a GTZAN.

Também utilizando transformadores, outros autores propõem um método de treinamento supervisionado utilizando o *Swin Transformer* para a classificação musical, tendo como objetivo aprender representações musicais significativas a partir de grandes quantidades de dados musicais não rotulados [Zhao et al. 2022]. O método utiliza um tipo de espectrograma (CQT-espectrograma) que oferece uma resolução de frequência variável, com maior precisão para as frequências mais baixas e menos precisão para as frequências mais altas. Os autores analisaram além da tarefa de classificação de gêneros musicais, a etiquetagem musical da música, que também é um campo de pesquisa MIR. Para a atividade de classificação de gêneros musicais, utilizando a base GTZAN, os autores conseguiram uma acurácia de 81,1%.

Por fim, um trabalho mais recente aborda o problema da classificação automática de gêneros musicais têm como principal proposta a combinação de uma arquitetura de Redes Neurais Convolucionais Residuais (*Res-Gated CNN*) com mecanismos de atenção para capturar características relevantes nas faixas musicais utilizando a base GTZAN [Wijaya et al. 2024]. A principal contribuição desse estudo foi o uso da atenção para focar em partes importantes dos espectrogramas gerados a partir do áudio, utilizando uma arquitetura de transformador com *encoder* e *decoder*, resultando em um modelo que conseguiu melhorar o desempenho da classificação em comparação com abordagens anteriores. O modelo apresentou uma acurácia de 96,8%, precisão de 89,4%, Recall 91,3%, e F1-Score 90,3%.

3.2.2. Bases de Dados

A Tabela 3.2 resume as principais bases de dados utilizadas em pesquisas de classificação de gêneros musicais, com a principal referência de cada base, destacando suas características e a disponibilidade de acesso. Cada uma dessas bases possui pontos fortes específicos e limitações que influenciam diretamente seu uso em estudos de MGR.

A ISMIR 2004 Genre Dataset é uma das primeiras bases de dados usadas para a classificação automática de gêneros musicais, contendo 1000 faixas de áudio distribuídas entre 10 gêneros, como Blues, Classical, Country, Disco, Hip-Hop, Jazz, Metal, Pop, Reggae, e Rock, sendo uma referência. No entanto, seu conteúdo limitado e a falta de músicas contemporâneas podem reduzir sua aplicabilidade em estudos modernos [ISMIR 2004]. Seguindo uma abordagem similar de grande escala, o Million Song Dataset tam-

Tabela 3.2. Tabela comparativa das principais bases de dados utilizadas em problemas de reconhecimento de gênero musical.

Tipo da Base	Base de Dados	Referência	Título	Gêneros Musicais	Número de Faixas	Duração da Faixa	Disponibilidade para Uso
Base Global	ISMIR 2004 Genre Dataset	ISMIR 2004	ISMIR 2004 Genre Classification Contest	Blues, Classical, Country, Disco, Hip-Hop, Jazz, Metal, Pop, Reggae, Rock	729	30s	Sim
	Million Song Dataset	Bertin-Mahieux et al. 2011	The Million Song Dataset	Diversos gêneros musicais	1M	-	
	GTZAN	Sturm 2013	The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use	Blues, Classical, Country, Disco, Hip-Hop, Jazz, Metal, Pop, Reggae, Rock	1.000	30s	
	FMA (Free Music Archive)	Defferrard et al. 2016	FMA: A dataset for music analysis	Diversos gêneros musicais	106.574	Variável	
	4MuLA	da Silva et al. 2020	4MuLA: A Multitask, Multimodal, and Multilingual Dataset of Music Lyrics and Audio Features	Diversos gêneros musicais	96.458	30s	
Base Regional	Latin Music Database (LMD)	Silla Jr et al. 2008	The Latin Music Database	Axé, Bachata, Bolero, Forró, Gaúcha, Merengue, Pagode, Salsa, Sertaneja, Tango	3.227	Variável	Não
	Base Iraniana	Abdoli 2011	Iranian Traditional Music Dastgah Classification	Dastgahs da música tradicional iraniana	210	1m	
	Greek Music Dataset	Makris et al. 2015	The Greek Music Dataset	Música tradicional e folclórica grega	1.000	Variável	
	Base Indonésia	Furqon et al. 2019	Indonesian Music Genre Classification on Indonesian Regional Songs	Gamelan Javanês, Gamelan Balinês, Kulintang Minahasa, Angklung Sundaês	1.000	30s	
Base Regional Brasileira	Base do Nordeste Brasileiro	Barbosa et al. 2015	Evaluating Automated Classification Techniques for Folk Music Genres from the Brazilian Northeast	Maracatu Solto, Maracatu Virado, Ciranda, Coco, Cavalo-Marinho	75	30s	Sim
	Base Brasileira	De Sousa et al. 2016	A Robust Music Genre Classification Approach for Global and Regional Music Datasets Evaluation	Forró, Rock, Repente, MPB(Música Popular Brasileira), Brega, Sertanejo e Disco.	120	30s	Não
	Base Brasileira	Conceição et al. 2020	Reconhecendo Gêneros Musicais Brasileiros com Técnicas de Aprendizagem de Máquina Supervisionada	Axé, Forró, Rock, MPB(Música Popular Brasileira), Samba e Sertanejo	613	30s, 40s, 1m, 2m	Parcial
	Base Brasileira	Silva et al. 2021a	Modelo Automático de Classificação de Gêneros Musicais Amazônicos	Andino, Brega, Carimbó, Cúmbia, Marabaixo, Merengue, Pasillo, Salsa, Vaqueirada	900	30s	Não

bém se destaca por oferecer uma vasta quantidade de metadados musicais, permitindo análises em larga escala sobre tendências e padrões na música popular. Entretanto, diferentemente da ISMIR, o Million Song Dataset não inclui as faixas de áudio, mas apenas seus metadados e recursos derivados, o que limita certos tipos de estudos que dependem de características acústicas detalhadas [Bertin-Mahieux et al. 2011].

A GTZAN dataset é uma das bases de dados mais reconhecidas na área de MGR. Contendo 1000 faixas de áudio divididas igualmente entre 10 gêneros musicais, esta base tem sido amplamente usada como baseline. A facilidade de acesso à GTZAN permite que pesquisadores em todo o mundo utilizem essa base para comparar seus resultados, promovendo a consistência nas avaliações [Sturm 2013]. Em contraste com a GTZAN, a FMA (Free Music Archive) oferece uma escala significativamente maior, contendo uma vasta diversidade de gêneros musicais. Este volume expansivo de dados permite uma análise mais abrangente e detalhada, embora exija recursos computacionais mais robustos. A riqueza de metadados e o acesso a faixas completas tornam a FMA particularmente útil para estudos que exploram padrões musicais em profundidade [Defferrard et al. 2016]. No entanto, a variabilidade na qualidade das gravações pode introduzir desafios adicionais, especialmente em análises que exigem uma uniformidade de dados.

Mais recentemente, a base de dados 4MuLA (Multitask, Multimodal, and Multilingual Dataset of Music Lyrics and Audio Features) tem ganhado destaque por sua abrangência e diversidade cultural. Composta por mais de 96 mil faixas distribuídas em 76 gêneros musicais, a base inclui músicas em português, inglês e espanhol, cobrindo artistas brasileiros e internacionais. Um diferencial importante da 4MuLA é a combinação de diferentes modalidades, como letras de músicas, metadados e características acústicas extraídas automaticamente, o que a torna adequada para tarefas multimodais de classificação, recomendação e análise semântica [da Silva et al. 2020]. No entanto, apesar de seu tamanho expressivo, a base fornece apenas trechos fixos de 30 segundos e não explora segmentações temporais refinadas ao longo das músicas, o que pode limitar aplicações mais detalhadas que envolvam análise de seções específicas das composições.

Abordando os trabalhos que propõem bases regionais, a Latin Music Database (LMD) se destaca por sua especialização em gêneros musicais latinos. Essa curadoria específica oferece uma representação culturalmente rica e diversa, essencial para estudos que buscam explorar a música latina em profundidade [Silla Jr et al. 2008]. Contudo, a

LMD enfrenta uma limitação significativa: sua indisponibilidade pública devido a restrições de direitos autorais. Esta barreira impede que a comunidade de pesquisa acesse amplamente a base, limitando a replicação de estudos e a validação externa. A base iraniana apresenta uma abordagem mais regional para o estudo dos modos musicais tradicionais iranianos [Abdoli 2011]. Embora essa base seja valiosa para a preservação e análise da música iraniana, a falta de informações sobre sua disponibilidade pública limita seu uso em contextos de pesquisa mais amplos. Ainda em uma linha de pesquisa cultural, a base de dados de músicas regionais da Indonésia [Furqon et al. 2019] e a base de músicas gregas [Makris et al. 2015], oferecem uma coleção de clipes musicais representativos regionais. Essas bases são ideais para estudos de diversidade cultural e musical, permitindo uma exploração mais profunda de tradições musicais regionais. No entanto, assim como outras bases regionais, a falta de clareza sobre sua disponibilidade pública pode ser um obstáculo significativo para sua utilização em pesquisas acadêmicas.

Complementando o foco em bases regionais, autores apresentam bases regionais brasileiras, como a pesquisa que apresenta uma base regional nordestina [Barbosa et al. 2015], empregando uma base de dados própria com 75 amostras de músicas, todas em formato de áudio com 30 segundos de duração extraídas do YouTube. As amostras foram convertidas em arquivos .WAV para garantir uma maior qualidade de análise, e estão disponibilizadas para uso no GitHub. De maneira similar, outros trabalhos apresentam bases regionais brasileiras [De Sousa et al. 2016, Conceição et al. 2020], explorando gêneros conhecidos por sua diversidade cultural. Similar aos anteriores, existe outro trabalho que foca em gêneros brasileiros, entretanto com um olhar para os gêneros característicos da Amazônia Legal [Silva et al. 2021a], o qual utiliza uma base própria, composta por músicas extraídas do YouTube.

3.2.3. Trabalhos com Estratégias de Segmentação Temporal e Redução de Dimensionalidade

A maior parte dos trabalhos relacionados utiliza apenas um trecho por faixa musical, geralmente extraído do início da música, o que limita a análise de variações temporais ao longo da estrutura musical. A Tabela 3.3 apresenta uma comparação entre esses estudos, destacando as estratégias de segmentação adotadas, as durações dos segmentos analisados e a ausência, na maioria dos casos, de uma consideração explícita da posição temporal.

Tabela 3.3. Comparação entre os trabalhos da literatura com foco em estratégias de segmentação, duração dos trechos analisados e posicionamento temporal.

Trabalho	Estratégia de Segmentação	Duração do Segmento	Técnica	Duração da Faixa	Posição Temporal	Base de Dados
Barbosa et al. 2015	Fixa	30s	SVM, KNN, outros	30s	✗	Brasileira
De Sousa et al. 2016			SVM		✗	GTZAN + Brasileira
Cheng et al. 2020			CNN		✗	GTZAN
Medhat et al. 2020			MCNN		✗	GTZAN, ISMIR
Wijaya et al. 2024			BiLSTM		✗	GTZAN, ISMIR
Zhuang et al. 2020	Curta	3s	Transformer	30s	✗	GTZAN
Xie et al. 2024		5s	CNN + Transformer		✗	GTZAN
Conceição et al. 2020	Variável	30s, 40s, 60s, 120s	KNN, SVM, outros	30s, 40s, 60s, 120s	✓	Brasileira
Dai et al. 2016	Não especificada	Não mencionada	LSTM	Faixa completa	✗	ISMIR
Silva et al. 2021a			KNN, SVM, outros	Faixa completa	✗	Brasileira
Zhao et al. 2022			Swin Transformer	30s	✗	GTZAN, FMA

Alguns estudos optam por utilizar a faixa musical completa ou segmentos cuja posição não é claramente especificada. É o caso dos trabalhos de [Dai et al. 2016], [Silva et al. 2021a] e [Zhao et al. 2022], nos quais não há uma exploração sistemática da posição temporal dos trechos utilizados na classificação. Essa ausência de controle sobre o posicionamento dificulta a análise de como diferentes partes da música influenciam a tarefa de reconhecimento de gênero. Outros trabalhos adotam uma abordagem mais comum na literatura, que consiste no uso de segmentos fixos de trinta segundos. Esses trechos são geralmente extraídos do início da faixa, como observado em [Barbosa et al. 2015], [De Sousa et al. 2016], [Cheng et al. 2020], [Medhat et al. 2020] e [Wijaya et al. 2024]. Embora essa estratégia ofereça certa padronização, ela ignora a diversidade estrutural presente ao longo da música.

Além disso, algumas pesquisas exploram o uso de segmentos curtos, com duração entre três e cinco segundos, como nos estudos de [Zhuang et al. 2020] e [Xie et al. 2024]. No entanto, mesmo nesses casos, a posição dos segmentos ao longo da faixa permanece fixa, sem considerar diferentes momentos temporais da música. Dessa forma, ainda há uma limitação importante quanto à análise de variações representativas ao longo da estrutura musical. Entre os estudos analisados, apenas [Conceição et al. 2020] adota uma abordagem mais abrangente, avaliando múltiplas durações e posições dos trechos musicais, com foco em gêneros brasileiros combinados à base GTZAN.

A Tabela 3.4 compara os principais trabalhos e técnicas de similaridade aplica-

das à análise de *embeddings* musicais ou não. A coluna MGR indica se o estudo aborda reconhecimento de gênero musical. Similaridade mostra se há análise explícita de proximidade entre representações. Modelo descreve a arquitetura utilizada, enquanto Redução de Dimensionalidade informa se foram aplicadas técnicas para visualização dos *embeddings*. Métrica de Similaridade especifica a métrica empregada na comparação vetorial, e Base de Dados apresenta os conjuntos utilizados em cada estudo.

Tabela 3.4. Comparação entre os trabalhos relacionados quanto ao uso de similaridade e técnica adotada.

Trabalho	MGR	Similaridade?	Modelo	Redução de Dim.	Métrica Similaridade	Base de Dados
Armijo and Rivas 2025	✗	✓	Vision Transformer	UMAP	Agrupamento Visual	Craigslist e OfferUp
Barnett et al. 2024		✓	Jukebox (OpenAI)	PCA, t-SNE	Euclidiana e Cosseno	Million Song Dataset, Spotify
Yonay et al. 2025		✗	Vision Transformer	✗	✗	AudioSet, GTZAN, MTT, EmoMusic
Sheikh Fathollahi and Razzazi 2021	✓	✓	CNN	✗	Euclidiana e Cosseno	GTZAN, Emotify
Spijkervet and Burgoyne 2021		✓	CNN	✗	Cosseno	MagnaTagATune
Shu et al. 2024		✗	CNN, Vision Transformer	PCA, t-SNE, UMAP	Agrupamento Visual	EDM Benchmark
Luo et al. 2024		✓	CLMR	UMAP, PCA, t-SNE	Cosseno	GTZAN, MagnaTagATune, outros

Embora alguns desses estudos não estejam diretamente voltados para o reconhecimento de gêneros, eles oferecem contribuições relevantes ao analisarem visualmente os agrupamentos ou calcularem métricas de proximidade entre representações musicais. O trabalho de [Armijo and Rivas 2025], por exemplo, utiliza Vision Transformers e UMAP para visualizar agrupamentos de *embeddings*, com foco na inspeção qualitativa, sem aplicar diretamente métricas como distância cosseno ou euclidiana. Já [Barnett et al. 2024] investigam relações estilísticas entre artistas a partir de *embeddings* extraídos do modelo Jukebox, comparando-os por meio de distâncias vetoriais e técnicas como PCA e t-SNE. Outro estudo relevante é o de [Yonay et al. 2025], que adota um modelo baseado em Vision Transformers treinado com aprendizado contrastivo sobre espectrogramas, focando na construção de representações musicais gerais, avaliadas posteriormente em bases como GTZAN e MagnaTagATune.

Entre os trabalhos que lidam diretamente com classificação de gêneros, observa-se diversidade nas estratégias adotadas para o uso de similaridade. O estudo de [Sheikh Fathollahi and Razzazi 2021] propõe um classificador baseado em CNNs e utiliza as representações intermediárias para construir um sistema de recomendação musical, com comparação vetorial via distância cosseno e euclidiana. Já [Spijkervet and Burgoyne 2021] treinam o modelo CLMR com aprendizado contrastivo e aplicam o reconhecimento de gênero sobre a base MagnaTagATune, empregando uma função de perda baseada em si-

milaridade cosseno. O trabalho de [Shu et al. 2024] propõe um benchmark de subgêneros de EDM, combinando CNNs e Vision Transformers, e utiliza PCA, t-SNE e UMAP para visualização das separações entre os gêneros, ainda que não calcule diretamente a similaridade. Por fim, [Luo et al. 2024] combina as representações do CLMR com mecanismos interativos de busca musical via similaridade cosseno, integrando técnicas de redução de dimensionalidade e interfaces visuais para exploração por gênero.

As Tabelas 3.3 e 3.4 evidenciam que poucos estudos integram análise temporal, representações aprendidas e medidas de similaridade. A abordagem adotada nesta dissertação, apresentada na próxima seção, busca justamente preencher essa lacuna.

3.2.4. MGR Utilizando Arquitetura de Transformadores

Para lidar com a dificuldade de classificação de gêneros musicais similares, este trabalho reformula a tarefa de classificação de áudio como um problema de visão computacional, utilizando espectrogramas na escala Mel como entrada, como utilizado na literatura de acordo com a Tabela 3.1. Essa representação permite capturar padrões espectrais e temporais relevantes, possibilitando o uso de arquiteturas originalmente projetadas para tarefas visuais, como os transformadores, que vêm demonstrando grande eficácia na modelagem de padrões complexos e dependências de longo alcance.

Embora transformadores tenham se destacado em tarefas de linguagem e áudio, seu uso ainda é pouco explorado no reconhecimento de gêneros musicais regionais, como evidenciado na Tabela 3.1. No que diz respeito às bases de dados utilizadas em tarefas de reconhecimento de gênero, a Tabela 3.2 mostra que grande parte dos conjuntos adotados na literatura é composta por trechos fixos de 30 segundos, muitas vezes extraídos apenas do início da faixa. Além disso, diversos desses conjuntos não estão publicamente disponíveis, o que dificulta a reprodutibilidade dos experimentos. A Tabela 3.3 reforça essas limitações ao evidenciar que a maioria dos trabalhos utiliza apenas um único segmento por faixa, com pouca ou nenhuma consideração da posição temporal ao longo da música. Mesmo abordagens que utilizam janelas curtas [Zhuang et al. 2020, Xie et al. 2024] não exploram a posição dos segmentos dentro da faixa, dificultando a identificação dos trechos mais representativos de cada gênero.

Outro ponto pouco abordado na literatura está relacionado à análise de similaridade entre representações musicais. A Tabela 3.4 mostra que poucos estudos combinam

o uso de *embeddings* com técnicas de visualização e métricas de similaridade para investigar a proximidade entre estilos musicais. Mesmo quando modelos como CNNs ou Vision Transformers são utilizados, as representações vetoriais geralmente são exploradas apenas para fins classificatórios, sem uma análise mais aprofundada sobre a organização interna dos gêneros no espaço latente aprendido.

Neste contexto, propõe-se a criação de uma base de dados composta por gêneros musicais regionais brasileiros, caracterizados por elevada similaridade estilística. Os dados são extraídos a partir de segmentos com diferentes durações (3s, 5s e 10s) e em distintas posições temporais dentro das faixas musicais, permitindo investigar a influência da estrutura musical na percepção e classificação dos gêneros. A avaliação envolve tanto métricas quantitativas de desempenho quanto análises qualitativas das representações internas, explorando a organização dos estilos no espaço vetorial por meio de técnicas como t-SNE, PCA e UMAP.

Dessa forma, esta dissertação propõe uma nova abordagem ao combinar segmentação temporal ao longo da música, múltiplas durações por segmento, análise qualitativa dos *embeddings* e uso de arquiteturas de transformadores no reconhecimento de gêneros musicais regionais brasileiros. Além disso, o trabalho disponibiliza uma base de dados representativa voltada a gêneros regionais brasileiros, contribuindo para a padronização de experimentos e para o desenvolvimento de pesquisas comparáveis na área.

Capítulo 4

Metodologia

Este capítulo discorre sobre os procedimentos metodológicos e experimentais realizados nesta pesquisa.

4.1. Visão Geral

A presente pesquisa é de **natureza aplicada**, com foco no desenvolvimento e avaliação de um modelo de aprendizado profundo baseado em Vision Transformer para a classificação de gêneros musicais regionais brasileiros. Adota-se uma abordagem **quali-quantitativa**, que combina métodos **quantitativos**, voltados à medição do desempenho do modelo em cenários controlados, com métodos **qualitativos**, voltados à análise interpretativa dos resultados obtidos e das relações de similaridade entre os gêneros. Para isso, foi realizado um **estudo de caso** que utiliza um conjunto de gêneros musicais representativos de diversas regiões do Brasil como objeto experimental. O objetivo central é investigar se uma rede do tipo Vision Transformer é capaz de capturar características estilísticas e estruturais que permitam não apenas classificar corretamente os gêneros musicais, mas também refletir suas relações de similaridade no espaço latente.

Para responder às questões de pesquisa deste trabalho, foi desenvolvida a base de dados **BYRM (Brazilian YouTube Regional Music)** [Guimarães et al. 2025a], composta por 1.082 faixas musicais de dez gêneros brasileiros, com trechos extraídos de diferentes partes das músicas. Cada faixa foi segmentada em blocos de 30 segundos, permitindo a análise temporal ao longo da música. Esses blocos foram posteriormente divididos em segmentos menores de 3, 5 e 10 segundos, com sobreposição de 50%, a fim de investigar

o impacto da duração do segmento no desempenho do modelo.

A metodologia contempla diferentes fases experimentais: (1) criação da base de dados BYRM, (2) identificação dos blocos mais e menos representativos das músicas por meio de testes com o Vision Transformer, (3) análise do impacto da duração dos segmentos nesses blocos, (4) comparação do desempenho entre modelos (Vision Transformer, ResNet50 e SVM), (5) análise da organização dos gêneros no espaço de embeddings aprendido e (6) cálculo da similaridade entre os embeddings presentes no espaço vetorial original. A estrutura metodológica adotada segue uma adaptação do modelo CRISP-DM (Cross Industry Standard Process for Data Mining) [Wirth and Hipp 2000], cobrindo as fases de compreensão dos dados, preparação, modelagem, avaliação e interpretação. Essa abordagem permite conduzir experimentos reproduzíveis e bem fundamentados, proporcionando uma compreensão aprofundada do desempenho e das limitações dos modelos de classificação aplicados à música regional brasileira. Cada uma das etapas descritas será detalhada nas seções seguintes.

4.2. Experimento Inicial

Antes da consolidação da metodologia apresentada nessa dissertação, foi conduzido um experimento exploratório utilizando uma versão inicial da base BYRM combinada com faixas da base pública GTZAN. Essa combinação foi necessária porque a BYRM era composta por apenas cinco gêneros musicais, enquanto os outros dez foram complementados pela GTZAN, de forma a totalizar quinze classes de gêneros musicais. O objetivo dessa etapa inicial era validar a viabilidade da arquitetura Vision Transformer (ViT) para a tarefa de classificação de gêneros musicais.

Durante essa etapa preliminar, a combinação da base BYRM com a GTZAN permitiu realizar experimentos exploratórios com a arquitetura Vision Transformer (ViT), cujo objetivo era verificar o potencial do modelo para a tarefa de classificação de gêneros musicais. No entanto, ao longo do desenvolvimento da pesquisa, foi identificado que essa abordagem poderia ser aprimorada em diversos aspectos, de modo a tornar os resultados mais robustos e representativos ao contexto brasileiro.

Entre os fatores que motivaram a reformulação da metodologia, destacam-se:

1. A ausência de técnicas de regularização, como *dropout* e *early stopping*, resultando em maior tendência ao *overfitting*.

2. A impossibilidade de realizar uma comparação justa com trabalhos anteriores que utilizaram a base GTZAN, devido à diferença nos critérios de separação dos dados e no pré-processamento das instâncias.
3. A presença de trechos da mesma faixa musical em diferentes divisões (treino, validação e teste), o que poderia comprometer a avaliação da capacidade de generalização do modelo.

Diante disso, optou-se por reformular a metodologia, construindo a versão definitiva da base BYRM, composta exclusivamente por gêneros regionais brasileiros representativos. Além disso, foi adotado um novo protocolo experimental, cujo foco passou a abranger não somente a classificação dos gêneros musicais, mas também uma análise quali-quantitativa da similaridade entre gêneros e uma investigação qualitativa da evolução temporal das suas características.

4.3. Base de Dados BYRM

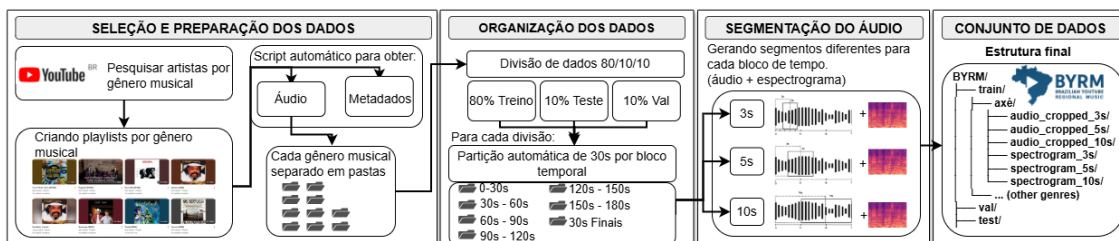
A base de dados **BYRM (Brazilian YouTube Regional Music)**¹ foi construída a partir da adaptação das etapas do processo de Descoberta de Conhecimento em Bases de Dados (KDD) [Fayyad 1997], dado o seu enfoque em extração de conhecimento a partir de dados brutos. Foram adaptadas as quatro fases do KDD: seleção e pré-processamento, organização, transformação e apresentação. Cada uma dessas fases será detalhada ao longo desta seção. A BYRM reúne dez gêneros musicais representativos de diferentes regiões do Brasil: toada, carimbó, axé, forró, sertanejo, rock brasileiro, samba, pagode, vaneira e xote gaúcho. A Figura 4.1 mostra a distribuição geográfica desses gêneros, reforçando o compromisso com a diversidade cultural do país.



Figura 4.1. Gêneros musicais regionais brasileiros que integram a base de dados BYRM por região característica.

¹Disponível em: <https://zenodo.org/records/16617888>

A base foi criada com o objetivo de viabilizar estudos sobre a classificação automática de gêneros musicais brasileiros, levando em consideração aspectos temporais e estilísticos das faixas. A Figura 4.2 apresenta o pipeline de construção da base, desde a coleta das músicas até a geração das segmentações finais utilizadas nos experimentos.



4. **Versão Final da Base:** Todas as versões mantêm a divisão original entre treino, validação e teste. Os segmentos estão organizados por gênero e por duração temporal, incluindo tanto os clipes de áudio quanto os espectrogramas correspondentes, permitindo diferentes estratégias de treinamento e análise.

A base BYRM é composta por 1.082 faixas distribuídas entre dez gêneros musicais brasileiros: Axé (101), Carimbó (103), Forró (104), Pagode (102), Rock Brasileiro (107), Samba (103), Sertanejo (104), Toada (147), Vaneira (107) e Xote Gaúcho (104). A Tabela 4.1 apresenta os artistas representativos de cada gênero presente na base, destacando a diversidade cultural e estilística do conjunto de dados.

Tabela 4.1. Principais artistas representativos por gênero musical no BYRM.

Gênero	Artistas Principais
Axé	Babado Novo; Bell Marques; Carlinhos Brown; Claudia Leitte; Daniela Mercury; Harmonia do Samba; Ivete Sangalo; Jammil e Uma Noites.
Sertanejo	Chitãozinho & Xororó; Zezé Di Camargo & Luciano; Bruno & Marrone; Rick & Renner; Rionegro & Solimões; Victor & Leo; Almir Sater.
Toada	Tadeu Garcia; David Assayag; Arlindo Júnior; Prince do Caprichoso; Sebastião Júnior.
Vanera	Irmãos Bertussi; José Mendes.
Xote Gaúcho	Teixeirinha; Os Filhos do Rio Grande; Bando Gaúcho; Embalo Nativo; Os Monarcas; Gaúcho da Fronteira; Os Mirins; Os Serranos.
Carimbó	Mestre Verequete; Fruta Quente; Dona Onete; Mestre Cupijó; Lucas Estrela; Nilson Chaves; Alípio Martins; Pinduca.
Forró	Trio Virgulino; Luiz Gonzaga; Santanna O Cantador; Dominginhos; Falamansa; Os Gonzagas; Dorgival Dantas; Trio Forrozão Oficial.
Pagode	Sorriso Maroto; Turma do Pagode; Exaltasamba; Dilsinho; Grupo Soweto; Grupo Pixote; Thiaguinho; Ferrugem.
Rock BR	Legião Urbana; Raimundos; Plebe Rude; Capital Inicial; Barão Vermelho; Os Mutantes.
Samba	Jorge Aragão; Beth Carvalho; Alcione; Martinho da Vila; Zeca Pagodinho; Xande de Pilares; Dona Ivone Lara; Arlindo Cruz.

As durações das faixas variam, com uma média de aproximadamente 207,77 segundos. Para manter a consistência na segmentação, apenas os primeiros 180 segundos de cada faixa foram utilizados, possibilitando a extração de múltiplos trechos de 30 se-

gundos para análise temporal. A Tabela 4.2 apresenta a quantidade de segmentos gerados por bloco temporal, considerando janelas de 3, 5 e 10 segundos com sobreposição de 50%. Essa abordagem visa capturar variações rítmicas, harmônicas e tímbricas em diferentes resoluções temporais, ao mesmo tempo em que garante uma cobertura equilibrada ao longo das músicas.

Tabela 4.2. Quantidade de segmentos gerados por duração e por bloco temporal.

Duração	Bloco Temporal						
	0–30	30–60	60–90	90–120	120–150	150–180	30s Final
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

Dessa forma, a base BYRM foi organizada para permitir a análise da classificação de gêneros musicais em diferentes trechos e segmentos temporais das faixas, proporcionando uma estrutura adequada para observar como os embeddings se agrupam ao longo do tempo e, também, para identificar quais partes da música e quais durações de segmento concentram as características mais representativas de cada gênero. Devido às limitações legais relacionadas aos direitos autorais das faixas musicais coletadas do YouTube, a base BYRM não poderá disponibilizar publicamente os áudios originais. No entanto, para garantir a reprodutibilidade dos experimentos, foram disponibilizados os espectrogramas gerados em formato png e os vetores de características extraídas de cada faixa, além dos metadados das músicas incluindo o ID do vídeo utilizado para reprodução.

4.4. Preparação dos Dados

Os dados de áudio utilizados na pesquisa passaram por um processo de pré-processamento com o objetivo de garantir que as características musicais relevantes fossem preservadas e devidamente capturadas. Cada faixa de áudio foi inicialmente segmentada em trechos de 30 segundos e, em seguida, dividida em clipes menores de 3, 5 e 10 segundos, com uma sobreposição de 50% entre os segmentos. Esse procedimento permite transformar cada trecho de 30 segundos em 19 clipes de 3 segundos, 11 clipes de 5 segundos e 5 clipes de 10 segundos, aproximadamente.

A escolha dessas janelas temporais foi definida empiricamente, com o intuito de

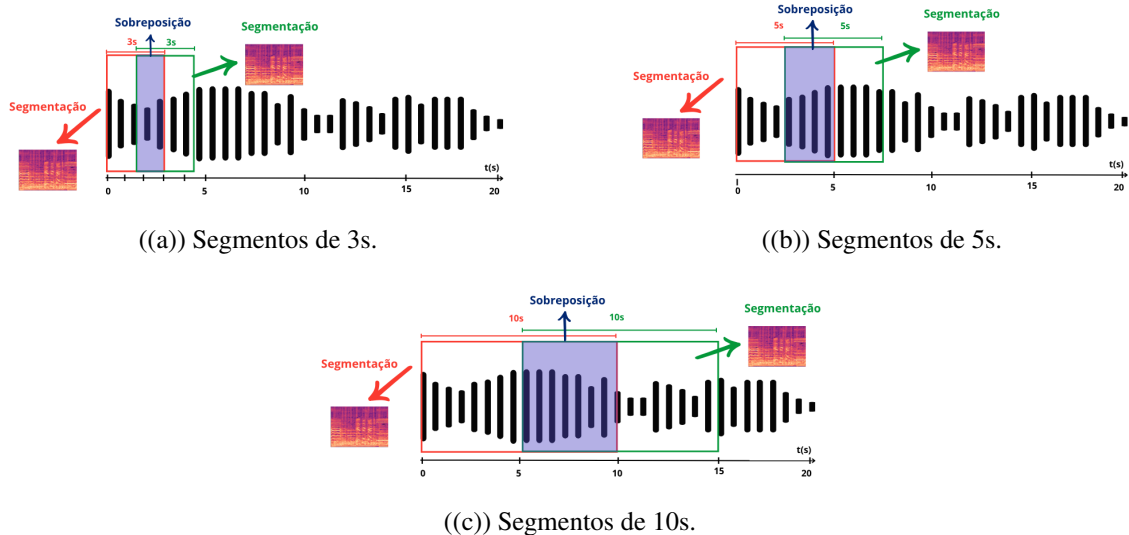


Figura 4.3. Representação da segmentação e sobreposição de clipes de áudio com diferentes durações.

capturar características musicais essenciais em diferentes resoluções temporais, considerando que as faixas podem apresentar variações rítmicas e tímbricas ao longo do tempo. Essa segmentação aumentou significativamente o número de amostras disponíveis para os conjuntos de treinamento, validação e teste. Por exemplo, considerando que cada classe possui 100 faixas de áudio, o processo gera aproximadamente 1.900 clipes de 3 segundos por classe, além de 1.100 clipes de 5 segundos e 500 clipes de 10 segundos.

A sobreposição de 50% garante que cada clipe capture informações presentes tanto no segmento anterior quanto no subsequente, minimizando a perda de conteúdo relevante devido à segmentação. Essa abordagem permite que o modelo analise diferentes partes da música de forma detalhada, reforçando a capacidade de identificação de padrões musicais mesmo em curtas janelas temporais. Após a segmentação, cada clipe de áudio foi transformado em um espectrograma Mel, que é uma representação visual das frequências do som ao longo do tempo, baseada na escala Mel. A Figura 4.3 ilustra o processo de segmentação de um trecho de 30 segundos em janelas de diferentes durações, evidenciando como a estrutura de entrada é preparada para a geração dos espectrogramas, garantindo a preservação das características musicais durante a análise.

O processo de conversão do áudio para mel espectrograma foi realizado utilizando uma transformada de Fourier de curto prazo (*Short-Time Fourier Transform - STFT*), com parâmetros específicos para garantir uma análise precisa das frequências ao longo do

tempo, com tamanho da janela de 2048 samples, deslocamento da janela de 512 samples, 128 bandas de mel, e uma taxa de amostragem de 22.050 Hz. Esses parâmetros foram escolhidos para garantir que as nuances e os detalhes musicais fossem capturados de forma eficaz, permitindo ao modelo de aprendizado profundo explorar as características visuais do espectrograma, como padrões de frequência e variações temporais, para realizar a classificação dos gêneros musicais com alta precisão. Por fim, a Figura 4.4 apresenta exemplos de instâncias de 3 segundos extraídas da base BYRM, representando um espectrograma Mel de cada um dos dez gêneros musicais. Esses exemplos foram obtidos a partir de trechos centrais das faixas, já segmentados e processados segundo os parâmetros descritos anteriormente.

Visualmente, nota-se que muitos espectrogramas compartilham padrões semelhantes de densidade espectral e distribuição de frequências, o que reforça o desafio da tarefa de classificação automática, especialmente em contextos de gêneros similares. Essa semelhança visual entre diferentes classes destaca a importância do uso de modelos capazes de capturar nuances tímbricas e estruturais sutis.

4.5. Divisão dos Dados

A divisão dos dados foi organizada com base nas segmentações descritas anteriormente na Seção 4.4. O particionamento dos dados em treino, validação e teste foi realizado no momento da criação da base BYRM, garantindo que todas as variações temporais fossem derivadas de faixas previamente atribuídas a um único grupo. Essa decisão garante a reprodutibilidade dos experimentos e evita vazamento de dados entre conjuntos, especialmente considerando a extração subsequente de múltiplos segmentos por faixa.

Como as faixas da base BYRM possuem durações distintas, nem todas geram instâncias em todos os blocos temporais. Faixas mais curtas, por exemplo, não alcançam os trechos finais da música. Além disso, cada versão da base, com janelas de 3, 5 e 10 segundos, gera um número diferente de segmentos por faixa, devido à variação na duração das janelas e à sobreposição de 50%. Para evitar vazamento de informação entre os conjuntos, foi garantido que todos os segmentos originados de uma mesma faixa fossem alocados no mesmo conjunto de treinamento, validação ou teste. Essa estratégia representa um cuidado metodológico fundamental para assegurar a confiabilidade dos resultados. A Tabela 4.3 apresenta a distribuição completa dos dados por bloco temporal e

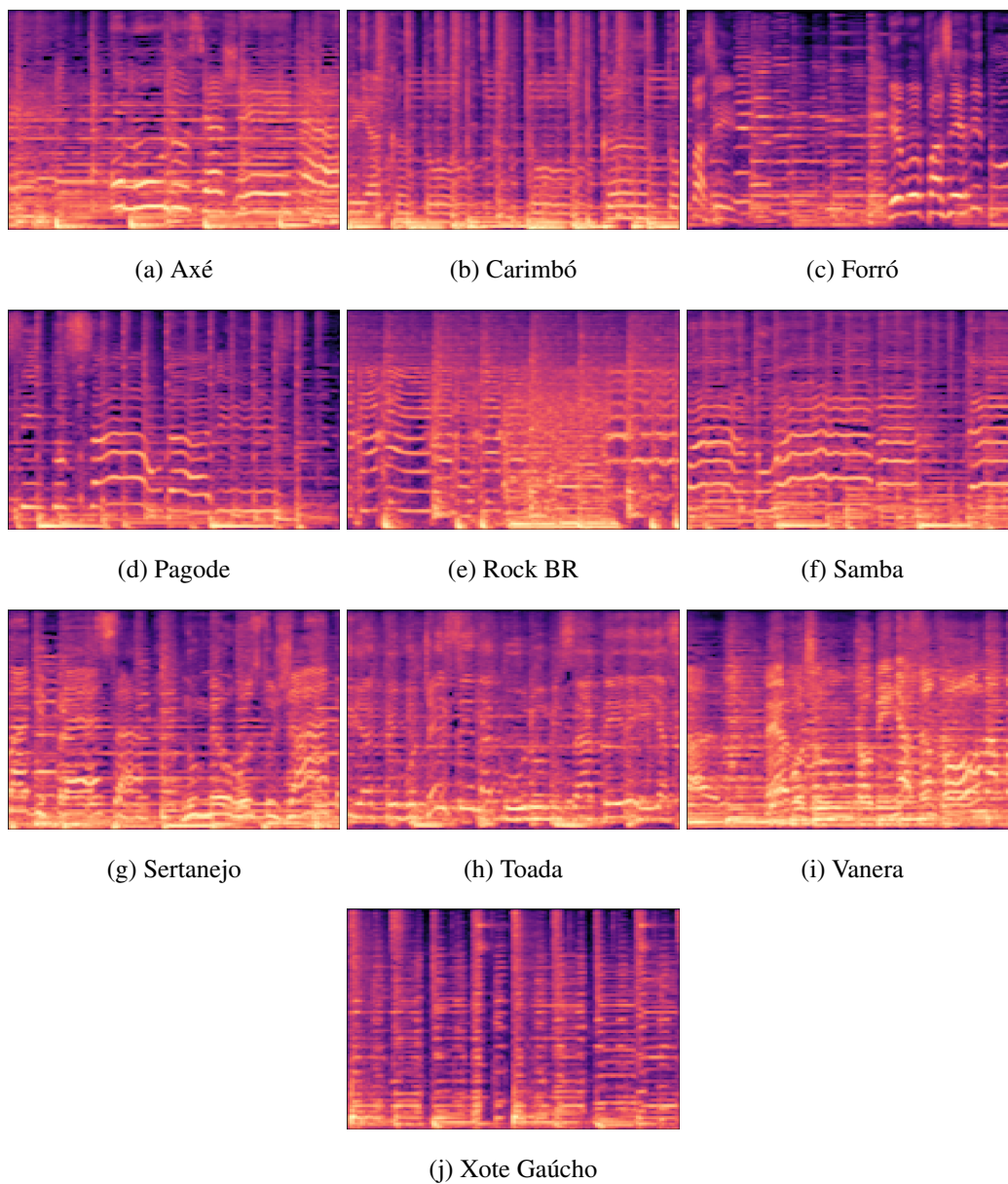


Figura 4.4. Exemplos visuais dos gêneros musicais presentes na base BYRM.

por duração de segmento.

Tabela 4.3. Divisão dos dados por bloco temporal e duração dos segmentos.

Bloco Temporal	Treinamento			Validação			Teste			Total		
	3s	5s	10s	3s	5s	10s	3s	5s	10s	3s	5s	10s
0–30	16.359	9.471	4.305	1.976	1.144	520	2.223	1.287	585	20.558	11.902	5410
30–60	16.359	9.471	4.305	1.976	1.144	520	2.223	1.287	585	20.558	11.902	5410
60–90	16.314	9.443	4.291	1.976	1.144	520	2.221	1.285	584	20.511	11.872	5395
90–120	16.206	9.377	4.259	1.961	1.134	514	2.191	1.268	576	20.358	11.779	5349
120–150	15.581	8.998	4.061	1.850	1.067	482	2.077	1.199	541	19.508	11.264	5084
150–180	13.386	7.698	3.437	1.590	918	409	1.739	999	442	16.715	9.615	4.288
30 Finais	16.359	9.471	4.305	1.976	1.144	520	2.223	1.287	585	20.558	11.902	5410

É possível observar que os blocos iniciais (como 0–30 e 30–60 segundos) mantêm um volume constante e elevado de amostras, enquanto blocos posteriores, como 150–180 segundos, apresentam redução expressiva na quantidade de dados. Esse padrão reflete a limitação da duração de parte das faixas, o que restringe a presença nos trechos finais. Além disso, nota-se que segmentos mais longos (como os de 10 segundos) geram naturalmente menos amostras por faixa em comparação aos segmentos de 3 segundos, o que impacta diretamente o tamanho dos conjuntos de dados utilizados nos experimentos. Esses fatores devem ser considerados na análise dos resultados, uma vez que influenciam tanto o volume de treino quanto a variabilidade das amostras disponíveis por trecho e por duração.

4.6. Configuração dos Experimentos

Os experimentos foram estruturados para investigar como diferentes posições e durações de segmentos afetam o desempenho na classificação de gêneros musicais. Para isso, a base BYRM foi utilizada em diferentes configurações de entrada. Inicialmente, todos os trechos de 30 segundos extraídos da base foram divididos em segmentos de 3 segundos com sobreposição de 50%. O modelo Vision Transformer (ViT), principal foco desta pesquisa, foi treinado separadamente em cada bloco temporal (por exemplo, 0–30s, 30–60s, ..., 150–180s) com esses segmentos, a fim de identificar os trechos mais e menos representativos em termos de desempenho de classificação.

Na etapa seguinte, o ViT foi treinado exclusivamente nos trechos mais e menos representativos identificados, utilizando segmentos de 3, 5 e 10 segundos, com o obje-

tivo de investigar o impacto da duração do segmento na capacidade discriminativa do modelo. Os modelos ResNet50 e SVM também foram avaliados nesses dois blocos, mas apenas com segmentos de 3 segundos, para possibilitar uma comparação consistente entre abordagens. Os detalhes técnicos do treinamento de cada modelo, incluindo otimização, regularização e configuração das janelas temporais, são apresentados na Seção 4.7.

Além da tarefa de classificação, foi realizada uma análise qualitativa aprofundada dos embeddings gerados pelo ViT em todos os experimentos conduzidos, com o objetivo de investigar a organização dos gêneros musicais no espaço vetorial aprendido. O modelo SVM não foi incluído nessa etapa, por não gerar embeddings internos. Também foi conduzida uma análise quantitativa de similaridade entre os gêneros, com base na similaridade cosseno entre os embeddings. Os resultados foram organizados em uma matriz de similaridade e visualizados por meio de um mapa de calor, possibilitando avaliar a proximidade e a distinção entre os estilos musicais regionais brasileiros no espaço vetorial aprendido.

4.7. Modelos de Classificação Utilizados

A Tabela 4.4 apresenta um resumo das principais configurações de treinamento adotadas para os três modelos avaliados neste trabalho: Vision Transformer (ViT), ResNet50 e Support Vector Machine (SVM). Cada modelo representa uma abordagem distinta na tarefa de classificação de gêneros musicais: modelos baseados em transformadores, redes convolucionais profundas e métodos clássicos com extração manual de características.

Tabela 4.4. Resumo das configurações dos modelos utilizadas nos experimentos.

Aspecto	ViT	ResNet50	SVM
Pré-treinamento	✓ (ImageNet)	✓ (ImageNet)	✗
Entrada	Mel-espectrograma	Mel-espectrograma	Características manuais
Segmentos utilizados	3s, 5s, 10s	3s	3s
Trechos utilizados	Todos	Melhor e pior	Melhor e pior
Características	Espectrograma	Espectrograma	MFCC, cromograma, centróide espectral, largura de banda, rolloff, ZCR
Épocas milestones	10/30/40	15/100/150	✗
Dropout	30% na cabeça, 7% interno	Padrão	✗
Épocas	100	200	✗
Paciência	10	15	✗
Otimização	Adam (lr = 1e-4)	Adam (lr = 1e-4)	✗
Decaimento do LR	0.8 → 0.6 → 0.3	0.8 → 0.6 → 0.3	✗

A partir das configurações resumidas na Tabela 4.4, observa-se que cada modelo adotou estratégias distintas de entrada, pré-processamento e treinamento. Para garantir a reprodutibilidade dos experimentos, foi utilizada a seed fixa 42 na divisão dos dados durante a validação cruzada aplicada ao modelo SVM. Nos modelos ViT e ResNet50, a mesma seed foi utilizada para controlar os elementos de aleatoriedade presentes no treinamento, como a inicialização dos pesos e o uso de *dropout*. No caso do SVM, sua configuração não incorpora aleatoriedade interna, o que assegura comportamento determinístico mesmo sem a definição explícita da seed no classificador. Ainda assim, reconhece-se que não foram realizadas múltiplas execuções com diferentes seeds, tampouco foi conduzida uma análise estatística formal dos resultados, o que representa uma limitação a ser abordada em trabalhos futuros. Para garantir clareza e reprodutibilidade, os procedimentos específicos de preparação e ajuste de cada modelo serão apresentados de forma detalhada nas subseções a seguir.

4.7.1. Procedimento de Treinamento SVM

O modelo Support Vector Machine (SVM) foi utilizado como *baseline* clássico para fins comparativos, operando sobre características acústicas extraídas diretamente dos cliques de áudio de 3 segundos. A extração foi realizada com a biblioteca *librosa*, e as features computadas incluíram MFCCs, cromograma, centróide espectral, largura de banda espectral, *rolloff* e taxa de cruzamento por zero (ZCR), resultando em um vetor representativo de cada segmento.

Os vetores de características foram normalizados com a técnica de padronização z-score, utilizando a classe *StandardScaler* da biblioteca *scikit-learn*, com o objetivo de garantir que todas as variáveis tivessem média zero e desvio padrão unitário, evitando o viés de escala durante o treinamento da SVM.

A SVM foi configurada com kernel radial (RBF), sendo os hiperparâmetros C e γ ajustados por meio de busca em grade (*GridSearchCV*) com validação cruzada estratificada de 5 folds. O modelo foi treinado apenas nos trechos de melhor e pior desempenho da base BYRM, previamente identificados com o modelo ViT. Por não se tratar de um modelo de aprendizado profundo, o SVM não realiza ajuste de pesos ao longo de épocas nem emprega técnicas como *dropout* ou *fine-tuning*.

4.7.2. Procedimento de Treinamento ResNet50

A ResNet50 foi empregada como modelo *baseline* de arquitetura convolucional profunda, utilizando pesos pré-treinados na base ImageNet. A camada de saída foi substituída por uma camada densa com dez unidades, correspondente ao número de classes da base BYRM. Os espectrogramas Mel foram redimensionados para 224×224 pixels e normalizados para atender ao padrão de entrada da ResNet, mantendo consistência com o pré-processamento adotado no ViT. O modelo foi treinado exclusivamente com segmentos de 3 segundos nos trechos previamente identificados como de melhor e pior desempenho da base. Para preservar o conhecimento aprendido no pré-treinamento e permitir uma adaptação gradual ao novo domínio, foi adotada uma política de descongelamento progressivo das camadas convolucionais.

Inicialmente, todas as camadas foram congeladas, com exceção da nova camada de saída. A partir de marcos definidos (épocas 15, 100 e 150), as camadas da rede foram descongeladas, iniciando pela *layer4* e prosseguindo até a *layer1*. A cada marco, também foi aplicada uma redução na taxa de aprendizado, seguindo os fatores de decaimento previamente definidos. O otimizador utilizado foi o Adam, com taxa inicial de 1×10^{-4} , e a função de perda foi a *Cross-Entropy Loss*.

Durante o treinamento, as estatísticas internas das camadas de normalização em lote (*batch normalization*) foram mantidas fixas nas camadas ainda congeladas, por meio de uma rotina que colocava essas camadas em modo de avaliação. A cada época, foram extraídos os embeddings da penúltima camada da rede para análise exploratória da organização dos gêneros no espaço vetorial, utilizando técnicas de redução de dimensionalidade como PCA, t-SNE e UMAP. Ao final do treinamento, o modelo com menor perda de validação foi selecionado para avaliação no conjunto de teste. A estratégia incluiu *early stopping* com paciência de 15 épocas para evitar sobreajuste.

4.7.3. Procedimento de Treinamento ViT

Todos os experimentos com o modelo Vision Transformer (ViT) foram conduzidos de forma padronizada e reproduzível em todas as configurações da base BYRM. Isso inclui tanto os diferentes blocos temporais (como 0–30s, 90–120s etc.) quanto as três variações de segmentos testadas (3s, 5s e 10s). O mesmo procedimento foi replicado para todas as variações com consistência.

O treinamento foi realizado utilizando a biblioteca *PyTorch*, e o processo de treinamento seguiu uma abordagem baseada em aprendizado por transferência, com inicialização dos pesos a partir da versão pré-treinada ViT-B/16 no ImageNet. A arquitetura ViT-B/16 utilizada possui 12 camadas, cada uma com uma dimensão de *embedding* de 768 unidades, e um *feedforward* com dimensão de 3072 unidades. O modelo emprega 12 cabeças de atenção, totalizando 86 milhões de parâmetros. Essas características proporcionam um balanceamento eficaz entre capacidade de modelagem e eficiência computacional, permitindo ao ViT lidar com tarefas de classificação complexas, como o reconhecimento de gêneros musicais em espectrogramas.

Os experimentos de treinamento do ViT foram realizados localmente em uma GPU NVIDIA GeForce RTX 4090, escolhida pela sua capacidade de lidar com modelos complexos e grandes volumes de dados. Os pesos pré-treinados foram carregados utilizando o pacote *torchvision*, especificamente o conjunto de pesos padrão do ViT-B/16. O modelo foi inicialmente congelado, mantendo fixas todas as camadas internas do encoder. A partir deste ponto, foi aplicado um ajuste fino progressivo, no qual as camadas do encoder foram sendo descongeladas uma a uma ao longo do treinamento, conforme marcos (*milestones*) predefinidos eram atingidos. Foram utilizados três marcos principais (épocas 10, 30 e 40), com o descongelamento de uma nova camada a cada marco, sempre iniciando pela última camada do encoder e avançando até camadas mais iniciais.

Em paralelo ao descongelamento, a taxa de aprendizado foi ajustada de forma progressiva a cada marco, seguindo os fatores de decaimento definidos experimentalmente. O otimizador utilizado foi o **Adam**, com taxa de aprendizado inicial de 1×10^{-4} . A função de perda utilizada foi a *Cross-Entropy Loss*, e o modelo foi treinado com *early stopping*, empregando uma paciência de 10 épocas para evitar sobreajuste.

Durante o treinamento, a cada época, foram extraídos os *embeddings* do modelo e gerados gráficos de visualização com PCA, t-SNE e UMAP, a partir dos dados de validação. Esses embeddings foram obtidos diretamente do token [CLS] da última camada do encoder do ViT, antes da classificação. Essa etapa permitiu monitorar a evolução da organização dos gêneros musicais no espaço vetorial ao longo do treinamento, como parte da análise qualitativa. Ao final do treinamento, o melhor modelo foi selecionado com base na menor perda de validação. Os pesos correspondentes foram salvos e utilizados para avaliação final no conjunto de teste e para a extração definitiva dos *embeddings*.

4.8. Arquitetura do Vision Transformer Aplicada a Espectrogramas

A arquitetura ViT-B/16 foi a principal utilizada para realizar os procedimentos experimentais desta dissertação. Um dos fatores que justifica essa escolha é sua eficácia como uma rede voltada para o reconhecimento de imagens, característica que a torna adequada para a análise de espectrogramas, que são representações visuais de sinais de áudio. Além disso, a arquitetura foi avaliada em comparação com modelos *baseline* amplamente utilizados em tarefas de classificação, como a ResNet50 e a SVM, e obteve os melhores resultados em todos os cenários experimentais testados, comprovando sua superioridade na tarefa proposta. A Figura 4.5 exibe as camadas do modelo ViT processando como entrada do modelo um espectrograma de áudio, diferente de uma imagem convencional.

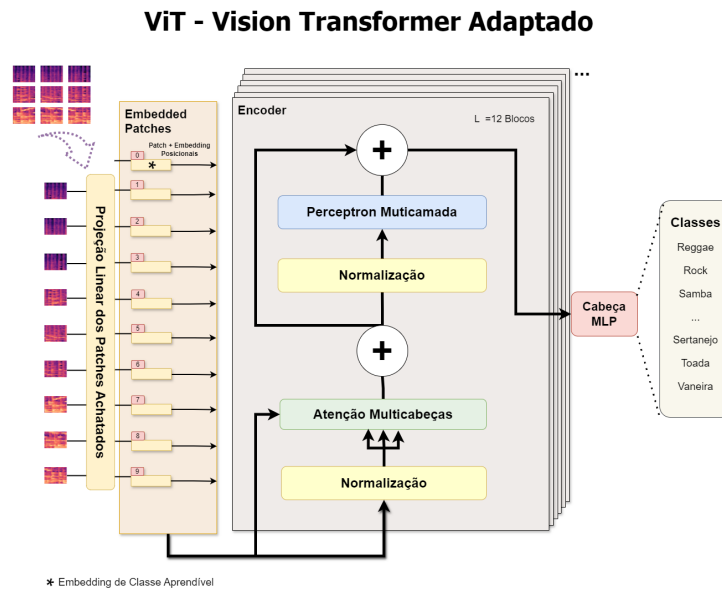


Figura 4.5. Arquitetura ViT.

O ViT particiona o espectrograma de áudio em várias imagens menores chamadas de *patches*. Cada *patch* é uma região quadrada da imagem, com 16x16 pixels. No ViT, o espectrograma de entrada tem uma dimensão padrão de 224x224 pixels, e o tamanho dos patches também é fixado em 16x16 pixels. Portanto, o espectrograma de 224x224 pixels será dividido em patches menores de 16x16 pixels cada, que serão processados pelo ViT para a tarefa de classificação de gêneros musicais.

Além dos *patches* extraídos do espectrograma, o modelo ViT insere um token especial chamado [CLS] no início da sequência de entrada. Esse token é um vetor apren-

dível que não representa um *patch* da imagem, mas sim um agregador global de informações. Durante o processamento pelas camadas do encoder, o [CLS] interage com todos os demais tokens por meio dos mecanismos de atenção. Ao final do encoder, o vetor correspondente ao [CLS] contém uma representação compacta e contextualizada de toda a entrada, sendo utilizado como entrada da MLP final para a tarefa de classificação. Neste trabalho, os *embeddings* extraídos para análise de similaridade entre gêneros musicais foram obtidos diretamente desse token, por refletirem a síntese global do espectrograma.

Após o espectrograma ser convertido em *patches*, cada *patch* precisa ser achatado (*flattened*) de um vetor bidimensional (2D) para um vetor unidimensional (1D). Esse processo é necessário porque o transformador requer que cada *patch* seja representado como um vetor unidimensional para ser tratado como um *token* de entrada. Como cada *patch* possui dimensões de 16x16 pixels, transformar esse *patch* em um vetor 1D consiste em multiplicar as dimensões. Em seguida, esse vetor é passado por uma camada de projeção linear (também chamada de *embedding*), que transforma o vetor achatado em um vetor com uma dimensão específica de 768 unidades, no caso do ViT-B/16, para ser usado pelo transformador. A projeção linear realiza essa transformação aplicando uma multiplicação dos vetores de entrada pelos pesos, seguida da adição de um termo de *bias*, similar ao que ocorre em camadas convolucionais. Especificamente, cada vetor 1D é transformado em um vetor com dimensão maior e fixada. Essa operação preserva a informação relevante do *patch*, enquanto ajusta sua dimensionalidade para facilitar o processamento pelo modelo.

Após o processo de projeção linear, são adicionados *embeddings* posicionais aos vetores resultantes. Esses *embeddings* posicionais indicam a localização de cada *patch* no espectrograma original, o que é crucial porque os *patches* não são alimentados sequencialmente na rede. A adição dos *embeddings* posicionais permite que o modelo mantenha a informação de localização, essencial para compreender a posição no espectrograma original. Finalmente, os *patches* com *embeddings* posicionais são alimentados na rede, que então processa esses vetores para realizar a classificação de gêneros musicais.

O ViT-Base/16 incorpora 12 blocos de *encoder*, que formam o núcleo do processamento de dados da arquitetura. Cada bloco aplica um mecanismo de **autoatenção com múltiplas cabeças** (do inglês, *Multi-Head Self-Attention*), permitindo que cada *patch* interaja com todos os outros *patches* do espectrograma. Esse mecanismo é crucial para capturar relações complexas tanto ao nível local quanto global, enriquecendo a repre-

sentação dos dados. Após a autoatenção, uma camada de **normalização** é aplicada para estabilizar o processo de treinamento e melhorar a robustez do modelo, assegurando que o treinamento seja mais eficiente e evitando problemas como a explosão ou o desaparecimento de gradientes. Em seguida, os dados processados passam por uma rede neural *feedforward* identificada na Figura 4.5 como MLP, que refina ainda mais a representação dos *patches*. A saída dessa rede é novamente normalizada e somada à entrada original do bloco através de uma conexão residual, promovendo uma integração mais eficaz das informações.

Após o processamento pelos blocos do *encoder*, o *token* de classe, que agora contém uma representação abrangente das informações globais da sequência, é passado por uma MLP. A **cabeça MLP** é composta por uma ou mais camadas densas, que mapeiam o *token* de classe para o espaço de saída, gerando a representação final necessária para a classificação. A última camada da *MLP Head* produz a predição para cada uma das classes de gêneros musicais. Geralmente, essa saída é passada por uma função *Softmax*, que converte as pontuações em probabilidades, permitindo ao modelo determinar a classe com maior probabilidade como a predição final.

4.9. Análise dos Embeddings e Similaridade entre Gêneros Musicais

Com o objetivo de compreender como os gêneros musicais são organizados no espaço vetorial aprendido pelos modelos, foi conduzida uma análise exploratória dos *embeddings* gerados pelas arquiteturas ViT e ResNet50. A extração dos embeddings foi realizada sobre os dados de validação ao longo do treinamento. No caso do Vision Transformer (ViT), foram extraídos os vetores correspondentes ao token [CLS] da última camada do codificador, antes da camada final de classificação. Essa extração foi realizada diretamente após o bloco de codificação, utilizando os espectrogramas como entrada, conforme descrito na Seção 4.7. Os vetores extraídos representam a síntese global do conteúdo do espectrograma processado pelo ViT.

Para a ResNet50, os embeddings foram obtidos a partir da saída da penúltima camada da rede, imediatamente após o bloco de *average pooling* e antes da camada totalmente conectada responsável pela classificação. Essa extração permite capturar representações ricas aprendidas ao longo das camadas convolucionais. A visualização da distribuição dos embeddings foi realizada utilizando técnicas de redução de dimensio-

nalidade: PCA, t-SNE e UMAP. Cada técnica foi aplicada de forma independente aos embeddings extraídos, e os resultados foram representados graficamente em dispersões bi-dimensionais com representação por cor para cada gênero musical. Essa visualização foi repetida em diferentes estágios do treinamento para avaliar como o aprendizado influencia a separabilidade entre as classes no espaço latente.

Além da análise visual, foi realizada uma avaliação quantitativa de similaridade entre os gêneros musicais, utilizando a métrica de similaridade cosseno. Para cada gênero, foi calculado o vetor médio dos embeddings, e então a similaridade cosseno foi computada entre todos os pares de gêneros. Os valores obtidos foram organizados em uma matriz de similaridade e visualizados por meio de um mapa de calor. Essa abordagem permite observar a proximidade e a distinção entre os estilos musicais regionais brasileiros no espaço vetorial aprendido pelos modelos. O modelo SVM não foi incluído nesta etapa, pois não possui representações internas aprendidas, sendo baseado em características manuais extraídas previamente.

Capítulo 5

Resultados

Este capítulo apresenta em detalhes os experimentos conduzidos e os resultados finais obtidos, conforme os procedimentos estabelecidos na seção de metodologia. Os resultados descritos refletem a fase conclusiva da pesquisa, incluindo análises sobre o desempenho dos modelos avaliados, comparação entre diferentes configurações e interpretações baseadas nas hipóteses formuladas para desenvolvimento da pesquisa.

5.1. Revisão dos Experimentos da Qualificação

Durante a etapa de qualificação, foram realizados dois experimentos exploratórios com o objetivo de investigar o uso da arquitetura Vision Transformer (ViT) na tarefa de reconhecimento de gêneros musicais. Naquele momento, a base BYRM era composta por apenas cinco gêneros e foi combinada à base GTZAN para compor um total de quinze classes.

No Experimento 1, o modelo foi treinado apenas com a GTZAN, enquanto no Experimento 2, foi utilizada a combinação GTZAN + BYRM. Em ambos os casos, o treinamento foi conduzido em duas etapas: inicial com a cabeça classificadora congelada e posterior ajuste fino das últimas camadas do encoder, onde o objetivo era verificar a viabilidade do uso de um ViT para classificação de gêneros musicais a partir de espectrogramas. As métricas globais obtidas após o fine-tuning são apresentadas na Tabela [5.1](#).

Os resultados obtidos foram consistentes com o esperado para a arquitetura, no qual o segundo experimento, contando com a combinação das duas bases, apresentou certa melhoria em relação ao primeiro experimento. Além das métricas globais, foram

Tabela 5.1. Métricas gerais obtidas na qualificação após ajuste fino.

Métrica	Experimento 1	Experimento 2
Acurácia	86,28%	87,36%
Precisão	87,05%	87,70%
Revocação	86,28%	87,36%
F1-score	86,03%	87,41%

extraídas métricas por classe e observada a evolução do modelo durante o ajuste fino.

A Figura 5.1 mostra a curva de acurácia e perda na fase do treinamento com as três últimas camadas de encoder descongeladas. Observa-se uma rápida convergência da perda de treino, que atinge valores baixos logo nas primeiras épocas. No entanto, a perda de validação estabiliza em um patamar superior, enquanto a acurácia de validação cresce de forma mais lenta e se mantém distante da acurácia de treino. Esse distanciamento entre as curvas é indicativo de sobreajuste, possivelmente causado pelo descongelamento simultâneo de três camadas do encoder, o que aumentou abruptamente a capacidade do modelo de memorizar os dados de treino, sem mecanismos adequados de regularização.

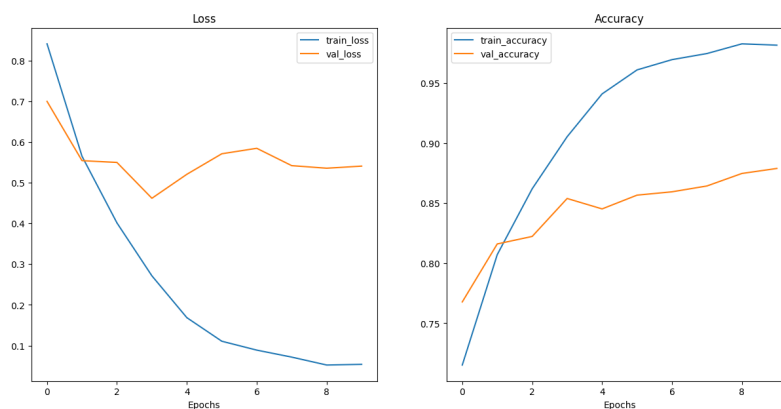


Figura 5.1. Funções de perda e acurácia ao longo do treinamento com a base GTZAN e BYRM.

A matriz de confusão do experimento, apresentada na Figura 5.2, reflete os altos índices observados nas métricas globais. Gêneros acusticamente distintos, como Classical e Metal, foram classificados com elevada precisão. Já categorias mais próximas, como Sertanejo, Toada, Vaneira e Axé, apresentaram alguns erros cruzados, ainda que em proporções reduzidas. Esse desempenho superior, mesmo entre classes similares, indica que

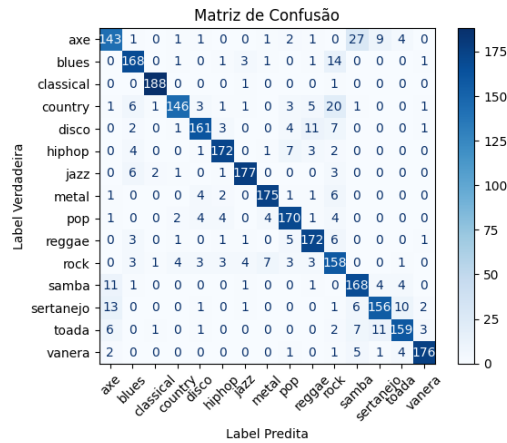


Figura 5.2. Matriz de confusão após o ajuste fino.

o modelo estava se beneficiando de um cenário mais simples do que o problema real.

Apesar das métricas elevadas, a metodologia apresentou uma limitação relevante: os segmentos de áudio foram extraídos antes da separação dos conjuntos de treino, validação e teste, o que resultou na presença de trechos da mesma faixa em diferentes divisões. Esse vazamento de dados comprometeu a avaliação da capacidade de generalização do modelo e acabou impulsionando melhorias na construção da base e na definição de um novo procedimento experimental. A base BYRM foi reformulada de modo que a divisão entre treino, validação e teste ocorresse antes da segmentação dos áudios, evitando sobreposição de trechos entre os conjuntos. Além disso, a base foi expandida e passou a focar exclusivamente nos gêneros regionais brasileiros. A seguir, são apresentados os resultados obtidos com essa nova abordagem experimental.

5.2. Análise da Influência Temporal na Classificação de Gêneros Musicais

Nesta seção, serão apresentados os resultados dos experimentos realizados com o modelo Vision Transformer (ViT), adotado como modelo principal desta pesquisa, sobre a base de dados BYRM. O objetivo da análise é investigar como a posição temporal do trecho influencia o desempenho da classificação de gêneros musicais regionais brasileiros.

Para isso, cada faixa foi dividida em blocos de 30 segundos, sendo extraídos segmentos de 3 segundos com sobreposição de 50% dentro de cada bloco, procedimento detalhado em 4.4. O modelo ViT foi treinado individualmente em cada um desses blocos temporais, com o objetivo de identificar quais trechos apresentaram os melhores e piores desempenhos. Os resultados foram avaliados com base em quatro métricas de classifica-

Tabela 5.2. Desempenho do ViT nos diferentes blocos de 30 segundos (segmentos de 3s).

Trecho	Acurácia (%)	Precisão (%)	Revocação (%)	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
30s finais	76,16	76,94	76,16	75,96

ção: acurácia, precisão, revocação e F1-score. A Tabela 5.2 apresenta os resultados obtidos pelo modelo ViT em cada bloco de 30 segundos da faixa musical. A análise permite observar diferenças significativas de desempenho entre os trechos temporais. O trecho de 90–120 segundos obteve o melhor desempenho global, atingindo 79,60% de acurácia e 79,51% de F1-score. Em contraste, o trecho inicial de 0–30 segundos apresentou os piores resultados, com acurácia de 74,76% e F1-score de 74,91%.

Apesar de o trecho de 150–180 segundos apresentar acurácia e F1-score ainda menor do que o inicial, optou-se por considerar o trecho 0–30 segundos como o de pior desempenho para fins de comparação. Isso se deve ao fato de que nem todas as faixas atingem 180 segundos de duração, o que resulta em um desequilíbrio entre classes no final das músicas. A métrica de F1-score foi adotada como principal critério de avaliação, por ser mais robusta em cenários de classificação multiclasse. Diferentemente da acurácia, que pode ser afetada por desbalanceamentos de classe, o F1-score harmoniza precisão e revocação, oferecendo uma medida mais equilibrada do desempenho do modelo em todas as classes. No entanto, métricas globais podem ocultar o comportamento individual dos gêneros musicais. A Figura 5.3 ilustra os valores de acurácia por gênero musical ao longo dos segmentos de 30 segundos, evidenciando variações consideráveis entre os trechos.

A Figura 5.4 apresenta os valores de F1-score por gênero em cada trecho de 30 segundos em que as músicas foram divididas, permitindo uma análise mais detalhada da performance do modelo ao longo do tempo.

Os resultados experimentais apoiam a segunda hipótese da pesquisa, que propõem se a posição do trecho dentro da música influencia o desempenho da classificação. O trecho central, de 90–120 segundos, apresentou o melhor desempenho geral, enquanto o

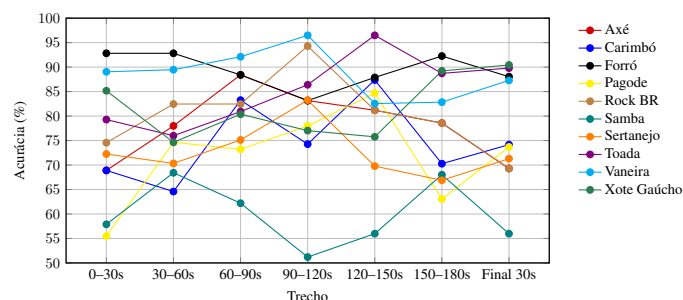


Figura 5.3. Acurácia por gênero em cada trecho de 30 segundos com o modelo ViT.

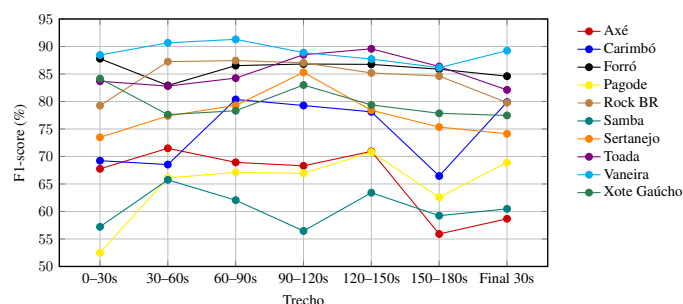


Figura 5.4. F1-score por gênero em cada trecho de 30 segundos com o modelo ViT.

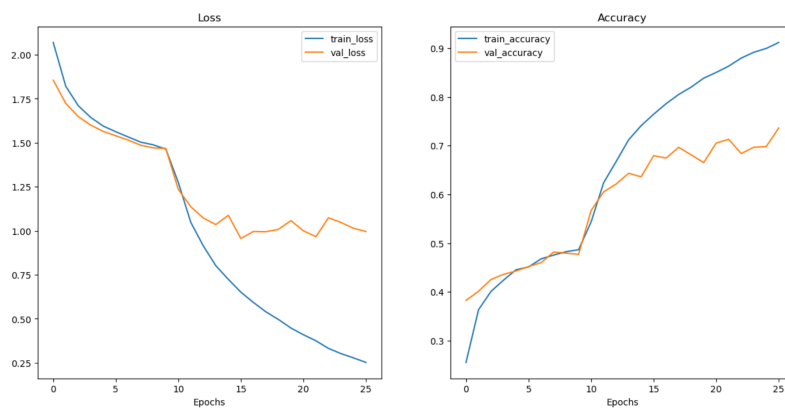
início da faixa 0–30 segundos apresentou o pior resultado. Esse comportamento sugere que trechos mais centrais carregam maior densidade de informações características dos gêneros, o que está alinhado com estudos da cognição musical que indicam que refrões e partes centrais da canção costumam ser mais representativos. Entretanto, a análise revelou variações entre os gêneros. Por exemplo, Samba e Axé atingiram seus picos de desempenho no trecho de 30–60 segundos, enquanto a vaneira apresentou seu melhor resultado no trecho de 60–90 segundos. Esses resultados indicam que cada gênero tende a revelar suas características mais distintivas em diferentes momentos da música, o que reforça a importância de uma análise temporal segmentada.

5.2.1. Comparação da Classificação entre Trechos: Melhor e Pior Posição

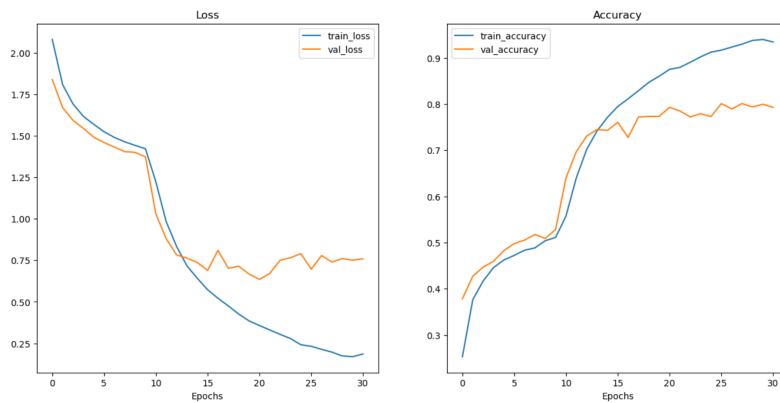
A partir da avaliação geral apresentada na Tabela 5.2, foram selecionados os trechos com melhor e pior desempenho para uma análise mais profunda: o segmento entre 90 e 120 segundos, que obteve os melhores resultados, e o trecho inicial de 0 a 30 segundos, que apresentou os piores índices de desempenho. Essa análise será explorada ao longo da seção de resultados, onde esses trechos serão analisados em outros experimentos.

A Figura 5.5(a) apresenta as curvas de perda e acurácia do modelo ViT, treinado

com os segmentos do trecho de 0–30 segundos. Nota-se que, embora a perda de treino (*train_loss*) reduza continuamente, a perda de validação (*val_loss*) começa a oscilar a partir da época 12, sem apresentar mais ganhos significativos. O mesmo padrão é observado na acurácia: a acurácia de treino ultrapassa 90% ao final do treinamento, enquanto a acurácia de validação estabiliza em torno de 70%, indicando uma discrepância de desempenho entre os conjuntos. Esse comportamento é um indicativo claro de sobreajuste ou *overfitting*, em que o modelo se ajusta aos padrões do conjunto de treino, mas perde capacidade de generalização para dados não vistos. O critério de *early stopping* foi ativado na época 25, quando a acurácia de validação deixou de apresentar melhorias consecutivas.



((a)) Trecho de 0–30 segundos.



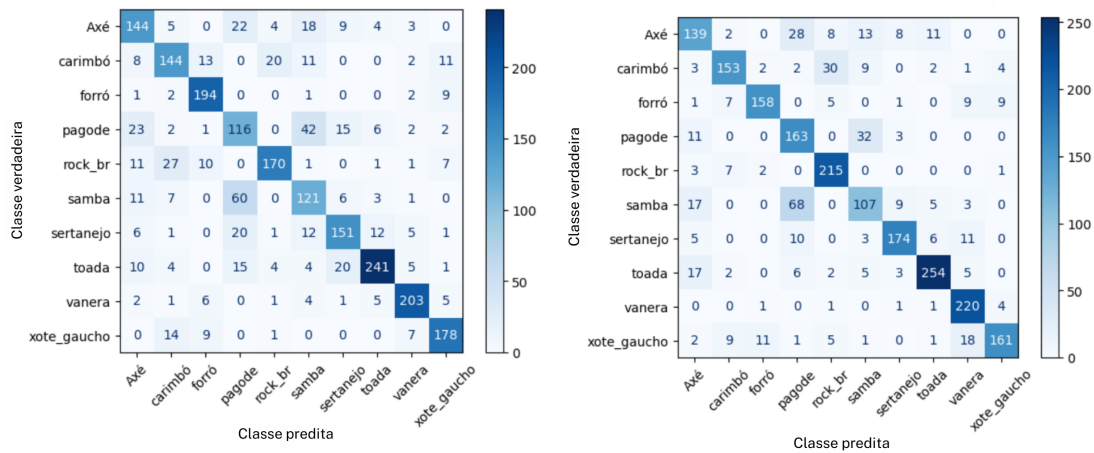
((b)) Trecho de 90–120 segundos.

Figura 5.5. Funções de perda e acurácia ao longo do treinamento para o melhor (90–120 segundos) e pior trecho (0–30 segundos).

Já a Figura 5.5(b) exibe as curvas de aprendizado para o melhor trecho identificado, de 90–120 segundos. Observa-se uma convergência mais estável: tanto a perda quanto a acurácia de validação seguem acompanhando a tendência do conjunto de treino até aproximadamente a época 20. Após esse ponto, a acurácia de validação estabiliza em

torno de 80%, enquanto a de treino segue crescendo. Ainda que essa diferença também sugira um início de sobreajuste, ela é bem menos acentuada do que no trecho inicial, e o modelo demonstra maior capacidade de generalização. O *early stopping* foi ativado na época 30, quando a acurácia de validação se estabilizou. É importante destacar que, nos dois casos, foi adotado um processo de ajuste fino em que apenas a última camada do ViT foi descongelada. Essa camada, por ser altamente especializada e diretamente ligada à saída do modelo, tende a se ajustar rapidamente aos dados de treinamento. Esse comportamento, por um lado, acelera a convergência, mas também aumenta a sensibilidade ao sobreajuste. Isso explica por que o modelo alcança rapidamente altos níveis de desempenho no treino, ao passo que a validação estabiliza precocemente. Ainda assim, o uso do *early stopping* foi eficaz em ambos os casos para interromper o treinamento antes que o desempenho em validação começasse a se degradar.

A Figura 5.6 apresenta as matrizes de confusão dos dois trechos analisados, permitindo observar os padrões de acertos e erros específicos em cada cenário.



((a)) Matriz de confusão do trecho de 0–30 segundos.

((b)) Matriz de confusão do trecho de 90–120 segundos.

Figura 5.6. Matrizes de confusão do melhor e pior trecho classificado com o modelo ViT.

As matrizes de confusão apresentadas nas Figuras 5.6(a) e 5.6(b) ilustram os resultados do modelo ViT treinado com segmentos de 3 segundos, extraídos dos trechos de 0–30 segundos e 90–120 segundos, respectivamente. De acordo com as métricas globais de desempenho (acurácia, precisão, recall e F1-score), o trecho de 90–120 segundos

foi identificado como o melhor ao longo da música, enquanto o trecho de 0–30 segundos apresentou o pior desempenho médio entre todos os blocos analisados. Apesar disso, uma análise mais cuidadosa revela que o trecho inicial obteve resultados superiores em algumas classes específicas. Por exemplo, o gênero axé teve 144 acertos no trecho de 0–30 segundos, contra 139 no de 90–120 segundos. De modo similar, o gênero forró apresentou desempenho comparável nos dois trechos, com ligeira vantagem no início, possivelmente influenciado pelo número maior de instâncias disponíveis nesse intervalo. Isso se deve ao fato de que nem todas as faixas da base possuem duração suficiente para atingir o trecho de 90–120 segundos, o que resulta em um número menor de amostras para alguns gêneros nesse bloco. Esse desbalanceamento temporal impacta a quantidade de dados por classe e, consequentemente, o desempenho do modelo em determinadas situações. Ainda assim, o modelo obteve melhorias substanciais no trecho final para diversos gêneros, como rock brasileiro (215 acertos contra 170 no início), pagode (163 contra 116) e xote gaúcho (161 contra 178, porém com menor dispersão de erros). Esses resultados evidenciam que as características mais distintivas de certos gêneros podem se manifestar com maior clareza em momentos posteriores da música, como refrões ou trechos melódicos centrais. Dessa forma, o experimento reforça a hipótese de que diferentes gêneros apresentam picos de expressividade em segmentos distintos da música, tornando a segmentação temporal um fator crítico na eficácia da classificação automática de gêneros musicais.

5.3. Comparação de Desempenho entre Diferentes Modelos de Classificação

Para investigar o impacto da arquitetura na tarefa de classificação de gêneros musicais, comparamos o desempenho de três modelos: Vision Transformer (ViT), Support Vector Machine (SVM) e ResNet50. Os experimentos foram conduzidos utilizando os trechos previamente identificados como os de melhor e pior desempenho: 90–120 segundos e 0–30 segundos, respectivamente. A Tabela 5.3 resume os resultados obtidos com base nas métricas de acurácia, precisão, revocação e F1-score.

Dentre os modelos avaliados, o ViT apresentou desempenho consistentemente superior em todos os cenários e métricas. Esse resultado pode ser atribuído à sua capacidade de capturar relações contextuais entre diferentes regiões do espectrograma por meio de mecanismos de autoatenção aplicados diretamente aos *patches* da imagem. Essa abordagem permite que o modelo compreenda padrões harmônicos, rítmicos e tímbricos distribuídos ao longo do espectro de frequências, favorecendo a identificação de características

Tabela 5.3. Comparação de desempenho entre ViT, SVM e ResNet50 nos melhores (90–120s) e piores (0–30s) trechos da base.

Métrica	90–120s			0–30s		
	ViT	SVM	ResNet50	ViT	SVM	ResNet50
Acurácia (%)	79,60	70,06	56,41	74,76	63,02	51,37
Precisão (%)	80,24	69,41	56,49	75,29	63,09	50,98
Revocação (%)	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

discriminativas entre os gêneros musicais.

O classificador SVM, embora baseado em técnicas clássicas de aprendizado supervisionado, demonstrou desempenho competitivo, especialmente no trecho de 90–120 segundos. Esse resultado pode ser atribuído ao fato de alguns gêneros apresentarem características acústicas bastante diferentes entre si, facilitando sua separação pelo classificador SVM. A Figura 5.7 apresenta as métricas de desempenho por classe. É possível observar que os gêneros toada, vaneira e forró apresentaram os melhores resultados gerais, com F1-scores superiores a 75%, indicando que o modelo foi capaz de aprender padrões consistentes para esses gêneros. Em contrapartida, gêneros como carimbó e xote gaúcho obtiveram desempenhos mais baixos, principalmente em termos de *revocação*, sugerindo maior dificuldade do modelo em identificar corretamente todas as ocorrências dessas classes.

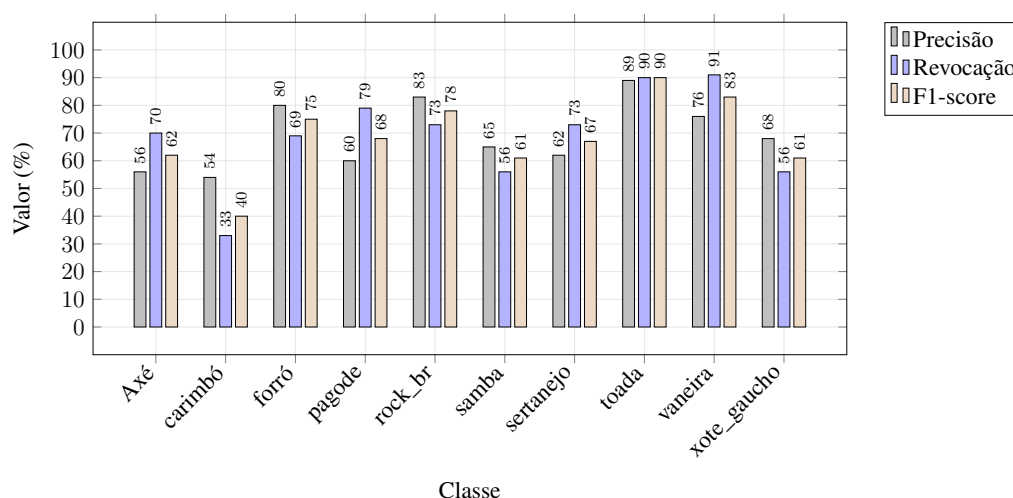


Figura 5.7. Métricas de desempenho por classe para o classificador SVM no trecho de 90–120 segundos.

A matriz de confusão, ilustrada na Figura 5.8, reforça essa análise ao mostrar como os erros de classificação se distribuem entre as classes. Nota-se que toada foi a classe mais bem reconhecida, com 265 instâncias corretamente classificadas e pouquíssima confusão com outras categorias. Já o carimbó foi frequentemente confundido com gêneros como axé, vaneira e xote gaúcho, evidenciando a sobreposição de padrões acústicos entre esses estilos. O gênero samba também apresentou confusões com pagode, o que é esperado dado a alta similaridade entre os dois gêneros.

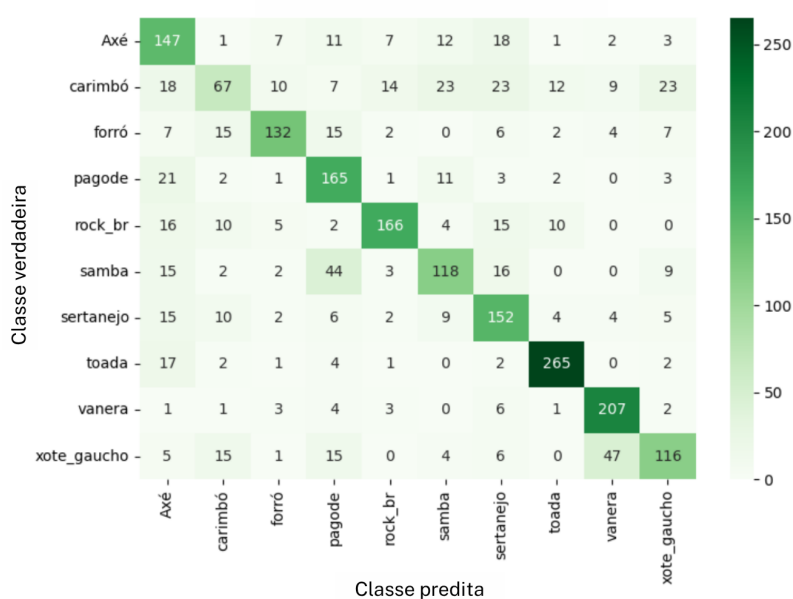


Figura 5.8. Matriz de confusão obtida com o SVM aplicado ao trecho de 90–120 segundos da base BYRM.

Por outro lado, a ResNet50 apresentou os menores valores de métricas em todos os testes, tanto no melhor quanto no pior trecho da base. Vale destacar que o modelo foi treinado com estratégias cuidadosamente definidas para competir em condições justas com o ViT, incluindo o uso dos mesmos dados de entrada, ajustes de fine-tuning e seleção criteriosa de hiperparâmetros. A Figura 5.9 apresenta as curvas de perda e acurácia da ResNet50 ao longo de 110 épocas de treinamento, tanto para os conjuntos de treino quanto de validação. Observa-se inicialmente uma redução consistente da perda e um aumento gradual da acurácia nas primeiras 40 épocas, o que indica que o modelo está aprendendo de forma eficiente nesse estágio.

Entretanto, após esse ponto, nota-se uma estabilização das métricas, com o conjunto de validação apresentando desempenho inferior ao de treinamento, uma diferença

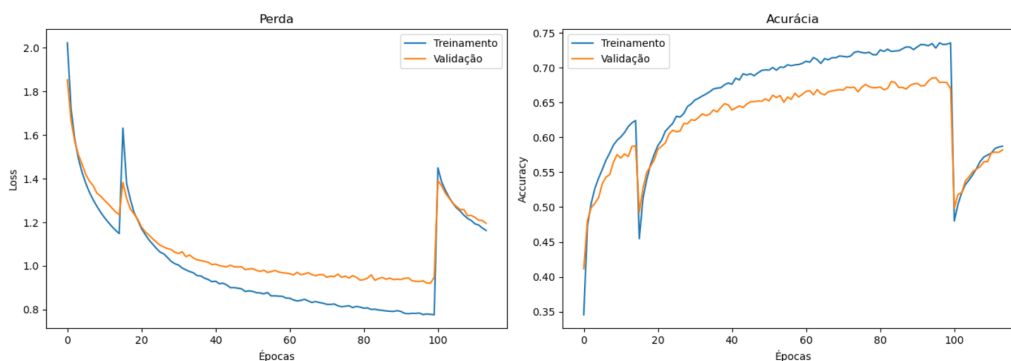


Figura 5.9. Funções de perda e acurácia ao longo do treinamento para o melhor trecho (90–120 segundos)

que se torna mais evidente a partir da época 50 até a época 100. Esse distanciamento entre as curvas de treino e validação é indicativo de sobreajuste, sugerindo que o modelo passou a memorizar os padrões do conjunto de treino sem generalizar bem para dados não vistos. Esse padrão reforça a hipótese de que, embora a ResNet50 possua uma arquitetura profunda e poderosa, ela é sensível a configurações específicas de treinamento, especialmente no contexto de espectrogramas musicais, que possuem características visuais distintas de imagens naturais. Estratégias de regularização mais agressivas, como *dropout*, *data augmentation* mais variado ou *early stopping* mais conservador, poderiam mitigar esse comportamento.

A Figura 5.10 apresenta a matriz de confusão obtida pelo classificador ResNet50 no trecho de 90 a 120 segundos da base BYRM. Observa-se que algumas classes como *rock_br*, *forró*, *vaneira* e *xote_gauche* obtiveram altos índices de acerto, com predominância de valores na diagonal principal, indicando boa capacidade do modelo em reconhecer esses gêneros. No entanto, há altas confusões entre classes com características similares, como no caso de *axé* e *toada*, onde muitos exemplos de *axé* foram classificados como *toada*. Também é possível observar que *xote_gauche* foi frequentemente confundido com *forró*, o que sugere que esses gêneros compartilham características visuais nos espectrogramas que dificultam a discriminação pelo modelo.

Gêneros como *pagode* e *samba* também apresentaram sobreposição, evidenciada pela quantidade significativa de predições cruzadas entre eles. Essa sobreposição reflete a proximidade estilística entre os gêneros, especialmente em termos de ritmo e instrumentação, o que reforça a complexidade do problema de classificação no contexto de gêneros musicais regionais.

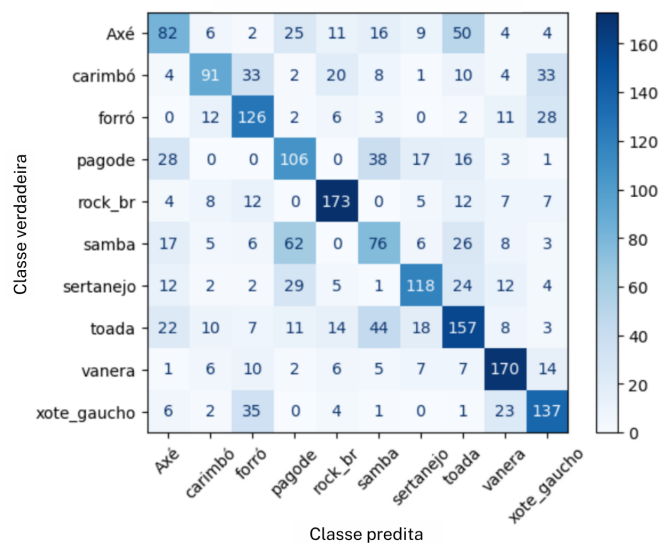


Figura 5.10. Matriz de confusão obtida com a ResNet50 aplicada ao trecho de 90–120 segundos da base BYRM.

A Figura 5.11 apresenta as métricas de desempenho por classe obtidas pela ResNet50 no segmento de 90 a 120 segundos da base BYRM. Observa-se que gêneros como rock_br, forró, vaneira e xote_gaucho alcançaram resultados mais expressivos, com F1-scores superiores a 60%, refletindo a capacidade do modelo em identificar padrões consistentes nesses estilos musicais. Por outro lado, gêneros como axé, carimbó, pagode e samba apresentaram desempenhos baixos em todas as métricas, o que evidencia maior dificuldade do modelo em distinguir entre eles. Esses gêneros contribuíram para a queda dos valores globais das métricas da ResNet50, indicando que o modelo possui maior sensibilidade a variações internas ou sobreposição de características entre as classes.

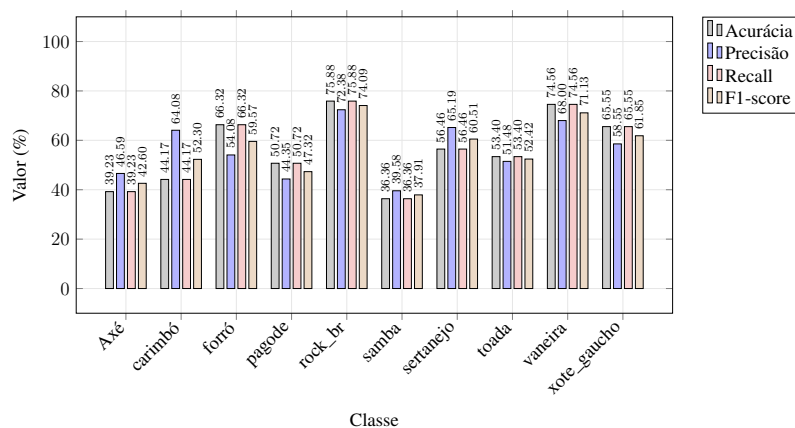


Figura 5.11. Métricas de desempenho por classe para o classificador ResNet50 no segmento de 90–120 segundos.

Mesmo com a aplicação dos mesmos dados de entrada do ViT, ajuste fino e seleção criteriosa de hiperparâmetros, o desempenho da ResNet50 permaneceu inferior ao do ViT. Esses resultados indicam possíveis limitações da arquitetura em capturar padrões discriminativos relevantes nos espectrogramas utilizados. Isso pode estar relacionado à sensibilidade do modelo a configurações como taxa de aprendizado, regularização ou mesmo à sua arquitetura, que pode não ser ideal para o tipo de representação visual utilizada. Ajustes mais refinados ou abordagens alternativas de treinamento podem ser necessários para que esse modelo atinja resultados mais expressivos.

5.4. Avaliação Baseada na Duração dos Segmentos

Esta seção investiga como a duração dos segmentos influencia a performance na tarefa de classificação utilizando apenas o modelo Vision Transformer (ViT), que apresentou o melhor desempenho nas etapas anteriores. Os experimentos buscam entender com os resultados se segmentos mais longos favorecem a representação dos gêneros musicais, especialmente em diferentes partes temporais das faixas. Segmentos de 3, 5 e 10 segundos foram extraídos com sobreposição de 50% dos trechos de melhor desempenho (90–120 segundos) e de pior desempenho (0–30 segundos), conforme identificado nas análises anteriores. A Tabela 5.4 apresenta os resultados obtidos pelo ViT para cada configuração de duração em ambos os trechos.

Tabela 5.4. Desempenho do ViT com diferentes durações de segmento nos trechos de melhor (90–120s) e pior (0–30s) desempenho.

Métrica	Trecho: 90–120s			Trecho: 0–30s		
	3s	5s	10s	3s	5s	10s
Acurácia (%)	79,60	78,94	81,94	74,76	73,50	73,16
Precisão (%)	80,24	79,48	82,85	75,29	76,00	75,35
Revocação (%)	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

Os resultados mostram que o melhor desempenho foi alcançado com segmentos de 10 segundos no trecho de 90–120 segundos, atingindo 81,94% de acurácia e 81,84% de F1-score. Esse desempenho sugere que janelas mais longas capturam padrões mais completos e recorrentes no espectro da música, como refrões, variações harmônicas ou instrumentais características de cada gênero.

Por outro lado, no trecho de 0–30 segundos, os segmentos de 3 segundos apresentaram ligeira vantagem sobre os demais, embora o desempenho geral tenha sido inferior ao trecho posterior. A Figura 5.12 detalha o F1-score por gênero no trecho de 90–120 segundos com as três durações de segmento avaliadas. É possível observar que cada gênero responde de forma distinta às variações na duração dos segmentos.

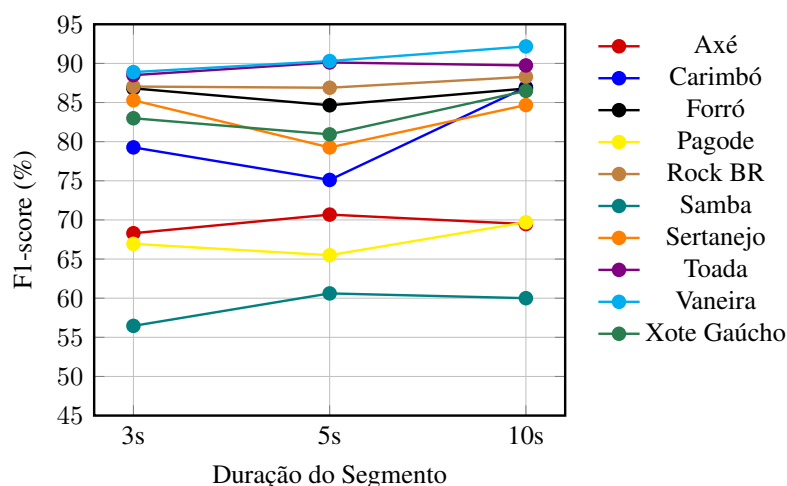


Figura 5.12. F1-score por gênero no trecho de 90–120s com diferentes durações de segmento (ViT).

Os dados ilustrados na Figura 5.12 evidenciam que não existe uma única duração de segmento ideal que maximize o desempenho para todos os gêneros musicais. A variação dos F1-scores ao longo dos diferentes segmentos, demonstra que cada gênero responde de maneira distinta ao tamanho da janela de análise, o que revela uma dependência direta da estrutura temporal e dos elementos musicais predominantes de cada estilo. Gêneros como Vaneira, Carimbó e Rock BR apresentaram crescimento consistente do F1-score à medida que a janela aumentava, alcançando os melhores desempenhos com 10 segundos. Esse padrão sugere que esses estilos se beneficiam de um maior contexto temporal, provavelmente devido à presença de elementos melódicos mais longos, introduções instrumentais ou progressões harmônicas mais complexas.

Por outro lado, gêneros como Forró, Toada e Xote Gaúcho atingiram F1-scores elevados já com segmentos curtos de 3 segundos, o que indica que possuem padrões rítmicos bem definidos e características acústicas marcantes que podem ser capturadas rapidamente. Esses estilos demonstram menor dependência de contexto estendido e são bem representados por janelas menores. Alguns gêneros apresentaram oscilações significativas. O Sertanejo, por exemplo, teve alto desempenho com segmentos de 3s e 10s,

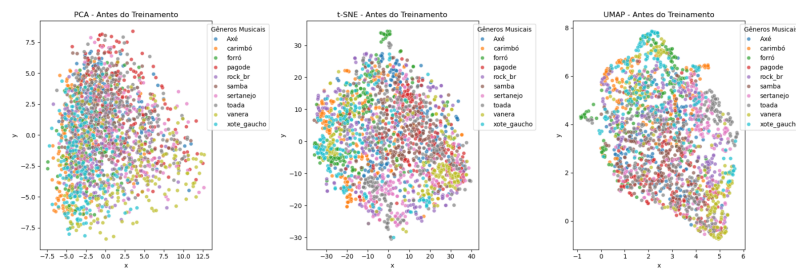
mas apresentou uma queda perceptível com 5s. O Carimbó teve desempenho inferior com 5s em comparação a 3s, mas mostrou uma melhora substancial com 10s. Esses comportamentos sugerem que janelas intermediárias nem sempre representam um ponto de equilíbrio e podem, em alguns casos, capturar partes menos informativas da faixa.

Já gêneros como Axé e Samba mantiveram desempenho relativamente estável entre 3s e 5s, sem ganhos evidentes com o aumento para 10s. Isso pode indicar que suas estruturas combinam elementos rítmicos e melódicos de forma equilibrada, sendo suficientemente representadas por janelas de duração moderada. Esses resultados reforçam a importância de estratégias de segmentação adaptativas, que considerem tanto as características estruturais de cada gênero quanto os momentos mais representativos da música. Trabalhos futuros podem explorar segmentações orientadas por análise de estrutura musical ou ainda desenvolver abordagens dinâmicas que identifiquem automaticamente as janelas mais informativas para cada faixa ou classe. A adoção de janelas variáveis por gênero ou até mesmo o uso combinado de múltiplas resoluções pode representar uma solução eficaz para capturar de maneira mais completa a diversidade presente nos estilos musicais analisados.

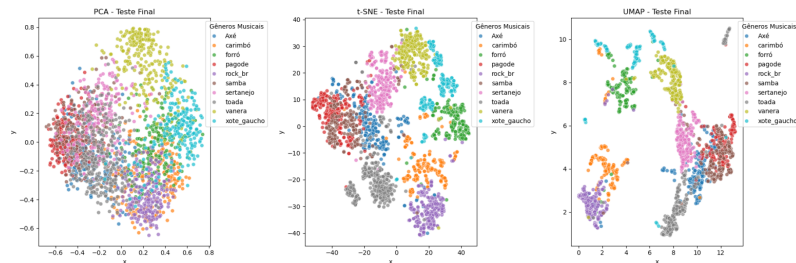
5.5. Visualização e Análise dos Embeddings no Espaço Vetorial

Com o objetivo de investigar como o modelo ViT organiza os dados musicais em seu espaço latente, foram aplicadas técnicas de redução de dimensionalidade para projetar os *embeddings* gerados pelo modelo em duas dimensões. Os métodos utilizados foram PCA, t-SNE e UMAP, sendo essa análise realizada tanto para o trecho de melhor desempenho (90–120s) quanto para o pior (0–30s), em três durações de segmento: 3s, 5s e 10s. Ainda que os segmentos e trechos já tenham sido previamente analisados em termos de desempenho, optou-se por aprofundar a investigação visual nesses dois extremos. A escolha permite avaliar como a posição temporal da amostra e a duração do segmento influenciam a separabilidade dos gêneros no espaço vetorial aprendido, complementando as análises quantitativas com uma perspectiva qualitativa. As projeções dos embeddings referentes ao trecho 90–120s, considerado o melhor segmento global, permitem observar como diferentes durações de segmentos influenciam na organização espacial dos gêneros musicais antes e após o treinamento do ViT. A análise considera três métodos de redução de dimensionalidade: PCA, t-SNE e UMAP.

Na Figura 5.13, com segmentos de 3 segundos, nota-se um alto número de instâncias, decorrente da granularidade da segmentação. Antes do treinamento, os dados encontram-se dispersos no espaço vetorial, especialmente nas projeções de t-SNE e UMAP, com sobreposições consideráveis entre os gêneros. Após o treinamento, a visualização via UMAP ainda revela sobreposições marcantes entre Samba, Pagode e Axé, com esses três estilos também aparecendo relativamente próximos de Toada e Sertanejo. No t-SNE, há um maior distanciamento entre os grupos, embora ainda se perceba a formação de um núcleo que inclui Sertanejo, Samba, Pagode e Axé. Vaneira aparece ligeiramente afastada, seguida de Xote Gaúcho e Forró, enquanto Carimbó e Rock BR ocupam regiões periféricas, com algumas instâncias sobrepostas entre si e com Forró.



((a)) Antes do treinamento, com segmentos de 3 segundos.

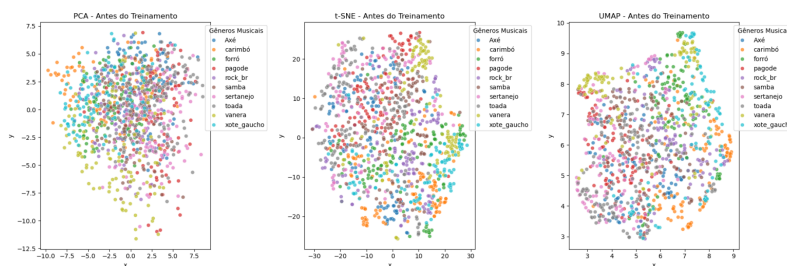


((b)) Após o treinamento, conjunto de teste com segmentos de 3 segundos.

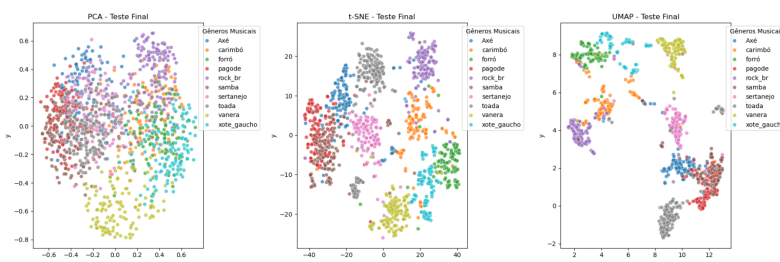
Figura 5.13. Projeções dos embeddings do trecho 90–120s com segmentos de 3s, antes e após o treinamento do ViT. Cada subfigura apresenta as representações via PCA, t-SNE e UMAP.

Com segmentos de 5 segundos (Figura 5.14), Vaneira se torna um *cluster* bem isolado, enquanto Xote Gaúcho aparece mais disperso e próximo de Forró. Carimbó mostra maior dispersão e sobreposição com vários outros gêneros, apresentando muitos *outliers*. Samba e Pagode mantêm alta sobreposição, e Sertanejo se distancia um pouco de Toada, sugerindo maior distinção estilística com a ampliação do segmento.

Na Figura 5.15, com segmentos de 10 segundos, os agrupamentos se tornam mais



((a)) Antes do treinamento, com segmentos de 5 segundos.

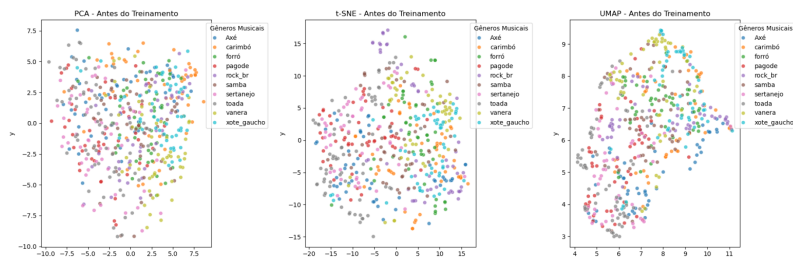


((b)) Após o treinamento, conjunto de teste com segmentos de 5 segundos.

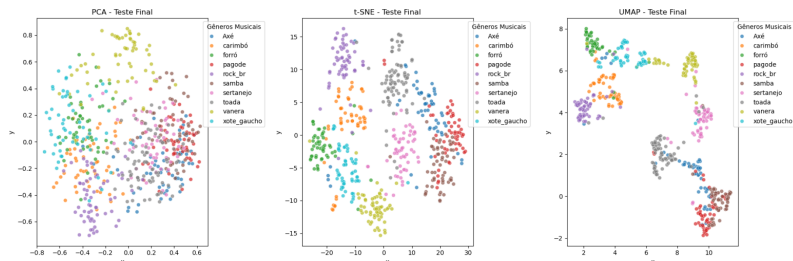
Figura 5.14. Projeções dos embeddings do trecho 90–120s com segmentos de 5s, antes e após o treinamento do ViT. Cada subfigura apresenta as representações via PCA, t-SNE e UMAP.

compactos devido à redução do número de instâncias. Vaneira aparece próxima de Sertanejo, enquanto Samba e Pagode seguem sobrepostos. Axé se mistura com esses dois e também com Toada. Carimbó se torna menos disperso, mas ainda se sobrepõe a Forró e Rock BR. Algumas instâncias de Axé também aparecem nas regiões de Forró.

A comparação entre as três figuras evidencia um equilíbrio delicado entre a duração do segmento e o número de instâncias. Segmentos mais curtos, como os de 3 segundos, oferecem uma maior quantidade de dados e promovem uma diversidade estatística mais ampla, mas podem dificultar a formação de representações vetoriais estáveis. Por outro lado, segmentos mais longos ampliam o conteúdo musical analisado por instância, o que tende a favorecer a construção de *embeddings* mais completos e informativos, embora à custa de uma menor quantidade de exemplos disponíveis, o que pode comprometer a representatividade estatística da base. Em todos os cenários analisados, mesmo após o treinamento, observam-se regiões de sobreposição vetorial entre alguns gêneros específicos. A sobreposição entre Samba e Pagode, assim como a presença do Axé distribuído entre esses dois gêneros, aparece com frequência nas projeções. Da mesma forma, há proximidade significativa entre Xote Gaúcho, Forró e Vaneira, com regiões parcialmente



((a)) Antes do treinamento, com segmentos de 10 segundos.

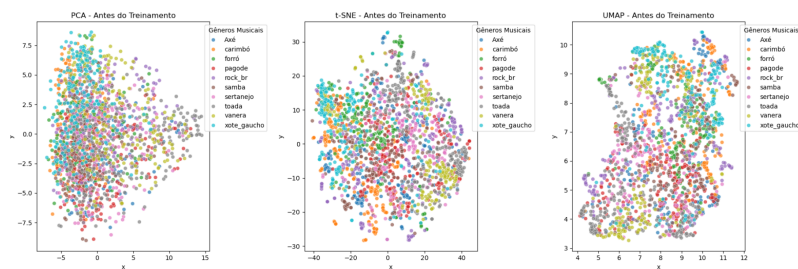


((b)) Após o treinamento, conjunto de teste com segmentos de 10 segundos.

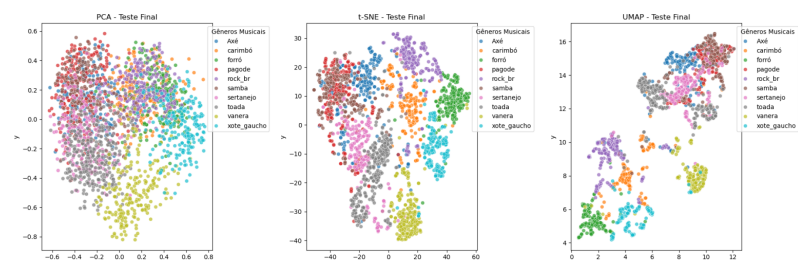
Figura 5.15. Projeções dos embeddings do trecho 90–120s com segmentos de 10s, antes e após o treinamento do ViT. Cada subfigura apresenta as representações via PCA, t-SNE e UMAP.

compartilhadas entre os três estilos. Uma análise qualitativa desses padrões sugere que tais sobreposições não se devem apenas a limitações do modelo ou à dimensionalidade, mas possivelmente refletem semelhanças estruturais e estilísticas reais entre os gêneros musicais. Esses agrupamentos podem estar relacionados a aspectos como instrumentação, ritmo ou textura sonora comuns, especialmente no caso de gêneros com raízes culturais próximas ou influências mútuas.

Embora as projeções dos embeddings do melhor trecho global já mostrem bastante sobre a organização espacial dos gêneros musicais, também é interessante apresentar os resultados obtidos no pior trecho global, com o objetivo de aprofundar a análise comparativa. A inclusão desse trecho, identificado anteriormente como o de pior desempenho, permite observar como a posição temporal impacta a qualidade das representações vetoriais geradas pelo ViT. Na Figura 5.16, Samba, Pagode, Axé e Sertanejo seguem próximos, com Sertanejo ainda mais misturado aos demais do que no melhor trecho. Vaneira aparece mais distante e isolada, enquanto Xote Gaúcho está mais concentrado e próximo de Carimbó. Rock BR e Carimbó têm instâncias sobrepostas, como também observado no melhor trecho.



((a)) Antes do treinamento, com segmentos de 3 segundos.



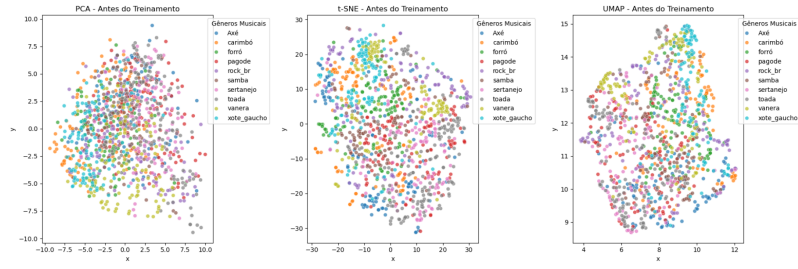
((b)) Após o treinamento, conjunto de teste com segmentos de 3 segundos.

Figura 5.16. Projeções dos embeddings do pior trecho 0–30s com segmentos de 3s, antes e após o treinamento do ViT. Cada subfigura apresenta as representações via PCA, t-SNE e UMAP.

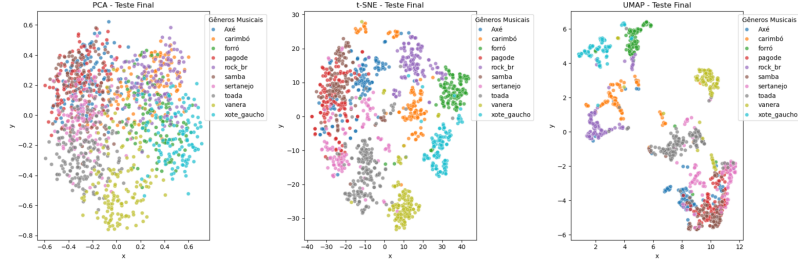
Na Figura 5.17, a sobreposição entre Sertanejo, Axé, Samba e Pagode se intensifica. Rock BR apresenta instâncias próximas de Forró e Carimbó. Xote Gaúcho está mais próximo de Forró e Vaneira se mantém como um grupo isolado. A presença de outliers é frequente em quase todos os gêneros, o que também ocorre no melhor trecho com 5 segundos, mas lá os agrupamentos são mais coesos.

Por fim, na Figura 5.18, a visualização mostra alta sobreposição entre Pagode e Samba. Vaneira forma um *cluster* mais denso e aparece próxima de Sertanejo e Toada. Axé mostra instâncias que se sobrepõem a Pagode, Samba, Toada e, de forma mais inesperada, a Rock BR. Xote Gaúcho está mais próximo de Carimbó e Forró.

Observando esses *embeddings*, é perceptível que mesmo com a mudança do trecho da música e a variação na duração dos segmentos, observam-se padrões recorrentes de proximidade e sobreposição entre determinados gêneros musicais. Samba e Pagode aparecem frequentemente sobrepostos, mantendo-se próximos em todos os cenários analisados. Axé também tende a ocupar a mesma região do espaço vetorial desses dois gêneros, formando um núcleo comum que o modelo tem dificuldade em separar, possivelmente

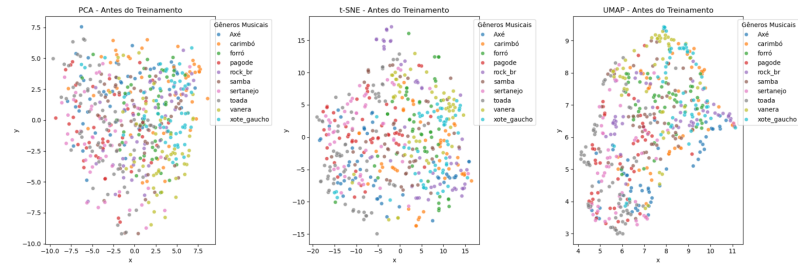


((a)) Antes do treinamento, com segmentos de 5 segundos.



((b)) Após o treinamento, conjunto de teste com segmentos de 5 segundos.

Figura 5.17. Projeções dos embeddings do pior trecho 0–30s com segmentos de 5s, antes e após o treinamento do ViT. Cada subfigura apresenta as representações via PCA, t-SNE e UMAP.



((a)) Antes do treinamento, com segmentos de 10 segundos.



((b)) Após o treinamento, conjunto de teste com segmentos de 10 segundos.

Figura 5.18. Projeções dos embeddings do pior trecho 0–30s com segmentos de 10s, antes e após o treinamento do ViT. Cada subfigura apresenta as representações via PCA, t-SNE e UMAP.

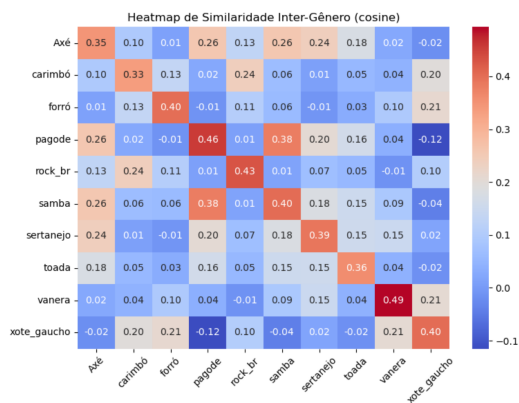
por compartilharem elementos como ritmo, percussão e estilo vocal. Em contrapartida, gêneros como Rock BR se posicionam mais distantes desses agrupamentos, sugerindo que o modelo consegue reconhecer sua maior distinção sonora.

Também é comum encontrar Xote Gaúcho, Forró e Carimbó em regiões próximas no espaço vetorial, reforçando características acústicas ou estruturais semelhantes entre esses estilos regionais. Por outro lado, em algumas projeções, ocorrem sobreposições inesperadas entre gêneros que, à primeira vista, parecem distintos, como Rock BR e Carimbó. Essas aproximações, embora menos intuitivas, indicam que o espaço latente capturado pelo ViT pode refletir relações acústicas que não são evidentes apenas pela categorização tradicional dos gêneros. Por isso, é interessante avaliar esses padrões qualitativos com as técnicas de redução de dimensionalidade em conjunto com os resultados quantitativos, pois eles enriquecem a compreensão sobre o comportamento do modelo e sobre as características compartilhadas entre diferentes gêneros musicais.

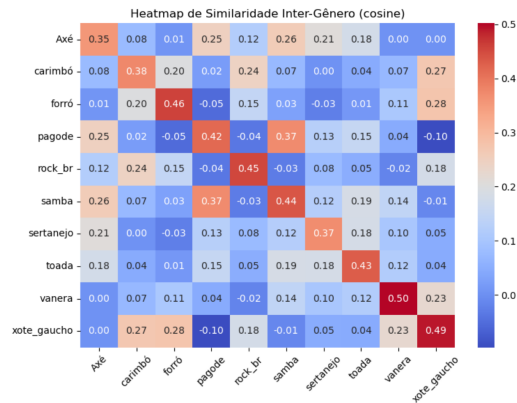
5.6. Análise de Similaridade entre Gêneros Musicais

Embora técnicas de redução de dimensionalidade como PCA, t-SNE e UMAP sejam úteis para fins de visualização, elas distorcem as distâncias originais entre os vetores e, portanto, não devem ser utilizadas para cálculos quantitativos de similaridade. Para uma análise mais precisa da proximidade entre os gêneros musicais, foi utilizada a versão original dos *embeddings* gerados pelo Vision Transformer (ViT) treinado com 768 dimensões. Para cada gênero, foi calculado o vetor médio de *embedding* a partir dos segmentos de 3, 5 e 10 segundos do conjunto de teste. Em seguida, aplicou-se a similaridade de cosseno entre todos os pares de vetores médios, gerando uma matriz de similaridade que reflete como o modelo relaciona internamente os diferentes gêneros no espaço vetorial. A Figura 5.19 apresenta essa matriz como um mapa de calor para o melhor trecho global de 90–120s. Valores mais próximos de 1 indicam alta similaridade entre os gêneros, valores próximos de 0 sugerem ausência de correlação vetorial, e valores negativos indicam oposição no espaço latente, ou seja, o modelo reconhece padrões que apontam em direções contrastantes.

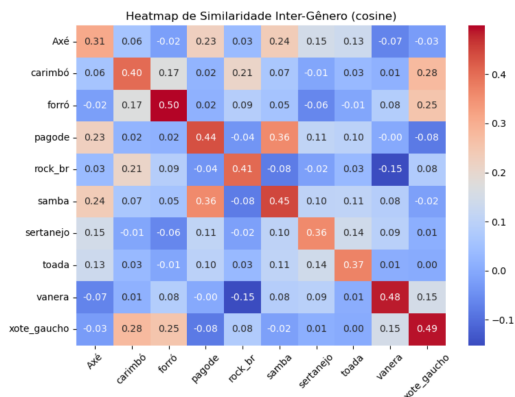
A matriz gerada a partir dos segmentos de 3 segundos, representada na Figura 5.19(a), revela agrupamentos vetoriais coerentes com as observações qualitativas feitas anteriormente. Um exemplo marcante é o par samba e pagode, que aparece consistentemente com



((a)) Segmentos de 3 segundos



((b)) Segmentos de 5 segundos



((c)) Segmentos de 10 segundos

Figura 5.19. Mapas de calor da similaridade de cosseno entre os gêneros musicais com embeddings do trecho 90–120s, considerando diferentes durações de segmentos.

alta similaridade, o que reforça tanto a análise visual dos embeddings quanto o entendimento musical sobre a proximidade rítmica, harmônica e cultural entre os dois estilos. Vaneira e xote gaúcho também formam um par fortemente relacionado, evidenciando afinidades regionais e instrumentais, especialmente com relação ao uso do acordeão. Forró aparece relacionado ao xote gaúcho e ao carimbó, o que pode refletir a presença de padrões percussivos e melódicos compartilhados.

O axé, por sua vez, apresenta uma postura vetorial híbrida, mantendo proximidade com estilos como samba, pagode, sertanejo e toada. Essa distribuição confirma sua natureza de fusão, combinando elementos de diferentes tradições musicais brasileiras. O carimbó se destaca por estabelecer relações não apenas com gêneros próximos geograficamente, como o forró, mas também com o Rock BR, uma curiosidade que já havia sido notada nas visualizações qualitativas. De modo geral, esses agrupamentos indicam que o ViT foi capaz de identificar estruturas sonoras compartilhadas entre os gêneros, mesmo em segmentos curtos.

Por outro lado, a matriz também evidencia pares com separação vetorial considerável. Gêneros como pagode e xote gaúcho, ou forró e sertanejo, apresentam relações negativas ou próximas de zero, indicando que o modelo aprendeu representações significativamente distintas para esses estilos. Vaneira se distancia de gêneros como axé, toada e carimbó, enquanto o xote gaúcho mantém separação clara em relação a estilos como axé, samba, pagode e toada. Esses contrastes vetoriais reforçam a capacidade discriminativa do modelo diante da diversidade estilística da base BYRM.

A matriz de similaridade com segmentos de 5 segundos, apresentada na Figura 5.19(b), mostra padrões bastante semelhantes aos encontrados com 3 segundos, mas com intensificações relevantes. Relações como samba e pagode permanecem fortemente presentes, assim como a associação entre axé e os gêneros que possuíam forte reação no segmento de 3 segundos. Algumas conexões, como carimbó com xote gaúcho e com forró, tornam-se ainda mais evidentes, sugerindo que segmentos ligeiramente mais longos permitem ao modelo capturar com mais clareza elementos rítmicos e harmônicos compartilhados. O xote gaúcho, por exemplo, apresenta maior aproximação com esses dois estilos neste cenário. O rock brasileiro mantém sua posição vetorial próxima ao carimbó, assim como na matriz anterior.

Com segmentos de 10 segundos, representados na Figura 5.19(c), observa-se a continuidade da alta similaridade entre samba e pagode e a consistência de forró com carimbó. No entanto, começam a surgir nuances importantes, como pode ser visto com o Axé, que apresenta reduções em sua similaridade com outros gêneros, o que pode indicar que, com janelas mais longas, o modelo consegue representar de forma mais precisa suas particularidades acústicas, tornando sua representação vetorial mais distinta. Além disso, a matriz apresenta uma distribuição mais visivelmente polarizada, com maior presença de tons frios no mapa de calor, indicando aumento nos valores baixos e negativos de similaridade, um sinal de que segmentos mais longos ajudam o modelo a separar melhor os estilos musicais mais contrastantes.

Por fim, uma análise intra-gênero observando a diagonal principal das matrizes revela que nenhum dos gêneros atinge valores de similaridade interna próximos de 1. O valor máximo encontrado foi de aproximadamente 0.5, o que indica que mesmo entre segmentos do mesmo gênero, há variações consideráveis no espaço vetorial. Esse comportamento é esperado, dado o caráter diversificado dos gêneros musicais brasileiros, que frequentemente misturam influências estilísticas e apresentam grande diversidade interna. Em outras palavras, os valores moderados na diagonal não enfraquecem a base, mas evidenciam sua riqueza e realismo. Gêneros como forró, samba e carimbó apresentaram seus maiores níveis de similaridade interna nos segmentos de 10 segundos, enquanto pagode e sertanejo mostraram maior coesão com segmentos curtos, e estilos como rock brasileiro, toada vaneira foram melhor representados com janelas de 5 segundos. Axé manteve seus níveis de coesão interna relativamente estáveis em 3s e 5s, e o xote gaúcho obteve melhores representações internas com 5s e 10s.

Esses resultados mostram que o modelo não apenas classifica, mas também organiza os gêneros musicais de forma significativa em seu espaço latente, permitindo interpretações musicais coerentes com a estrutura da base. A análise de similaridade, portanto, não apenas valida os agrupamentos qualitativos observados nas projeções dos *embeddings*, mas também reforça o potencial dos modelos baseados em atenção para capturar nuances estilísticas em dados musicais.

5.7. Discussão dos Resultados

Nesta seção, são discutidos os principais resultados obtidos nos experimentos, com ênfase em dois aspectos centrais: a influência da posição temporal dos segmentos no desempenho do modelo e a organização dos gêneros musicais no espaço vetorial de embeddings. A análise temporal permite compreender quais trechos da música são mais representativos para a tarefa de classificação, considerando também o impacto da duração dos segmentos. Em seguida, a análise vetorial contribui para entender como os gêneros se distribuem no espaço de características e quais apresentam maior sobreposição.

5.7.1. Temporalidade

Os experimentos realizados indicaram que o desempenho do modelo ViT na tarefa de classificação de gêneros musicais variou de forma significativa em função da posição temporal do trecho analisado e da duração dos segmentos utilizados. A Tabela 5.5 apresenta, para cada gênero, o melhor e o pior trecho de 30 segundos ao longo da faixa musical, com os respectivos valores de F1-score.

Tabela 5.5. Melhor e pior trecho por gênero, com F1-score máximo e mínimo.

Gênero	Melhor Trecho (30s)	F1-score Máximo	Pior Trecho (30s)	F1-score Mínimo
Forró	0–30s	87,78%	30–60s	82,91%
Axé	30–60s	71,49%	150–180s	55,91%
Samba	30–60s	65,75%	90–120s	56,46%
Carimbó	60–90s	80,37%	150–180s	66,45%
Rock BR	60–90s	87,44%	0–30s	79,25%
Vaneira	60–90s	91,30%	150–180s	86,15%
Sertanejo	90–120s	85,29%	0–30s	73,48%
Xote Gaúcho	90–120s	82,99%	30–60s	77,61%
Pagode	120–150s	70,80%	0–30s	52,49%
Toada	120–150s	89,58%	0–30s	83,68%

As definições foram feitas com base no F1-score, adotado como métrica principal nesta pesquisa por harmonizar precisão e revocação, fornecendo uma medida mais confiável em contextos de classificação multiclasse e sujeitos a desbalanceamentos entre classes. Essa escolha é especialmente relevante considerando a presença de sobreposições estilísticas entre alguns gêneros regionais brasileiros.

A análise da Tabela 5.5 revela que, embora o trecho de 90–120 segundos tenha

apresentado o maior desempenho global, com os melhores resultados médios considerando todos os gêneros, ele não foi o mais recorrente entre os melhores trechos individuais. O intervalo de 60–90 segundos apareceu com maior frequência como o melhor trecho por gênero para Carimbó, Rock BR, Vaneira. Por outro lado, o trecho de 0–30 segundos concentrou a maior parte dos piores desempenhos para os gêneros Rock BR, Sertanejo, Pagode, Toada. Ao observar os gêneros individualmente, nota-se que o Vaneira e o Rock BR alcançaram alto desempenho no trecho de 60–90 segundos, o que pode indicar que suas estruturas musicais apresentam elementos centrais nesse ponto da faixa, como refrões ou progressões harmônicas marcantes. O Carimbó também se destacou nesse trecho, sugerindo um padrão semelhante de concentração rítmica ou melódica.

Foi possível observar que gêneros como Sertanejo e Xote Gaúcho apresentaram melhor desempenho no trecho de 90–120 segundos, possivelmente devido à consolidação de temas melódicos e menor variação harmônica nesse ponto. O Forró, teve seu melhor resultado no início da faixa (0–30s), demonstrando que certos estilos podem apresentar sua identidade sonora já nos primeiros segundos. Em contraste, Pagode e Toada obtiveram os piores desempenhos nesse mesmo trecho inicial, evidenciando que a introdução musical nem sempre carrega informações suficientes para uma boa classificação automática.

No extremo oposto, o trecho de 150–180 segundos concentrou os piores resultados de gêneros como Axé, Carimbó e Vaneira. Esse desempenho inferior pode estar associado a dois fatores principais. Primeiro, esse intervalo temporal apresentou desbalanceamento na distribuição de amostras, uma vez que nem todas as faixas da base alcançam três minutos de duração, o que reduz a representatividade estatística desse trecho nos dados de treinamento e teste. Segundo, é possível que, em algumas músicas que efetivamente alcançam essa marca, o conteúdo corresponda a seções finais com repetições instrumentais, ou variações menos características, comprometendo a expressividade dos padrões musicais e dificultando a identificação do gênero. Esses resultados reforçam a importância de adotar estratégias de segmentação sensíveis à posição temporal, uma vez que a identidade de gênero não é distribuída de forma uniforme ao longo das faixas.

Complementarmente à análise da posição do trecho, investigou-se também o impacto da duração dos segmentos na performance do modelo ViT. Segmentos de 3, 5 e 10 segundos foram testados nos trechos de melhor (90–120s) e pior desempenho global

(0–30s). Os resultados revelam que segmentos mais longos tendem a favorecer a classificação no trecho de maior desempenho, enquanto janelas menores foram ligeiramente superiores no trecho inicial.

Especificamente, no intervalo de 90–120 segundos, o modelo atingiu o maior F1-score com segmentos de 10 segundos (81,84%), superando as versões com 3s e 5s. Isso indica que janelas mais amplas são capazes de capturar padrões musicais mais completos e contextuais, como refrões, modulações e variações instrumentais que caracterizam o gênero. Esse efeito é particularmente evidente em gêneros como Vaneira, Carimbó e Rock BR, que apresentaram crescimento contínuo de desempenho com o aumento da janela. Em contrapartida, no trecho de 0–30 segundos, o melhor resultado foi obtido com segmentos de 3 segundos (F1-score de 74,91%), embora a diferença para 5s e 10s tenha sido pequena. Gêneros como Forró, Toada e Xote Gaúcho se destacaram mesmo com janelas curtas, o que sugere a presença de elementos rítmicos marcantes logo no início da faixa, capazes de sustentar a classificação com menor contexto temporal.

Por fim, observou-se que janelas de 5 segundos nem sempre representaram um ponto de equilíbrio. Em alguns casos, como no Sertanejo e no Carimbó, essa configuração intermediária apresentou desempenho inferior a segmentos curtos ou longos, indicando que trechos menos informativos podem coincidir com essa janela, dependendo da estrutura interna da música. Essas evidências sugerem que uma abordagem dinâmica e adaptativa de segmentação, que considere tanto o estilo musical quanto a localização temporal do trecho, pode melhorar significativamente os sistemas de reconhecimento automático. A exploração futura de técnicas baseadas em análise estrutural musical ou segmentação multirresolução pode contribuir para representações mais fiéis e eficazes dos gêneros analisados.

5.7.2. Espaço Vetorial e Similaridade

Após as análises segmentadas ao longo do capítulo, esta seção retoma os principais padrões observados no espaço vetorial de *embeddings*, sintetizando as relações de similaridade aprendidas pelo modelo ViT entre os diferentes gêneros musicais. A seguir, destacam-se agrupamentos, sobreposições e separações identificadas tanto visual quanto quantitativamente:

- **Gêneros com alta proximidade visual e vetorial:** Samba e pagode formam clus-

ters densos e sobrepostos em todas as projeções, e apresentam alta similaridade vetorial, o que é coerente com sua afinidade rítmica, histórica e harmônica. Vaneira e xote gaúcho também demonstram alinhamento consistente, com as maiores similaridades absolutas nas matrizes de cosseno, mesmo que em algumas projeções 2D apareçam parcialmente afastados. Esses resultados indicam que o ViT consegue capturar padrões estruturais mesmo com variações de posicionamento visual.

- **Agrupamentos regionais com sobreposições parciais:** Axé se apresenta como um gênero limítrofe, ocupando uma posição intermediária entre samba, pagode, sertanejo e toada, com diferentes níveis de proximidade vetorial. Essa configuração reflete sua natureza híbrida, marcada pela fusão de ritmos afro-brasileiros com elementos pop. O carimbó, por sua vez, mostra-se relativamente disperso no espaço vetorial, apresentando sobreposições com forró e, curiosamente, com o rock brasileiro. Esse comportamento sugere uma diversidade tímbrica que pode gerar representações vetoriais ambíguas ou múltiplas.
- **Distanciamentos e contrastes:** Rock brasileiro emerge como o gênero mais isolado no espaço vetorial, com baixa similaridade em relação à maioria dos outros estilos. Esse afastamento pode ser atribuído à predominância de instrumentos elétricos, distorções e estruturas harmônicas distintas, características pouco frequentes nos gêneros regionais. Vaneira também mantém distâncias vetoriais relevantes de estilos como axé, samba e toada, reforçando sua identidade regional própria. Em alguns casos, gêneros como toada e sertanejo, que compartilham elementos melódicos e temáticos, apresentaram certa aproximação em janelas curtas, mas se distanciaram com o aumento da duração dos segmentos. Isso indica que segmentos mais longos ajudam o modelo a capturar nuances que diferenciam os estilos com mais clareza.

Essas observações reforçam que o modelo ViT não apenas atinge bons índices de classificação, mas também organiza os gêneros de maneira coerente com suas características musicais. A análise vetorial, combinando visualizações e medidas de similaridade de cosseno, evidencia como os estilos regionais se agrupam, se separam ou se sobrepõem dentro do espaço latente aprendido, permitindo inferências interpretáveis sobre a natureza da música representada na base BYRM.

Capítulo 6

Considerações Finais

Este trabalho investigou o uso da arquitetura de transformadores de visão computacional para o reconhecimento de gêneros musicais, com foco na música regional brasileira. A partir da construção da base BYRM, foram conduzidos experimentos que buscaram compreender não apenas o desempenho do modelo em termos de acurácia, mas também como ele organiza os estilos musicais no espaço vetorial. O modelo ViT mostrou-se capaz de lidar com a diversidade estilística dos dados, alcançando bons resultados na classificação de gêneros regionais e produzindo representações vetoriais interpretáveis de forma qualitativa.

A metodologia adotada foi baseada na avaliação das faixas musicais, considerando diferentes trechos temporais ao longo da música e diferentes durações de segmento (3s, 5s e 10s). Essa abordagem permitiu investigar quais partes das músicas são mais representativas para a tarefa de classificação. Os experimentos demonstraram que segmentos mais longos tendem a favorecer a separação entre estilos musicais no espaço vetorial, e que trechos centrais da música são os que melhor capturam os elementos estilísticos característicos. Como resultado no melhor trecho identificado, o modelo ViT alcançou acurácia de 81,94%, precisão de 82,85%, revocação de 81,94% e F1-score de 81,84%, evidenciando sua eficácia na identificação de padrões musicais mais informativos.

Além da avaliação quantitativa, foi conduzida uma análise qualitativa dos *embeddings* gerados, com projeções em duas dimensões (PCA, t-SNE e UMAP), revelando agrupamentos e sobreposições entre os gêneros. Essas visualizações foram complementadas por uma análise de similaridade vetorial com base na distância de cosseno entre

os vetores médios de cada gênero, evidenciando padrões consistentes: samba e pagode aparecem constantemente agrupados, vaneira e xote gaúcho formam pares fortemente relacionados, enquanto rock brasileiro tende a ocupar uma região mais isolada no espaço vetorial.

Apesar dos avanços alcançados, algumas limitações e oportunidades de aprimoramento foram identificadas ao longo do trabalho. Uma delas refere-se ao treinamento da ResNet50, que poderia ter sido mais cuidadosamente ajustada, a fim de se posicionar como um concorrente mais equilibrado em relação ao ViT, permitindo comparações mais justas entre as arquiteturas. Além disso, não foram aplicadas técnicas de aumento de dados nos espectrogramas, recurso que poderia ter contribuído para o aumento da acurácia geral e para a robustez do treinamento.

Outra limitação diz respeito à ausência de uma análise quantitativa direta dos *embeddings* reduzidos. Embora as projeções visuais tenham revelado padrões relevantes, a incorporação de métricas objetivas de agrupamento ou separabilidade poderia ter complementado essas observações com maior rigor analítico. Por fim, um dos principais desafios enfrentados foi a interpretação dos gêneros no espaço vetorial e nas matrizes de similaridade. Compreender de que forma essas representações traduzem nuances musicais nem sempre se mostrou trivial, exigindo uma abordagem multidisciplinar entre os campos da computação e da música.

Como contribuição direta desta pesquisa, destaca-se a publicação da base de dados BYRM, disponibilizada no repositório Zenodo [Guimarães and Freitas 2025]. A base contém mel-espectrogramas, vetores de características acústicas extraídos das faixas e metadados completos, incluindo os respectivos IDs dos vídeos utilizados no YouTube, possibilitando a reprodutibilidade dos experimentos.

Além disso, parte dos resultados obtidos foi aceita no Symposium on Knowledge Discovery, Mining and Learning (KDMiLe 2025), evidenciando o reconhecimento da proposta por pares da área [Guimarães et al. 2025a]. Nessa publicação, foram apresentadas tanto a construção da BYRM quanto a estratégia de segmentação temporal adotada para a avaliação dos modelos. Outra parte dos resultados foi também aceita no Simpósio Brasileiro de Computação Musical (SBCM 2025), onde foram discutidas as análises de similaridade entre gêneros musicais a partir dos *embeddings* extraídos pelo modelo Vision

Transformer, aprofundando a compreensão das relações entre estilos musicais regionais brasileiros [Guimarães et al. 2025b].

Além das contribuições diretamente associadas ao desenvolvimento desta dissertação, também se destacam pesquisas prévias realizadas em qualidade perceptual de áudio, que serviram como base conceitual e metodológica para o amadurecimento desta investigação. Dessa forma, o estudo sobre a detecção perceptual de artefatos induzidos por perda de pacotes em sistemas de áudio sem fio foi aceito e apresentado no Detection and Classification of Acoustic Scenes and Events (DCASE 2025), demonstrando domínio prévio em tarefas de processamento de sinais e classificação de áudio utilizando modelos de aprendizado de máquina supervisionados [Guimarães et al. 2025].

Trabalhos futuros incluem a aplicação dos modelos SVM e ResNet50 em todos os segmentos temporais da base BYRM, e não apenas nos trechos com melhor e pior desempenho. Essa ampliação poderá proporcionar uma avaliação mais abrangente da consistência do ViT em diferentes contextos e faixas temporais, viabilizando comparações mais robustas entre arquiteturas. Recomenda-se também a realização de execuções múltiplas de cada experimento, com diferentes sementes de aleatoriedade, a fim de obter resultados mais estáveis e confiáveis. Essa abordagem permitiria a aplicação de testes estatísticos sobre as métricas de desempenho, contribuindo para reforçar a validade das conclusões.

Uma linha de investigação motivada pelos resultados obtidos pelo SVM neste trabalho é a realização de fusão entre características extraídas manualmente e representações aprendidas automaticamente pelo ViT. Essa abordagem envolveria a combinação de descritores acústicos tradicionais (como MFCCs, cromagrama e ZCR) com *embeddings* gerados pelo Transformer. A integração dos modelos pode potencializar o desempenho da classificação, sobretudo em gêneros musicais específicos nos quais o SVM apresentou acertos que o ViT não conseguiu capturar. Dessa forma, a fusão entre ambos pode explorar forças distintas de cada abordagem, resultando em melhorias localizadas na classificação.

Outra vertente relevante para estudos futuros é a expansão da base BYRM com a inclusão de novos gêneros musicais brasileiros, o que tende a aumentar sua diversidade e complexidade. Isso permitiria investigações mais aprofundadas sobre o comportamento dos modelos diante de estilos com características sonoras semelhantes, reforçando o foco

na análise da organização vetorial da música regional. Os resultados obtidos até o momento também indicam caminhos para análises mais direcionadas por trecho e por segmento musical, aprofundando o entendimento da estrutura temporal na representação dos gêneros.

Referências

- Abdoli, S. (2011). Iranian traditional music dastgah classification. In *ISMIR*, pages 275–280.
- Armijo, C. and Rivas, P. (2025). Exploring visual embedding spaces induced by vision transformers for online auto parts marketplaces. *arXiv preprint arXiv:2502.05756*.
- Aucouturier, J.-J. and Pampalk, E. (2008). Introduction—from genres to tags: A little epistemology of music information retrieval research. *Journal of New Music Research*, 37(2):87–92.
- Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Barbosa, J., McKay, C., and Fujinaga, I. (2015). Evaluating automated classification techniques for folk music genres from the brazilian northeast. *Computer Music: Beyond the frontiers of signal processing and computational models*.
- Barnett, J., Garcia, H. F., and Pardo, B. (2024). Exploring musical roots: Applying audio embeddings to empower influence attribution for a generative music model. *arXiv preprint arXiv:2401.14542*.
- Basheer, I. A. and Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design, and application. *Journal of microbiological methods*, 43(1):3–31.
- Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. (2011). The million song dataset.
- Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.
- Byron, T. P., Rushworth, C. T., and Stewart, M. J. (2025). 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.
- Cassano, L., Cassano, E. d. O. M. N., and Pechman, R. M. (2024). Passado e presente da música carioca: Narrativas urbanas e evolução tecnológica. *DEBATES-Cadernos do Programa de Pós-Graduação em Música*, 28:e282408–e282408.

- Castro, A. A. (2010). Axé music: mitos, verdades e world music. *Per Musi*, pages 203–217.
- Cerati, G. (2021). Difficult to define, easy to understand: the use of genre categories while talking about music. *SN Social Sciences*, 1(12):288.
- Cheng, Y.-H., Chang, P.-C., and Kuo, C.-N. (2020). Convolutional neural networks approach for music genre classification. In *2020 International Symposium on Computer, Consumer and Control (IS3C)*, pages 399–403. IEEE.
- Colliot, O. (2023). Machine learning for brain disorders.
- Conceição, J. L., de Freitas, R., Gadelha, B., Kienen, J. G., Anders, S., and Cavalcante, B. (2020). Applying supervised learning techniques to brazilian music genre classification. In *2020 XLVI Latin American Computing Conference (CLEI)*, pages 102–107. IEEE.
- Correa, M. G. (2018). 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).
- da Silva, A. C. M., Silva, D. F., and Marcacini, R. M. (2020). 4mula: A multitask, multi-modal, and multilingual dataset of music lyrics and audio features. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, pages 145–148.
- Dai, J., Liang, S., Xue, W., Ni, C., and Liu, W. (2016). 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)*, pages 1–5. IEEE.
- de Araújo Lima, R., de Sousa, R. C. C., Lopes, H., and Barbosa, S. D. J. (2020). Brazilian lyrics-based music genre classification using a blstm network. In *International Conference on Artificial Intelligence and Soft Computing*, pages 525–534. Springer.
- De Sousa, J. M., Pereira, E. T., and Veloso, L. R. (2016). 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.
- Defferrard, M., Benzi, K., Vandergheynst, P., and Bresson, X. (2016). Fma: A dataset for music analysis. *arXiv preprint arXiv:1612.01840*.

- Dieleman, S. and Schrauwen, B. (2014). End-to-end learning for music audio. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6964–6968. IEEE.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Downie, J. S. (2003). Music information retrieval. *Annual review of information science and technology*, 37(1):295–340.
- Esposito, A., Faundez-Zanuy, M., Morabito, F. C., Pasero, E., et al. (2021). *Progresses in Artificial Intelligence and Neural Systems*. Springer.
- Fabbri, F. et al. (1982). A theory of musical genres. two applications. In *Popular music perspectives*, volume 1, pages 52–81. Iaspm.
- Fayyad, U. (1997). Knowledge discovery in databases: An overview. In *International Conference on Inductive Logic Programming*, pages 1–16. Springer.
- Furqon, M. N., Khadijah, K., Suhartono, S., and Kusumaningrum, R. (2019). Indonesian music genre classification on indonesian regional songs using deep recurrent neural network method. In *2019 3rd International Conference on Informatics and Computational Sciences (ICICoS)*, pages 1–5. IEEE.
- Gold, B., Morgan, N., and Ellis, D. (2011). *Speech and audio signal processing: processing and perception of speech and music*. John Wiley & Sons.
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.
- Guimarães, V., Kienen, J. G., and de Freitas, R. (2025a). Segment-based evaluation of music genre classification models with the byrm dataset. In *Symposium on Knowledge Discovery, Mining and Learning (KDMiLe)*, pages 97–104. SBC.
- Guimarães, V., Kienen, J. G., and de Freitas, R. (2025b). Understanding genre similarity in brazilian music through vision transformer embeddings. In *Brazilian Symposium on Computer Music (SBCM)*. SBC. Presented; in press.
- Guimarães, V., Bentes, L., Pires, A., and Freitas, R. (2025). Perceptual detection of packet loss-induced audio artifacts in black-box wireless music systems. In *Proceedings of the*

- 10th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2025)*, pages 215–219, Barcelona, Spain.
- Guimarães, V. d. S. and Freitas, R. d. (2025). Byrm: Brazilian youtube regional music dataset.
- Haykin, S. (1994). *Neural networks: A comprehensive foundation*, macmillan college publishing co. *New York*.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558.
- ISMIR (2004). Ismir 2004 genre classification contest. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*, Barcelona, Spain.
- Junior, J. (2006). Música popular massiva e gêneros musicais: produção e consumo da canção na mídia. *Comunicação Mídia e Consumo*, 3(7):31–47.
- Krogh, M. (2019). Music-genre abstraction, dissemination and trajectories. *Dansk Musikforskning Online*, 9:83–108.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.
- Luo, X., Hoveling, V., and Eisemann, E. (2024). Musicon: Glyph-based design for music visualization and retrieval.
- Makris, D., Karydis, I., and Sioutas, S. (2015). The greek music dataset. In *Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS)*, pages 1–7.
- Manning, C. D. (2009). *An introduction to information retrieval*.
- McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

- Medhat, F., Chesmore, D., and Robinson, J. (2020). Masked conditional neural networks for sound classification. *Applied Soft Computing*, 90:106073.
- Meng, Y. (2024). Music genre classification: A comparative analysis of cnn and xgboost approaches with mel-frequency cepstral coefficients and mel spectrograms. *arXiv preprint arXiv:2401.04737*.
- Müller, M. (2015). *Fundamentals of music processing: Audio, analysis, algorithms, applications*, volume 5. Springer.
- Muszynski, S. and Tarapata, Z. (2023). Methods of automated music comparison based on multi-objective metrics of network similarity. *Applied Sciences*, 13(6):3567.
- Pachet, F., Cazaly, D., et al. (2000). A taxonomy of musical genres. In *RIAO*, volume 2, pages 1238–1245. Citeseer.
- Percino, G., Klimek, P., and Thurner, S. (2014). Instrumentational complexity of music genres and why simplicity sells. *PloS one*, 9(12):e115255.
- Powers, D. M. (2011). Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63.
- Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.-Y., and Sainath, T. (2019). Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13(2):206–219.
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.
- Russell, S. and Norvig, P. (2020). *Artificial Intelligence: A Modern Approach*. 4th edition.
- Salamon, J., Rocha, B., and Gómez, E. (2012). Musical genre classification using melody features extracted from polyphonic music signals. In *2012 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 81–84. IEEE.
- Sammut, C. and Webb, G. I. (2011). *Encyclopedia of machine learning*. Springer Science & Business Media.
- Schlüter, J. and Grill, T. (2015). Exploring data augmentation for improved singing voice detection with neural networks. In *ISMIR*, pages 121–126.

- Sheikh Fathollahi, M. and Razzazi, F. (2021). Music similarity measurement and recommendation system using convolutional neural networks. *International Journal of Multimedia Information Retrieval*, 10:43–53.
- Shu, H., Li, X., Jiang, H., Fu, M., and Li, X. (2024). Benchmarking sub-genre classification for mainstage dance music. *arXiv preprint arXiv:2409.06690*.
- Silla Jr, C. N., Kaestner, C. A., and Koerich, A. L. (2007). Automatic music genre classification using ensemble of classifiers. In *2007 IEEE International Conference on Systems, Man and Cybernetics*, pages 1687–1692. IEEE.
- Silla Jr, C. N., Koerich, A. L., and Kaestner, C. A. (2008). The latin music database. In *ISMIR*, pages 451–456.
- Silva, D., Zampar, L., Rodrigues, F., and Gomes, C. (2021a). Modelo automático de classificação de gêneros musicais amazônicos. In *Anais do XVIII Simpósio Brasileiro de Computação Musical*, pages 225–228. SBC.
- Silva, D. F., da Silva, A. C. M., Ortolan, L. F., and Marcacini, R. M. (2021b). 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.
- Silva, D. F., da Silva, M. V., Szram Filho, R., and da Silva, A. C. M. (2021c). On the fusion of multiple audio representations for music genre classification. In *Simpósio Brasileiro de Computação Musical (SBCM)*, pages 37–44. SBC.
- Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.
- Spijkervet, J. and Burgoyne, J. A. (2021). Contrastive learning of musical representations. *arXiv preprint arXiv:2103.09410*.
- Sturm, B. L. (2013). The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use. *arXiv preprint arXiv:1306.1461*.
- Sturm, B. L. (2014). 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.
- Tzanetakis, G. and Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302.

- Van den Oord, A., Dieleman, S., and Schrauwen, B. (2013). Deep content-based music recommendation. *Advances in neural information processing systems*, 26.
- Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Wijaya, N. N., Muslikh, A. R., et al. (2024). Music-genre classification using bidirectional long short-term memory and mel-frequency cepstral coefficients. *Journal of Computing Theories and Applications*, 1(3):243–256.
- Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1, pages 29–39. Manchester.
- Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.
- Xie, C., Song, H., Zhu, H., Mi, K., Li, Z., Zhang, Y., Cheng, J., Zhou, H., Li, R., and Cai, H. (2024). Music genre classification based on res-gated cnn and attention mechanism. *Multimedia Tools and Applications*, 83(5):13527–13542.
- Yonay, O., Hammond, T., and Yang, T. (2025). Myna: Masking-based contrastive learning of musical representations. *arXiv preprint arXiv:2502.12511*.
- Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27.
- Zhao, H., Zhang, C., Zhu, B., Ma, Z., and Zhang, K. (2022). 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)*, pages 606–610. IEEE.
- Zhuang, Y., Chen, Y., and Zheng, J. (2020). Music genre classification with transformer classifier. In *Proceedings of the 2020 4th international conference on digital signal processing*, pages 155–159.