A Multimodal Approach for Music Genre Classification Using Audio and Lyrics Embeddings

J. M. L. Silva[1] **C. M. S. Figueiredo**[2] **E. B. Guedes**[2] **T. De Melo**[2]

[1]Instituto de Pesquisas Eldorado – Manaus
Av. Mário Ypiranga, 315 - Adrianópolis, Manaus - AM, 69057-070

[2]Laboratório de Sistemas Inteligentes – LSI
Escola Superior de Tecnologia (EST) – Universidade do Estado do Amazonas (UEA),
Av. Darcy Vargas, 1.200 - Parque Dez de Novembro, 69050-020 – Manaus – AM – Brazil

{jmlds.snf21, cfigueiredo, ebgcosta, tmelo}@uea.edu.br

Abstract. *Music genre classification is an important task for music recommendation and organization. This work explores deep learning methods for automatic classification, combining convolutional neural networks (CNNs) for audio spectral analysis with large language model (LLM) embeddings from lyrics. Experiments on the GTZAN dataset show that audio-only methods achieved higher overall accuracy, while the multimodal approach redistributed performance across classes, improving some genres (e.g., Country, Pop).*

Resumo. *A classificação de gêneros musicais é uma tarefa importante para sistemas de recomendação e organização. Este trabalho investiga métodos de aprendizado profundo para classificação automática, combinando redes neurais convolucionais (CNNs) para análise espectral de áudio e embeddings de grandes modelos de linguagem (LLMs) extraídos das letras. Os experimentos no conjunto GTZAN mostram que os métodos apenas de áudio alcançaram maior acurácia geral, enquanto a abordagem multimodal redistribuiu o desempenho entre as classes, melhorando alguns gêneros (ex.: Country, Pop).*

1. Introdução

Gêneros musicais são rótulos categóricos para agrupar músicas com características comuns, tais como estrutura rítmica, conteúdo harmônico, emocional e instrumentação [Tzanetakis and Cook 2002]. A classificação automática de gênero musical é crucial para organizar coleções e aprimorar sistemas de recomendação e busca. Essa tarefa, embora complexa, tem gerado pesquisas que vão além dos métodos tradicionais, incluindo extração e análise de características acústicas e uso de modelos de Inteligência Artificial, para enfrentar os desafios da subjetividade e da natureza multifacetada dos gêneros musicais [Sturm 2014].

A Recuperação de Informação Musical (MIR, do inglês *Music Information Retrieval*) consiste em uma área de pesquisa multidisciplinar que utiliza técnicas de *Machine Learning* para abordar a tarefa mencionada [Bahuleyan 2018], incluindo métodos avançados de extração de características [Seo et al. 2024]. Os avanços no uso de *Deep Learning*, conforme reportado na literatura, têm demonstrado progressos em tarefas de

classificação de áudio [Muhammad Turab 2022, Hershey et al. 2017] e na extração de características musicais [Koh and Dubnov 2021].

Conforme o estado da arte, uma abordagem possível para a classificar gêneros musicais envolve a análise de sinais de áudio [Zhang and Zhang 2021, Lau and Ajoodha 2022]. Neste contexto, a proposição de *datasets*, como o GTZAN [Sharma et al. 2020], mostrou-se oportuna como *benchmark* para a avaliação comparativa de modelos. Outra abordagem possível para a classificação de gêneros musicais consiste na análise de letras [Silverman 2009]. Neste caso, se faz necessário o uso de métodos e técnicas de Processamento de Linguagem Natural (PLN), como o uso de *word embeddings* [Wang et al. 2021], para classificação textual. Nesta perspectiva, contribuições para a classificação de gêneros musicais com base em características textuais também foram propostas na literatura [de Araújo Lima et al. 2020, Marijić and Bagić Babac 2025].

Considerando duas vertentes distintas para a classificação de gênero musical, uma baseada em sinais de áudio e a outra nas características das letras, este trabalho propõe uma abordagem multimodal, combinando áudio e letras para a tarefa proposta. Particularmente, são explorados *embeddings* típicos de áudio, como o uso do VGGish [Koh and Dubnov 2021, Campana et al. 2023], em conjunto com *embeddings* de modelos generativos modernos, no caso o LLama 3.2 [Jing et al. 2024]. Os resultados da abordagem multimodal proposta demonstram métricas de acurácia competitivas com o estado da arte (superiores a 90%), apresentando, adicionalmente, maior consistência no desempenho entre as classes avaliadas no *dataset* utilizado.

Para apresentar a abordagem multimodal proposta, o presente trabalho encontra-se estruturado da seguinte forma: a Seção 2 discorre sobre os fundamentos teóricos e os trabalhos correlatos, a Seção 3 detalha a metodologia empregada, a Seção 4 analisa e discute os resultados obtidos e, por fim, a Seção 5 sumariza as considerações finais e perspectivas futuras.

2. Trabalhos Relacionados

Métodos de *Deep Learning* [Goodfellow et al. 2016], especialmente Redes Neurais Convolucionais (CNNs), avançaram significativamente na classificação de sinais de áudio devido à sua capacidade de reconhecer padrões em dados espectrais. Arquiteturas clássicas como LeNet-5, AlexNet e VGG oferecem camadas adaptáveis para diversas representações de áudio, extraindo características em múltiplos níveis [Dogo et al. 2022]. Aplicações incluem classificação de sons respiratórios para diagnóstico [Cinyol et al. 2023] e monitoramento microssísmico [Yin et al. 2021].

Para classificação de áudio, uma etapa comum é a extração de características acústicas, como os Coeficientes Cepstrais na Escala de Mel (MFCC), que representam faixas de frequência dos sinais. No entanto, em *Deep Learning*, espectrogramas, especialmente log-mel espectrogramas [Liu et al. 2024], são mais usados, pois geram dados bidimensionais (frequência e tempo) adequados para CNNs originalmente projetadas para imagens.

No tocante à classificação de gêneros musicais a partir dos sinais de áudio, contribuições recentes elencam soluções de *Deep Learning* para esta tarefa [Zhang and Zhang 2021, Lau and Ajoodha 2022]. Neste contexto, destacam-se dois eixos

principais de investigação: (1) a análise e seleção de características acústicas para identificar descritores mais eficientes na distinção entre gêneros musicais [Sharma et al. 2020]; e (2) estratégias de fusão de características (*feature fusion*), como a fusão tardia (*late fusion*), que processa separadamente STFT, MFCC e espectrogramas Mel antes da classificação. Avanços recentes que combinam múltiplos modelos especializados em diferentes características musicais mostraram desempenho superior a métodos baseados em um único modelo [Muhammad Turab 2022].

Com o avanço das técnicas de *Deep Learning*, a obtenção de *embeddings* a partir de modelos pré-treinados em grandes volumes de dados de áudio tornou-se uma prática comum. O modelo VGGish [Hershey et al. 2017], em particular, ganhou popularidade na classificação de sons em diversas aplicações, como para a classificação de áudios capturados por *smartphones* [Campana et al. 2023]. O presente trabalho emprega o modelo VGGish como arquitetura de referência (*baseline*), adaptando-o à tarefa específica de classificação de gênero musical.

A análise de letras com técnicas de Processamento de Linguagem Natural constitui a segunda abordagem para classificação de gêneros musicais, na qual métodos de *Machine Learning* também são amplamente utilizados. Zhang *et al.* [2022] , por exemplo, classificaram emoções musicais com um modelo híbrido CNN-LSTM com mecanismo de atenção, utilizando TF-IDF e Word2Vec, alcançando 84.8% de acurácia.

O advento de modelos profundos generativos, notadamente os LLMs (do inglês, *Large Language Models*), viabilizou a utilização de *embeddings* textuais para a classificação genérica de textos [Wang et al. 2021]. A título de ilustração, Oliveira *et al.* [2020] aplicaram essa abordagem à classificação de gêneros em letras de músicas brasileiras, enquanto Marijic *et al.* [2025] investigaram a classificação de gêneros musicais em um escopo mais amplo.

Na tarefa de classificação de gêneros musicais, este trabalho propõe uma abordagem multimodal baseada na fusão de representações vetoriais (*embeddings*) de diferentes modalidades. A metodologia integra características acústicas extraídas pela arquitetura VGGish com *embeddings* textuais derivados do modelo de linguagem LLaMA [Jing et al. 2024], explorando complementaridades informativas entre áudio e texto para aumentar a robustez e a acurácia da classificação.

3. Metodologia

O estudo implementou dois pipelines para classificação de gêneros musicais, abrangendo desde o pré-processamento de dados e modelo de classificação. O primeiro baseia-se na análise de áudio com extração de características com *embeddings* VGGish. Já o segundo, combina cacterísticas de áudio e letras, acrescentando um *embedding* textual ao processo.

3.1. Conjunto de Dados

O conjunto de dados utilizado no processo de classificação é o *GTZAN*, reconhecido como um *benchmark* na área de pesquisa em classificação de gêneros musicais. Desenvolvido por George Tzanetakis e Perry Cook em 2002, esse *dataset* tornou-se um padrão de avaliação para algoritmos de reconhecimento e categorização musical [Seo et al. 2024].

O *GTZAN* é composto por 1.000 arquivos de áudio no formato `WAV`, com 30 segundos de duração e taxa de amostragem de 22.050 Hz. As faixas estão organizadas

de forma balanceada entre 10 gêneros musicais: *blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae* e *rock*, sendo 100 faixas para cada categoria [Sharma et al. 2020]. A Tabela 1 apresenta um resumo das principais características do conjunto de dados.

Tabela 1. Atributos e Descrições do *dataset* GTZAN

Atributo	Descrição
Total de faixas	1.000 arquivos de áudio
Gêneros musicais	10 (100 faixas por gênero)
Duração por faixa	30 segundos
Formato dos arquivos	WAV
Taxa de amostragem	22.050 Hz
Canais	Mono
Tamanho total aproximado	Cerca de 1.2 GB
Atributos disponíveis	Nome do arquivo, gênero, caminho relativo
Metadados adicionais	Não disponíveis originalmente

Embora o GTZAN seja amplamente utilizado, ele apresenta limitações reconhecidas na literatura, como faixas duplicadas, baixa qualidade de gravação e inconsistências nos rótulos de gênero, o que pode comprometer a validade da avaliação dos modelos [Sturm 2014].

Além disso, a ausência de metadados estruturados dificulta análises mais contextuais. Para contornar essa limitação, podem ser integradas fontes externas, como a API Genius, para enriquecer o conjunto com letras e informações relevantes.

Apesar desses problemas, o GTZAN permanece popular devido à sua acessibilidade, equilíbrio entre classes e aceitação pela comunidade científica, servindo como base sólida para testes iniciais em tarefas de classificação de gêneros musicais.

3.2. Modelo com VGGish como Extrator de *Features* de Áudio

Para a classificação de gêneros musicais por meio do áudio da música, utilizou-se o modelo VGGish pré-treinado como extrator de características, conforme ilustra a Figura 1.

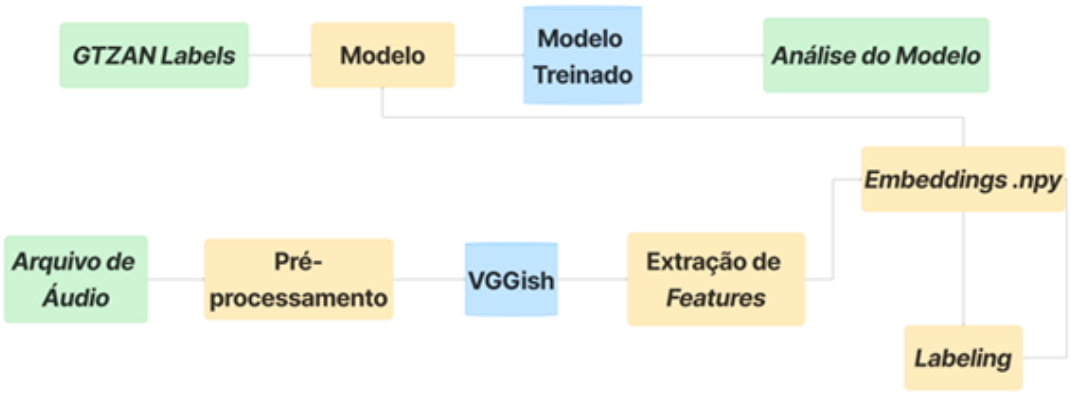

Figura 1. Fluxo do Modelo com Embeddings de Áudio - VGGish

Para isso, uma etapa de pré-processamento foi realizada com o intuito de obter as representações dos áudios em formato visual por meio de seus Mel-Espectrogramas. Essa etapa fez uso da da biblioteca *python* Librosa e um exemplo pode ser visto na Figura 2.

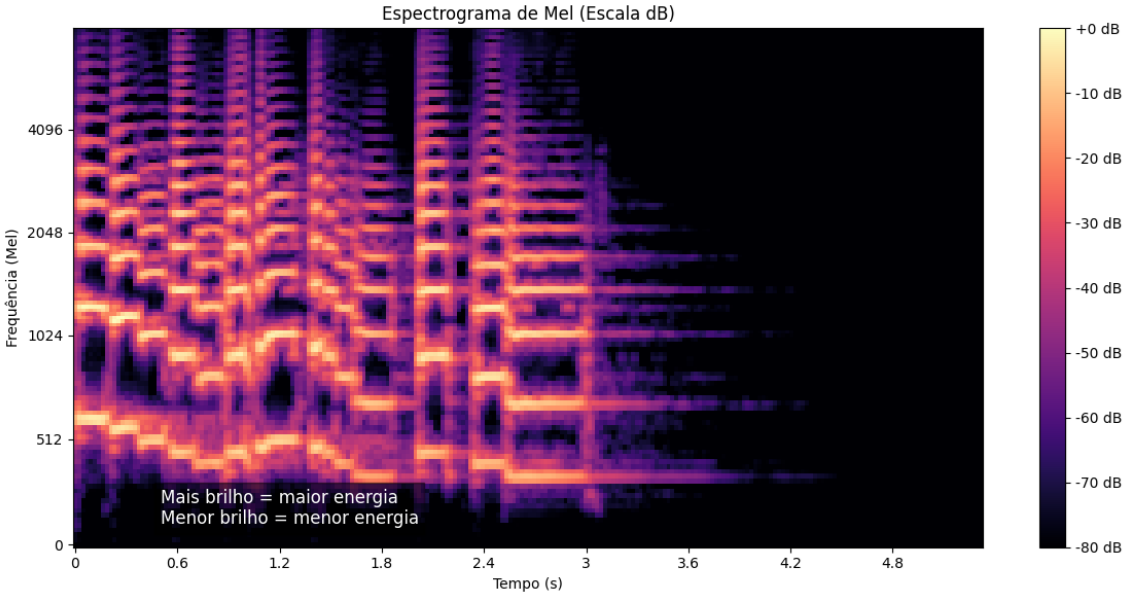

Figura 2. Representação de um *Mel-Spectrogram* de um arquivo de áudio.

A arquitetura VGGish adapta a rede VGG para áudio, operando sobre espectrogramas por meio de quatro blocos convolucionais com filtros 3×3, ativações ReLU e *max pooling*, seguidos por uma camada densa de 128 unidades que gera os *embeddings* acústicos. Esses vetores de dimensão 128 são extraídos a partir de janelas de $0{,}96$ segundos, produzindo matrizes $(N, 128)$, padronizadas para (156×128) via corte ou preenchimento. A extração utilizou o modelo pré-treinado no YouTube-8M com a biblioteca `vggish` do TensorFlow.

Após a extração das *features* pelo modelo VGGish, foi empregada uma rede neural convolucional (CNN) com três camadas `Conv1D` (64, 128 e 256 filtros), seguidas por `BatchNormalization`, `MaxPooling1D` e `Dropout` (30%). A fase densa inclui duas camadas `Dense` (512 e 256 unidades), com regularização L2 e `Dropout` (40%). A saída é gerada por uma camada `Dense` com ativação `softmax`.

O modelo foi treinado com `Adam` (taxa de aprendizado de 0,0001), utilizando `categorical_crossentropy` e métricas de acurácia, precisão e revocação. Adotou-se `EarlyStopping` (paciência de 5 épocas) e pesos de classe balanceados. O treinamento foi realizado em até 100 épocas, com *batch size* de 16 e 20% dos dados reservados para validação.

3.3. Modelo Multimodal com Embeddings de Letras

A arquitetura multimodal, representada na Figura 3, integrou informações acústicas e textuais através de dois ramos paralelos de processamento. O ramo de áudio manteve a estrutura do modelo VGGish descrito anteriormente, enquanto o ramo textual processou *embeddings* gerados pelo LLaMA 3.2 a partir de letras obtidas via API Genius. Para

isso, foi necessário construir uma tabela de mapeamento entre os arquivos do GTZAN e os nomes das músicas e artistas, com base no estudo de Bob L. Sturm - "*The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use*". Em seguida, um processo de *scraping* via API Genius permitiu recuperar as letras correspondentes, que foram armazenadas em um arquivo `.csv` para futura vetorização.

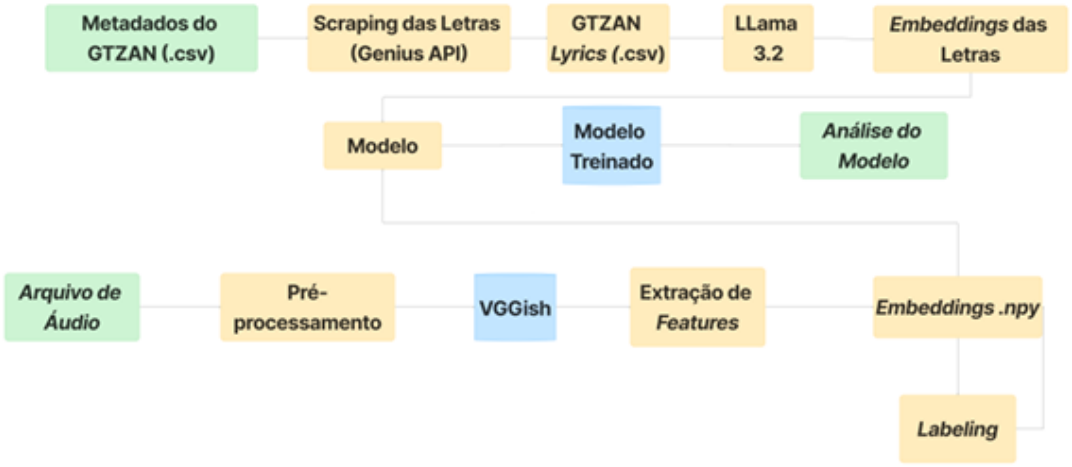

Figura 3. Fluxo do Modelo Multimodal com Embeddings de Áudio e Letras

Como nem todas as faixas do GTZAN têm letras disponíveis, especialmente gêneros instrumentais, foi feito um downsampling estratificado para equilibrar 1.000 amostras por modalidade. Para faixas sem letras, usou-se vetor textual nulo para evitar viés, e aplicaram-se pesos para compensar desequilíbrios. Os embeddings textuais foram gerados com LLaMA 3.2 a partir de letras obtidas por scraping. Os dados foram divididos em 80% treino, 10% validação e 10% teste. A fusão multimodal usou forte regularização e ajuste de pesos entre os ramos textual e acústico.

4. Resultados e Discussões

O estudo avaliou duas abordagens para classificação de gêneros musicais, obtendo os seguintes resultados globais mostrados na Tabela 2, mostrando um grande desempenho por parte do modelo utilizando áudio puro com extração de *features* feita pelo modelo VGGish. O modelo multimodal se mostrou melhor em diferenciar gêneros específicos que possuem características acústicas similares. Detalha-se nas subseções a seguir os resultados de cada modelo.

Tabela 2. Desempenho comparativo das abordagens testadas

Abordagem	Acurácia Total	Gênero com Melhor Desempenho
VGGish (puro)	96.6%	Classical, Hip-hop, Metal (100%)
Multimodal	93.5%	Country, Hip-hop, Jazz, Metal, Pop (100%)

4.1. Modelo de Áudio com VGGish

Como pode ser observado na Tabela 3, a abordagem com VGGish demonstrou superioridade, alcançando 96.6% de acurácia, onde todos os gêneros tiveram desempenho acima

de 90%.

Tabela 3. Métricas de desempenho por gênero com base na matriz de confusão

Gênero	Acurácia	Precisão	Revocação	F1-Score
Blues	95.0%	95%	95%	95%
Classical	100.0%	100%	100%	100%
Country	92.0%	92%	92%	92%
Disco	96.0%	96%	96%	96%
Hiphop	100.0%	100%	100%	100%
Jazz	95.0%	95%	95%	95%
Metal	100.0%	100%	100%	100%
Pop	93.0%	93%	93%	93%
Reggae	96.0%	96%	96%	96%
Rock	99.0%	99%	99%	99%
Acurácia Média (Total)	**96.6%**	-	-	-

Como pode-se observar com auxílio na matriz de confusão da Figura 4, percebe-se que as categorias de pior desempenho foram os gêneros musicais *country*, com confusões distribuídas entras as classes blues, *classical* e *jazz*, e *pop*, com confusão concentrada com *rock*. No primeiro caso, sugere-se que o gênero *country* realmente tenha características musicais diversas, facilitando confusões mais distribuídas. Já no segundo caso, é natural a influência em alguns sub-gêneros de *rock* com *pop*. Essa mesma razão ajuda a justificar as confusões entre *blues* e *rock*.

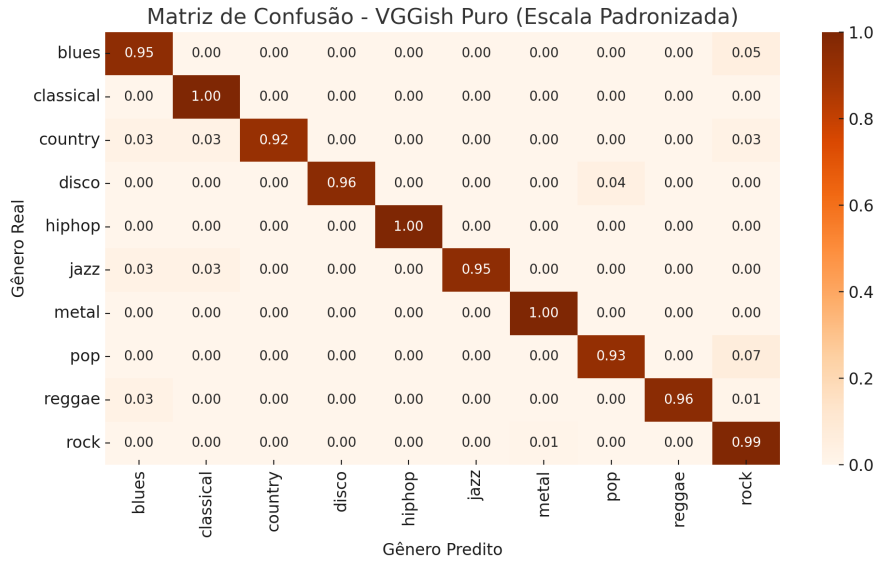

Figura 4. Matriz de confusão - VGGish puro (96% acurácia)

A Figura 5 apresenta a acurácia por gênero obtida na abordagem baseada exclusivamente em áudio, com destaque para a linha de referência que indica a acurácia média geral do modelo (96,6%). Nota-se que alguns dos gêneros musicais superaram

essa média, sendo eles: *Classical*, *Hiphop*, *Metal*, *Rock* e *Reggae*, demonstrando que o modelo foi especialmente eficaz na identificação desses estilos.

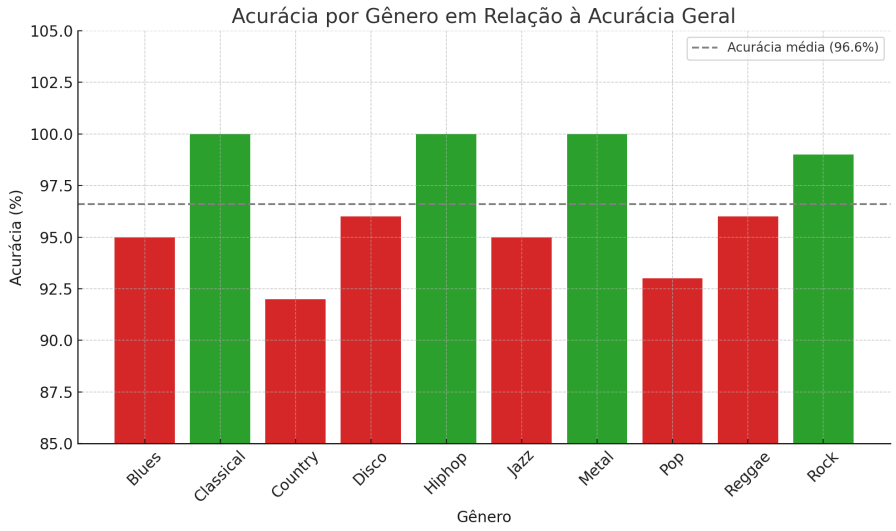

Figura 5. Acurácia por gênero na abordagem utilizando apenas áudio, com linha indicando a média geral (96,6%).

4.2. Modelo Multimodal

A integração de *embeddings* de áudio e letras alcançou 93.5% de acurácia, com cinco gêneros (*Country, Hiphop, Jazz, Metal, Pop*) atingindo 100% de precisão. Como pode-se observar na Tabela 4, as principais confusões ocorreram entre *disco* e *country* (15%) e *reggae* e *rock* (10%).

Tabela 4. Métricas de desempenho por gênero com base na matriz de confusão (Multimodal)

Gênero	Acurácia	Precisão	Revocação	F1-Score
Blues	90.0%	90%	90%	90%
Classical	85.0%	85%	85%	85%
Country	100.0%	100%	100%	100%
Disco	80.0%	80%	80%	80%
Hiphop	100.0%	100%	100%	100%
Jazz	100.0%	100%	100%	100%
Metal	100.0%	100%	100%	100%
Pop	100.0%	100%	100%	100%
Reggae	85.0%	85%	85%	85%
Rock	90.0%	90%	90%	90%
Acurácia Total	**93.5%**	-	-	-

Os resultados sugerem que enquanto características das letras possam ajudar a resolver ambiguidades entre alguns gêneros, como por exemplo, *country* e *pop*, ela findou

por aumentar a confusão de gêneros como *classical*, *disco* e *reggae* como mostra a Figura 6.

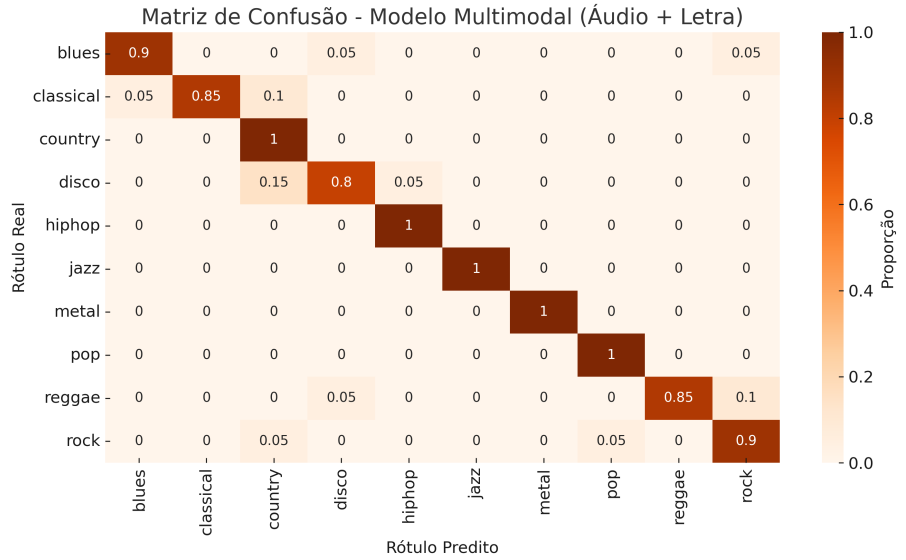

Figura 6. Matriz de confusão - Abordagem multimodal (93.5% acurácia)

Particularmente, na primeira comparação, espera-se que as temáticas das letras realmente tenham diferenças significativas, enquanto que na segunda comparação, temos gêneros com poucas ou nenhuma letra. A Figura 7 mostra detalhadamente estes resultados:

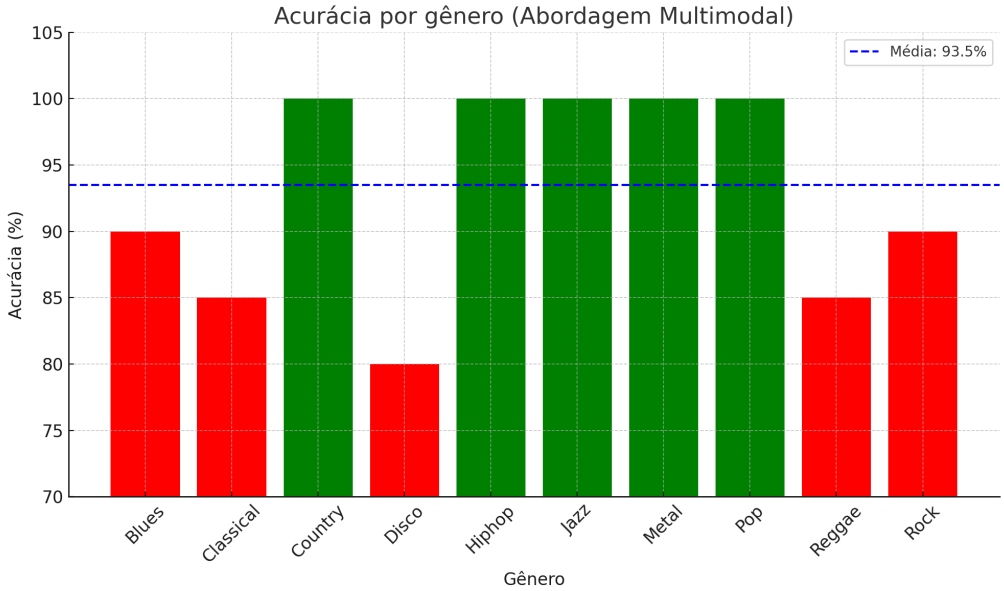

Figura 7. Acurácia por gênero na abordagem multimodal, com linha indicando a média geral (93,5%). Gêneros com desempenho acima da média estão destacados em verde; abaixo da média, em vermelho.

4.3. Principais Achados

Os resultados revelaram padrões distintos de desempenho entre os gêneros musicais analisados. De forma geral, características de áudio das músicas são suficientes para separar bem a grande maioria dos gêneros, causando maiores confusões onde tipicamente há diversidade musical, como *country* e *jazz*, ou influência entre gêneros, como *blues*, *pop* e *rock*.

Quanto à abordagem multimodal, com *embeddings* textuais das letras, resolve-se muitas ambiguidades de gêneros próximos, como pode-se observar com *pop* e *country*, no entanto, acaba-se por interferir no conjunto de características extraídas acrescentando ruído para a classificação de gêneros onde predomina a parte musical, e não a letra, como nos gêneros *classical* e *disco*. Essas observações sugerem que modelos multimodais podem ter melhores resultados ao estruturarmos *ensembles* de modelos.

Em ambos os modelos, espera-se ainda que alguns gêneros possuam maior sobreposição musical, sendo causa ou alvo de confusões recorrentes. Exemplificamos o caso de *rock*, onde tem-se similaridade natural de algumas músicas do gênero com *reggae* e *country*, por exemplo.

Em relação a trabalhos da literatura, podemos observar na tabela 5 que o desempenho do modelo com VGGish em áudio obteve a melhor acurácia no dataset GTZAN, enquanto que o modelo multimodal ficou na mesma faixa de desempenho de outros trabalhos, mostrando-se competitivo em termos de acurácia.

Tabela 5. Desempenho comparativo entre as abordagens presentes nesse estudo e outras abordagens presentes na literatura que também utilizam o *dataset* GTZAN

Método / Estudo	Acurácia Total
VGGish/Multimodal	96,6% / 93,5%
[Bhalke et al. 2017]	96,05%
[Gan 2021]	93,1%
[Liu et al. 2021]	93,65%

5. Considerações Finais

Este trabalho avaliou diferentes abordagens para classificação automática de gêneros musicais, demonstrando que *embeddings* do *VGGish* alcançaram a melhor performance (96% de acurácia), seguido pela abordagem multimodal áudio-letras (93,5%). No entanto, a inclusão das características de letras ajudou a resolver melhor a identificação de gêneros com muita sobreposição musical, como *pop* e *country*. Estes resultados comprovam o potencial das abordagens baseadas em *deep learning* para aplicações em organização musical digital e sistemas de recomendação.

O estudo revelou que gêneros muito distintos, como *metal*, *hip-hop* e *jazz* apresentam menos desafios, seja pelas características acústicas ou textuais da música. Que alguns gêneros têm, nas letras, a introdução de mais confusão, como *disco* e *reggae*. E

que alguns gêneros têm proximidade natural, podendo acarretar em erros recorrentes de classificação, como *rock*.

Os resultados obtidos abrem possibilidades de novos modelos, *ensambles* ou híbridos que permitam a evolução das métricas de classificação, ficando como principal trabalho futuro. Como outras direções futuras, destacam-se: (1) emprego de arquiteturas com mecanismos de atenção, (2) técnicas avançadas de fusão multimodal, e (3) expansão do *dataset* com maior diversidade de gêneros.

Referências

[Bahuleyan 2018] Bahuleyan, H. (2018). Music genre classification using machine learning techniques. *CoRR*, abs/1804.01149.

[Bhalke et al. 2017] Bhalke, D. G., Rajesh, B., and Bormane, D. S. (2017). Automatic genre classification using fractional fourier transform based mel frequency cepstral coefficient and timbral features. *Archives of Acoustics*, vol. 42(No 2).

[Campana et al. 2023] Campana, M. G., Delmastro, F., and Pagani, E. (2023). Transfer learning for the efficient detection of covid-19 from smartphone audio data. *Pervasive and Mobile Computing*, 89:101754.

[Cinyol et al. 2023] Cinyol, F., Baysal, U., Köksal, D., Babaoğlu, E., and Ulaşlı, S. S. (2023). Incorporating support vector machine to the classification of respiratory sounds by convolutional neural network. *Biomedical Signal Processing and Control*, 79:104093.

[de Araújo Lima et al. 2020] de Araújo Lima, R., de Sousa, R. C. C., Barbosa, S. D. J., and Lopes, H. C. V. (2020). Brazilian lyrics-based music genre classification using a BLSTM network. *CoRR*, abs/2003.05377.

[Dogo et al. 2022] Dogo, E. M., Afolabi, O. J., and Twala, B. (2022). On the relative impact of optimizers on convolutional neural networks with varying depth and width for image classification. *Applied Sciences*, 12(23).

[Gan 2021] Gan, J. (2021). Music feature classification based on recurrent neural networks with channel attention mechanism. *Mobile Information Systems*, 2021(1):7629994.

[Goodfellow et al. 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

[Hershey et al. 2017] Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., and Wilson, K. (2017). Cnn architectures for large-scale audio classification.

[Jia 2022] Jia, X. (2022). Music emotion classification method based on deep learning and improved attention mechanism. *Computational Intelligence and Neuroscience*, 2022(1):5181899.

[Jing et al. 2024] Jing, H., Liu, Y., Ma, Y., and Zheng, N. (2024). Hidden states in llms improve eeg representation learning and visual decoding. In *ECAI*, pages 2130–2137.

[Koh and Dubnov 2021] Koh, E. and Dubnov, S. (2021). Comparison and analysis of deep audio embeddings for music emotion recognition.

[Lau and Ajoodha 2022] Lau, D. and Ajoodha, R. (2022). *Music Genre Classification: A Comparative Study Between Deep Learning and Traditional Machine Learning Approaches*, pages 239–247.

[Liu et al. 2024] Liu, H., Liu, X., Kong, Q., Wang, W., and Plumbley, M. D. (2024). Learning temporal resolution in spectrogram for audio classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(12):13873–13881.

[Liu et al. 2021] Liu, J., Wang, C., and Zha, L. (2021). A middle-level learning feature interaction method with deep learning for multi-feature music genre classification. *Electronics*, 10(18).

[Marijić and Bagić Babac 2025] Marijić, A. and Bagić Babac, M. (2025). Predicting song genre with deep learning. *Global Knowledge, Memory and Communication*, 74(1/2):93–110.

[Muhammad Turab 2022] Muhammad Turab, Tipparti Anil Kumar, M. B. T. S. (2022). Investigating multi-feature selection and ensembling for audio classification.

[Seo et al. 2024] Seo, W., Cho, S.-H., Teisseyre, P., and Lee, J. (2024). A short survey and comparison of cnn-based music genre classification using multiple spectral features. *IEEE Access*, 12:245–257.

[Sharma et al. 2020] Sharma, G., Umapathy, K., and Krishnan, S. (2020). Trends in audio signal feature extraction methods. *Applied Acoustics*, 158:107020.

[Silverman 2009] Silverman, M. J. (2009). The use of lyric analysis interventions in contemporary psychiatric music therapy: Descriptive results of songs and objectives for clinical practice. *Music Therapy Perspectives*, 27(1):55–61.

[Sturm 2014] 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 and Cook 2002] Tzanetakis, G. and Cook, P. (2002). Musical genre classification of audio signals. *IEEE Trans. Speech Audio Process.*, 10(5):293–302.

[Wang et al. 2021] Wang, C., Nulty, P., and Lillis, D. (2021). A comparative study on word embeddings in deep learning for text classification. In *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '20, page 37–46, New York, NY, USA. Association for Computing Machinery.

[Yin et al. 2021] Yin, X., Liu, Q., Huang, X., and Pan, Y. (2021). Real-time prediction of rockburst intensity using an integrated cnn-adam-bo algorithm based on microseismic data and its engineering application. *Tunnelling and Underground Space Technology*, 117:104133.

[Zhang and Zhang 2021] Zhang, Y. and Zhang, K. (2021). Music style classification algorithm based on music feature extraction and deep neural network. *Wireless Communications and Mobile Computing*, 2021:9298654.