Classificação de Gêneros Musicais Usando Redes Convolucionais

Matheus A. de Oliveira, Walter A. Gontijo, João P. Vieira, Pedro P. Santos, Danilo Silva e Eduardo L. O. Batista

***Resumo*— Este artigo explora abordagens baseadas em redes neurais convolucionais (CNNs) para a classificação de gêneros musicais utilizando o dataset GTZAN. São avaliadas três estratégias principais: (1) VGGish, uma arquitetura CNN desenvolvida para tarefas de áudio; (2) um modelo com *transfer learning* utilizando MobileNetV2 pré-treinado no ImageNet; e (3) um esquema de votação majoritária entre segmentos. Os experimentos foram conduzidos tanto na versão filtrada do GTZAN (GTZAN-FF), que corrige falhas estruturais, quanto em uma versão não filtrada e aleatória (GTZAN-Random). Os resultados mostram que a combinação de *transfer learning* com votação majoritária atinge a maior acurácia realista (76% no GTZAN-FF), enquanto a avaliação sobre dados não filtrados gera métricas infladas (88,5%). Tais resultados mostram que utilizar uma versão filtrada do GTZAN permite garantir a validade dos resultados em cenários mais realistas.**

***Palavras-Chave*— Classificação de Gêneros Musicais, CNN, Transfer Learning.**

***Abstract*— This paper explores approaches based on Convolutional Neural Networks (CNNs) for music genre classification using the GTZAN dataset. Three main strategies are evaluated: (1) VGGish, a CNN architecture developed for audio tasks; (2) a transfer learning model using MobileNetV2 pre-trained on ImageNet; and (3) a majority voting scheme among segments. Experiments were conducted on both the filtered version of GTZAN (GTZAN-FF), which corrects structural flaws, and an unfiltered and randomly split version (GTZAN-Random). Experimental results show that combining transfer learning with majority voting achieves the highest realistic accuracy (76% on GTZAN-FF), while evaluation on unfiltered data yields inflated metrics (88.5%). These findings indicate that using a filtered version of GTZAN helps ensure the validity of results in more realistic scenarios.**

***Keywords*— Music Genre Classification, CNN, Transfer Learning.**

I. Introdução

Com o vasto volume de músicas disponíveis na internet, plataformas de streaming como Spotify e iTunes enfrentam o grande desafio de classificar e recomendar faixas que melhor se alinhem às preferências de seus usuários. Dessa forma, o uso de ferramentas capazes de categorizar automaticamente gêneros musicais de forma eficiente tem se tornado essencial.

No campo da recuperação de informação musical (*Music Information Retrieval* — MIR), diversos estudos têm explorado a classificação de gêneros musicais utilizando redes neurais convolucionais (CNNs) [1]–[4]. No entanto, muitas dessas investigações não examinam adequadamente técnicas alternativas, nem abordam de forma satisfatória os desafios associados ao uso do conjunto de dados GTZAN [5].

O presente trabalho tem como objetivo desenvolver um modelo robusto para a classificação de gêneros musicais utilizando redes neurais convolucionais (CNNs), explorando diferentes técnicas e abordagens para maximizar seu desempenho. Além disso, adota-se uma metodologia mais rigorosa no uso do *dataset* GTZAN, aplicando filtros específicos desenvolvidos para mitigar problemas como o vazamento de dados. Para identificar e discutir os cuidados metodológicos inerentes à aplicação de CNNs de segmentos curtos, replicamos o estudo de Choudhury et al. [1] e avaliamos os resultados obtidos em diferentes cenários experimentais.

II. Trabalhos Relacionados

A classificação de gêneros musicais tem sido amplamente explorada por diversos pesquisadores, com o objetivo de desenvolver novas abordagens e técnicas para aprimorar *datasets*, métodos de extração de *features* e modelos de classificação. Nesse contexto, Tzanetakis e Cook introduziram o *dataset* GTZAN, o qual tem sido amplamente utilizado em estudos de classificação de gêneros musicais. Tal *dataset* contém 10 gêneros musicais, cada um com 100 amostras de áudio de 30 segundos, e reúne diversas *features* [5].

As limitações do *dataset* GTZAN são bem conhecidas pela comunidade científica. No estudo de Sturm [6], os principais problemas identificados incluem a super-representação de certos artistas, amostras duplicadas, baixa qualidade de áudio, faixas com rótulos incorretos e a ausência de um filtro por artista. Em resposta a essas questões, diversos pesquisadores propuseram versões personalizadas do GTZAN que procuram mitigar essas deficiências. Por exemplo, os estudos em [7]–[9] utilizaram uma versão filtrada do *dataset* GTZAN, a qual possui partes separadas para o treino, teste e validação. Adicionalmente, o estudo em [10] propôs três versões filtradas do *dataset* GTZAN, cada uma dividida em partes dedicadas para treino e teste.

No campo da recuperação de informação musical (MIR), diversos estudos investigaram o desempenho de diferentes modelos de aprendizado de máquina para a tarefa de classificação de gêneros musicais. Em [11], Srinivas et al. compararam SVM (*Suport Vector Machine*), CNN e redes neurais convolucionais recorrentes (RCNN), relatando que as CNNs obtiveram a maior acurácia. Em [12], Jahnavi et al. avaliaram modelos de aprendizado profundo em comparação com abordagens clássicas, como KNN (*K-Nearest Neighbors*), MLP (*Multi-Layer Perceptron*), SVM e *Naive Bayes*, concluindo que as CNNs superaram consistentemente os modelos tradicionais. Chillara et al. [13] examinaram CNNs, redes neurais convolucionais

Matheus Oliveira, Walter A. Gontijo, Eduardo L. O. Batista, LINSE Laboratório de Circuitos e Processamento de Sinais, Departamento de Engenharia Elétrica e Eletrônica, Universidade Federal de Santa Catarina, UFSC, Florianópolis-SC, Brasil, e-mails: matheusaop09@linse.ufsc.br, walter@linse.ufsc.br e eduardo.batista@ufsc.br. João Paulo Vieira, Pedro Pordeus Santos e Danilo Silva, Departamento de Engenharia Elétrica e Eletrônica, Universidade Federal de Santa Catarina, Florianópolis, SC, e-mails: j.p.vieira@posgrad.ufsc.br, pedro.pordeus@grad.ufsc.br, danilo.silva@ufsc.br.

recorrentes (CRNN) e modelos híbridos CNN-RNN, identificando as CNNs como a arquitetura mais eficaz. Em [14], Ramírez et al. avaliaram os modelos *Naive Bayes*, SVM, redes neurais e redes neurais recursivas utilizando o *dataset* AudioSet, e observaram que as redes neurais apresentaram o melhor desempenho. De forma semelhante, Deng et al. [15] compararam diversos classificadores — incluindo *Naive Bayes*, KNN, Árvores de Decisão, Florestas Aleatórias, SVM, Regressão Logística e Redes Neurais Totalmente Conectadas (FCNN) — e demonstraram que as FCNNs produziram os melhores resultados entre os métodos testados.

Em [16], Won et al. avaliaram sete abordagens de aprendizado profundo amplamente utilizadas na literatura, incluindo Redes Convolucionais Totalmente Conectadas (FCN), Musicnn, CNN em nível de amostra (*Sample-Level CNN*), redes convolucionais recorrentes (CRNN), um modelo baseado em *self-attention*, CNN Harmônica (*Harmonic* CNN) e CNN de segmentos curtos (*Short-Chunk* CNN). A avaliação foi realizada com três conjuntos de dados diferentes: MagnaTagATune (MTAT), *Million Song Dataset* (MSD) e MTG-Jamendo. Os resultados indicaram que o modelo *Short-Chunk* CNN obteve o melhor desempenho geral.

III. CNN de Segmentos Curtos

A CNN de segmentos curtos (*Short-chunk* CNN) é uma rede neural convolucional totalmente conectada, cuja principal diferença está na forma de entrada dos dados. Diferentemente de outras arquiteturas, esse modelo opera sobre *features* extraídas de pequenos segmentos de áudio (tipicamente trechos de 3 segundos), permitindo capturar padrões locais relevantes para a tarefa de classificação musical [16].

Neste trabalho, a arquitetura CNN de segmentos curtos foi implementada e seu desempenho aprimorado por meio de uma estratégia de votação majoritária. Esta seção discute os cuidados metodológicos essenciais ao empregar a arquitetura CNN de segmentos curtos. Para tal, é realizada a replicação do estudo de Choudhury et al. [1], que utiliza essa mesma abordagem. A partir da comparação entre os resultados obtidos na replicação e os apresentados no estudo original, foram identificadas e analisadas possíveis falhas metodológicas. Com base nessa análise, são apresentadas recomendações para a utilização adequada dessa arquitetura em tarefas de classificação de gêneros musicais.

O modelo desenvolvido por Choudhury et al. possui 88.330 paramêtros treináveis, e foi configurado com tamanho de lote (*batch size*) de 32, 100 épocas e taxa de aprendizado de 0,0001, empregando uma camada de dropout com taxa de 0,3. Adicionalmente, Choudhury et al. segmentaram o conjunto de dados GTZAN dividindo cada faixa musical em 10 segmentos, resultando em um total de 10000 amostras de áudio de três segundos cada, das quais 8000 foram usadas para treino, 1000 para teste e 1000 para validação. Foram utilizados coeficientes cepstrais em frequência mel (MFCCs) como *feature* de entrada para a CNN; entretanto, detalhes específicos sobre os parâmetros usados na geração dos MFCCs não foram fornecidos. Na replicação, foram utilizados os parâmetros padrão da biblioteca Librosa [17], com exceção do parâmetro n_mfcc, que foi definido como 128.

As Figuras 1, 2 e 3 apresentam as curvas de treinamento para três cenários experimentais: (1) com vazamento de dados, (2) sem vazamento e (3) sem vazamento e sem embaralhamento. Apenas no Cenário 1 (em que segmentos de uma mesma faixa são aleatoriamente distribuídos entre treino, validação e teste), os resultados coincidem com os reportados por Choudhury et al. [1].

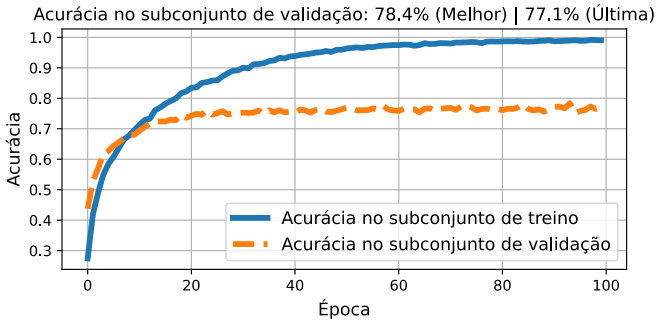

Fig. 1. Curva de Treinamento e Acurácia, Cenário com Vazamento.

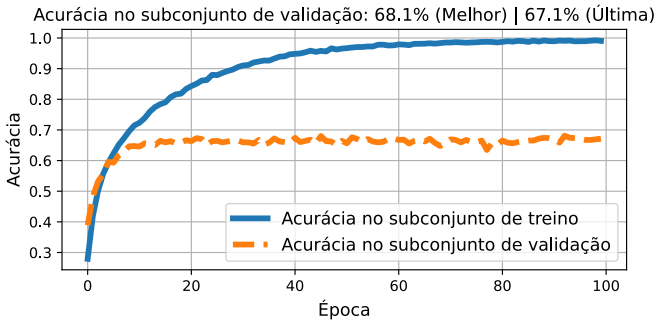

Fig. 2. Curva de Treinamento e Acurácia, Cenário sem Vazamento.

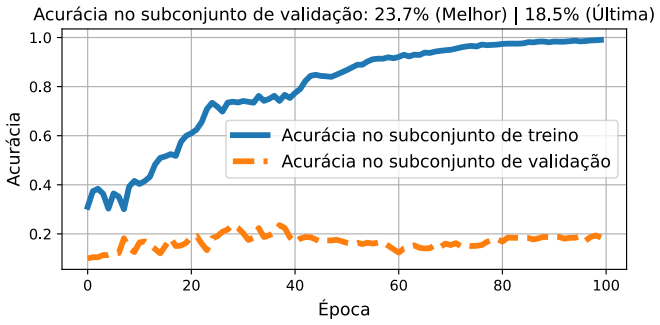

Fig. 3. Curva de Treinamento e Acurácia, Cenário sem Vazamento e sem o Embaralhamento do *Dataset*.

No segundo cenário, em que o vazamento de dados foi eliminado, observa-se uma queda de aproximadamente 10% na acurácia de validação. No terceiro cenário, em que não houve vazamento nem embaralhamento, a acurácia diminuiu ainda mais, caracterizando *overfitting*. O vazamento ocorre porque cada faixa de 30s é dividida em dez segmentos de 3s altamente correlacionados; quando esses segmentos são alocados aleatoriamente em diferentes conjuntos, o modelo passa a avaliar dados muito semelhantes em treino e validação, superestimando seu desempenho em ambiente de produção. Além disso, a ausência da etapa de embaralhamento após a segmentação resulta em lotes sequenciais, reduzindo a diversidade interna de cada *batch* e agravando o *overfitting*.

A partir desta análise, são apresentadas as seguintes recomendações: agrupar todos os segmentos de uma mesma faixa em um único subconjunto (treino, validação ou teste) e de embaralhar as amostras segmentadas antes do treinamento.

IV. Métodos e Abordagens

O presente trabalho propõe a implementação de três abordagens distintas para a tarefa de classificação de gêneros musicais, todas fundamentadas na arquitetura de CNN de segmentos curtos. Nesta seção, cada uma dessas abordagens é apresentada e discutida em detalhe. A primeira consiste em uma arquitetura convolucional projetada especificamente para análise de áudio. A segunda explora o uso de *transfer learning*, aproveitando redes pré-treinadas para aprimorar a extração de *features* relevantes. Por fim, é descrita uma estratégia de votação majoritária entre segmentos, com o objetivo de reforçar o desempenho global do modelo. A seguir, tais abordagens são descritas em mais detalhes.

A. VGGish

VGGish é uma arquitetura de CNN derivada da VGG16 e projetada especificamente para análise de áudio. Ela tem sido amplamente adotada em diversas tarefas de classificação de áudio [18]–[20]. Neste trabalho, o modelo VGGish é reimplementado com base no estudo [18] e avalia-se sua eficácia na tarefa de classificação de gêneros musicais. A Tabela I resume a arquitetura da rede.

TABELA I
ARQUITETURA VGGISH

Camada	Tipo	Neurônios/Filtros	Kernel
1	*Convolution (ReLU)*	64	3x3
2	*Max Pool, strides (2,2)*	-	2x2
3	*Convolution (ReLU)*	128	3x3
4	*Max Pool, strides (2,2)*	-	2x2
5	*Convolution (ReLU)*	256	2x2
6	*Convolution (ReLU)*	256	2x2
7	*Max Pool, strides (2,2)*	-	2x2
8	*Convolution (ReLU)*	512	2x2
9	*Convolution (ReLU)*	512	2x2
10	*Max Pool, strides (2,2)*	-	2x2
11	*Global Average Pooling*	-	-
12	*Flatten*	-	-
13	*Dense (ReLU)*	32	-
14	*Dense (ReLU)*	32	-
15	*Dense (Output) (Softmax)*	10	-
Número Total de Parâmetros Treináveis: 2,059,914			

Para a implementação utilizada neste trabalho, emprega-se a API Keras [21]. Como entrada, são utilizados espectrogramas Mel extraídos com a biblioteca Librosa [17], mantendo-se seus parâmetros padrão, exceto pelo número de bandas Mel (n_mels), que foi aumentado para 512.

B. Transfer Learning

Esta seção aborda a aplicação de *transfer learning* para a tarefa de classificação de gêneros musicais. Essa abordagem consiste em aproveitar uma rede neural pré-treinada no conjunto de dados ImageNet (originalmente desenvolvida para classificação de imagens) e adaptá-la para operar em *features* extraídas de sinais de áudio. Como os pesos pré-treinados provêm de um domínio não relacionado ao conjunto de dados alvo, não há correlação significativa entre os dados de origem e destino, o que ajuda a prevenir *overfitting*. Diversos estudos na literatura [22]–[25] adotaram essa estratégia e relataram resultados promissores.

Neste trabalho, utilizam-se as implementações disponíveis na API Keras [21], que fornece acesso a uma variedade de modelos pré-treinados por meio do módulo Keras *Applications*. Esses modelos esperam entradas na forma de tensores de imagem com três canais (RGB). Entretanto, em tarefas baseadas em áudio, as *features* de entrada (espectrogramas Mel) são tipicamente matrizes de um único canal. Para atender aos requisitos de entrada desses modelos, os espectrogramas foram repetidos nos três canais. Além disso, as matrizes de entrada foram redimensionadas para 224×224, formato de entrada mais comum esperado pelas arquiteturas pré-treinadas.

Entre os diversos modelos disponíveis no módulo Keras *Applications*, optou-se pelo MobileNetV2. Essa escolha baseou-se em seu menor número de parâmetros, o que oferece eficiência computacional, assim como no forte desempenho demonstrado em comparação a outros modelos considerados em nossas avaliações preliminares.

C. Método Baseado em Votação

As técnicas discutidas nas Subseções A e B foram treinadas usando segmentos de 3 segundos extraídos de faixas de áudio de 30 segundos. Entretanto, é possível (e frequentemente desejável) combinar as previsões desses segmentos individuais para inferir o gênero da música completa. Essa estratégia está em sintonia com cenários reais, onde geralmente dispõe-se de trechos mais longos ou da música inteira, e assume-se que todos os segmentos de uma mesma faixa compartilham o mesmo rótulo de gênero.

Essa técnica, conhecida como votação em nível de segmento, foi adotada em estudos anteriores [26], [27], os quais demonstraram que agregar previsões por votação majoritária pode levar a melhor desempenho de classificação. Neste trabalho, aplica-se a mesma estratégia nos modelos VGGish e o modelo MobileNet-V2 (*transfer learning*). Especificamente, durante o treinamento, os modelos foram treinados com segmentos de áudio de 3 segundos. Na fase de teste, cada música de 30 segundos foi dividida em dez segmentos de 3 segundos. O modelo então gerou uma previsão para cada segmento, e a previsão final para a música completa foi obtida aplicando-se votação majoritária entre as dez previsões.

V. Metodologia

Para avaliar de forma rigorosa o desempenho de diferentes abordagens baseadas em CNN para classificação de gêneros musicais, este trabalho adota um *pipeline* experimental em quatro etapas. Primeira, criam-se duas categorias distintas do *dataset* GTZAN para avaliar o impacto de suas falhas conhecidas nos resultados finais. Segunda, implementam-se duas arquiteturas de CNN: uma rede baseada em VGGish projetada especificamente para tarefas de áudio e uma rede baseada em MobileNetV2 aproveitando a abordagem de *transfer*

learning. Terceira, realiza-se otimização de hiperparâmetros para cada modelo usando o método *hold-out* com um conjunto de validação dedicado. Finalmente, aplica-se uma estratégia de votação majoritária para agregar previsões em nível de segmento em um único rótulo de gênero para cada faixa de 30 segundos, simulando condições de implantação em cenários reais.

A. Dataset Utilizado

O *dataset* GTZAN consiste em 10 gêneros musicais, cada um representado por 100 faixas de 30 segundos. Neste trabalho, cada faixa é segmentada em 10 trechos não sobrepostos de 3 segundos, resultando em 1000 segmentos por gênero. Embora amplamente utilizado na literatura, o GTZAN contém várias falhas bem documentadas, conforme apontado por Sturm [6]. Essas falhas incluem réplicas, amostras de áudio corrompidas e uma forte influência de artistas que pode enviesar a avaliação do modelo. Para mitigar esses problemas, adotam-se duas versões filtradas do *dataset*. A primeira versão é utilizada em [7]–[9] e está disponível publicamente em um repositório aberto[1]. A segunda versão é proposta por Foleiss et al. [10] e está disponível em outro repositório[2]. Ambas versões aplicam as seguintes correções: remoção de réplicas e arquivos corrompidos e um filtro de artistas para garantir que não ocorra vazamento de artistas entre os subconjuntos de treino, teste e validação. Neste trabalho, essas versões filtradas do GTZAN são denominadas como GTZAN-FF.

Para fins de comparação, também avaliam-se os modelos no GTZAN original não filtrado usando uma divisão aleatória, referida como GTZAN-Random. Conforme discutido na Seção III, o vazamento de dados é evitado agrupando os dez segmentos de cada faixa em um mesmo subconjunto (treino, teste ou validação) e, em seguida, embaralhando essas amostras segmentadas. Esses cuidados metodológicos foram rigorosamente aplicados em todos os experimentos.

B. Ajuste de Hiperparâmetros

A estratégia de busca de hiperparâmetros segue o esquema *hold-out* que utiliza a versão GTZAN-FF. Os modelos são treinados no subconjunto de treino e os hiperparâmetros são selecionados com base no desempenho no subconjunto de validação. Emprega-se o otimizador Adam e, inicialmente, realiza-se uma busca em grade (*grid search*) sobre taxa de aprendizado (*learning rate*), tamanho de lote (*batch size*) e número de épocas para ambos os modelos VGGish e MobileNetV2 (*transfer learning*). Em seguida, executa-se uma busca aleatória (*random search*) para ajustar finamente os parâmetros β_1 e β_2 do otimizador.

As combinações ótimas de hiperparâmetros para cada arquitetura estão resumidas na Tabela II. Esses ajustes são então fixados para o treinamento final do modelo e posterior avaliação no conjunto de teste.

[1]https://github.com/boblsturm/GTZAN

[2]https://github.com/julianofoleiss/gtzan_sturm_filter_3folds_stratified

TABELA II
HIPERPARÂMETROS UTILIZADOS

Hiperparâmetros	VGGish	Transfer Learning
Número de épocas	50	100
Tamanho do *batch*	64	64
Taxa de aprendizado	0,001	0,001
β_1	0,9	0,9
β_2	0,999	0,999

VI. RESULTADOS

A Tabela III apresenta a acurácia média e o desvio padrão de cada método, nas versões filtradas do GTZAN (GTZAN-FF) e aleatórias não filtradas (GTZAN-Random). Para obter a média e desvio padrão foram executadas 5 divisões do GTZAN-Random com diferentes *seeds* de aleatoriedade. Já para o caso do GTZAN-FF, foi utilizado 4 divisões distintas do *dataset*.

Observa-se que a estratégia combinando *transfer learning* e votação majoritária obteve o melhor desempenho geral, alcançando uma acurácia média de 76,09% no GTZAN-FF. No entanto, o mesmo método obteve 88,5% de acurácia no GTZAN-Random, um valor significativamente mais alto, mas que deve ser interpretado com cautela. A diferença entre os dois cenários evidencia que a versão filtrada do GTZAN oferece uma estimativa mais realista do desempenho em produção, ao passo que a versão aleatória não filtrada superestima os resultados devido a falhas estruturais no conjunto de dados, como duplicações e vazamento de informações entre os conjuntos de treino e teste.

TABELA III
ACURÁCIA MÉDIA E DESVIO PADRÃO

Dataset	Métodos	Acurácia (%)
GTZAN-FF	*Transfer Learning*	71.32 $\pm$ 2.48
GTZAN-FF	*Transfer Learning* + estratégia de votação	76.09 $\pm$ 3.21
GTZAN-FF	VGGish	62.15 $\pm$ 4.61
GTZAN-FF	VGGish + estratégia de votação	71.14 $\pm$ 7.54
GTZAN-Random	*Transfer Learning*	80.58 $\pm$ 1.53
GTZAN-Random	*Transfer Learning* + estratégia de votação	88.50 $\pm$ 1.43
GTZAN-Random	VGGish	73.82 $\pm$ 2.48
GTZAN-Random	VGGish + estratégia de votação	81.67 $\pm$ 3.22

A Fig. 4 apresenta a matriz de confusão da abordagem com *transfer learning* e votação majoritária no conjunto de teste do GTZAN-FF.

Nota-se que os gêneros reggae, rock e country apresentaram os piores desempenhos, com acurácia inferior a 68%, indicando maior confusão entre essas classes.

Esses resultados reforçam a importância do uso de versões filtradas do *dataset* GTZAN nos experimentos, não apenas para garantir a validade estatística das métricas obtidas, mas também para evitar conclusões excessivamente otimistas que não se sustentariam em contextos reais de aplicação. A combinação de *transfer learning* e votação majoritária mostra-se, portanto, uma solução eficaz e realista dentro de um protocolo experimental rigoroso.

VII. CONCLUSÕES

Este artigo avaliou três abordagens baseadas em CNN para classificação de gêneros musicais no *dataset* GTZAN. Primeiro, explorou-se a arquitetura VGGish, projetada especificamente para análise de áudio, seguida de um modelo de *transfer learning* com MobileNetV2 pré-treinado no ImageNet. Por fim, aplicou-se um esquema de votação majoritária entre

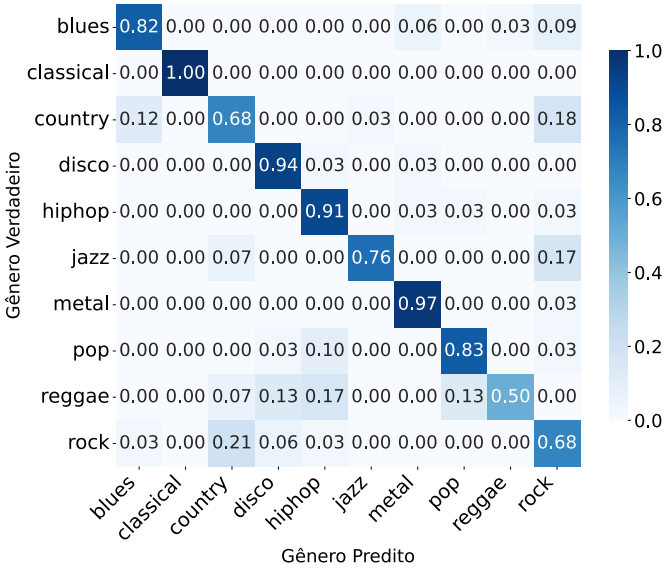

Fig. 4. Matriz de confusão da abordagem que combina *transfer learning* e votação majoritária no GTZAN-FF.

segmentos, permitindo um melhor desempenho de predição do gênero da musica. Os experimentos foram conduzidos tanto em versões filtradas do GTZAN (GTZAN-FF), que corrigem replicas e evitam vazamento de dados, quanto em uma versão aleatória (GTZAN-Random). Tais experimentos mostraram que a combinação de *tranfer learning* e votação majoritária alcançaram os melhores resultados.

AGRADECIMENTOS

Este trabalho foi realizado com apoio de recursos computacionais disponibilizados pela Fundação de Amparo à Pesquisa e Inovação do Estado de Santa Catarina, termo de outorga n. 2024TR000090.

REFERÊNCIAS

[1] N. Choudhury, D. Deka, S. Sarmah, and P. Sarma, "Music genre classification using convolutional neural network," in *Proc. 4th International Conference on Computing and Communication Systems (I3CS)*, Shillong, India, 2023, pp. 1–5.

[2] Y.-H. Cheng, P.-C. Chang, and C.-N. Kuo, "Convolutional neural networks approach for music genre classification," in *Proc. 2020 International Symposium on Computer, Consumer and Control (IS3C)*, Taichung City, Taiwan, 2020, pp. 399–403.

[3] M. Bohra, I. Kumar, and Shivam, "Automated music genre classification using modified mobilenet deep learning model," in *Proc. 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, Coimbatore, India, 2024, pp. 767–772.

[4] M. Shah, N. Pujara, K. Mangaroliya, L. Gohil, T. Vyas, and S. Degadwala, "Music genre classification using deep learning," in *Proc. 6th International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2022, pp. 974–978.

[5] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, Jul. 2002.

[6] B. L. Sturm, "The state of the art ten years after a state of the art: Future research in music information retrieval," *Journal of New Music Research*, vol. 43, no. 2, p. 147–172, Apr. 2014. [Online]. Available: http://dx.doi.org/10.1080/09298215.2014.894533

[7] C. Kereliuk, B. L. Sturm, and J. Larsen, "Deep learning, audio adversaries, and music content analysis," in *Proc. 2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, USA, 2015, pp. 1–5.

[8] ——, "Deep learning and music adversaries," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2059–2071, Nov. 2015.

[9] F. Rodríguez-Algarra, B. L. Sturm, and H. Maruri-Aguilar, "Analysing scattering-based music content analysis systems: Where's the music?." in *Proc. 17th International Society for Music Information Retrieval Conference*, New York City, USA, 2016, pp. 344–350.

[10] J. H. Foleis and T. F. Tavares, "Texture selection for automatic music genre classification," *Applied Soft Computing*, vol. 89, p. 106127, Mar. 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1568494620300673

[11] U. M. Srinivas, S. Rafi, T. V. Manohar, and M. V. Rao, "Classification of music genre using deep learning approaches," in *Proc. 4th International Conference on Artificial Intelligence and Signal Processing (AISP)*, Vijayawada, India, 2024, pp. 1–5.

[12] M. Jahnavi, A. Satapathy, C. Lokesh, and P. B. Likhitha, "A comparative performance evaluation of machine learning approaches for spectrogram-based music genre classification," in *Proc. IEEE 3rd International Conference on Technology, Engineering, Management for Societal impact using Marketing, Entrepreneurship and Talent (TEMSMET)*, Mysuru, India, 2023, pp. 1–7.

[13] S. Chillara, A. Kavitha, S. A. Neginhal, S. Haldia, and K. Vidyullatha, "Music genre classification using machine learning algorithms: a comparison," *Int Res J Eng Technol*, vol. 6, no. 5, pp. 851–858, May. 2019. [Online]. Available: https://www.irjet.net/archives/V6/i5/IRJET-V6I5174.pdf

[14] J. Ramírez and M. J. Flores, "Machine learning for music genre: multifaceted review and experimentation with audioset," *Journal of Intelligent Information Systems*, vol. 55, no. 3, pp. 469–499, Dec. 2020. [Online]. Available: https://doi.org/10.1007/s10844-019-00582-9

[15] D. Deng, Y. Gu, and Y. Zhu, "Comparison of multiple machine learning algorithms for music genre classification," *Applied and Computational Engineering*, vol. 8, pp. 768–774, Aug. 2023. [Online]. Available: https://www.ewadirect.com/proceedings/ace/article/view/2708

[16] M. Won, A. Ferraro, D. Bogdanov, and X. Serra, "Evaluation of cnn-based automatic music tagging models," 2020. [Online]. Available: https://arxiv.org/abs/2006.00751

[17] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proc. 14th Python in Science Conference*, Austin, USA, 2015, pp. 18–25.

[18] M. A. L. Sawadogo, F. Pala, G. Singh, I. Selmi, P. Puteaux, and A. Othmani, "Ptsd in the wild: a video database for studying post-traumatic stress disorder recognition in unconstrained environments," *Multimedia Tools and Applications*, vol. 83, no. 14, pp. 42 861–42 883, Apr. 2024. [Online]. Available: https://doi.org/10.1007/s11042-023-17203-x

[19] L. Shi, K. Du, C. Zhang, H. Ma, and W. Yan, "Lung sound recognition algorithm based on vggish-bigru," *IEEE Access*, vol. 7, pp. 139 438–139 449, Sep. 2019.

[20] Mohammed, A. L. Swapnil, M. D. Peris, I. H. Nihal, R. Khan, and M. A. Matin, "Multimodal deep learning for violence detection: Vggish and mobilevit integration with knowledge distillation on jetson nano," *IEEE Open Journal of the Communications Society*, vol. 6, pp. 2907–2925, Dec. 2024.

[21] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[22] J. Mehta, D. Gandhi, G. Thakur, and P. Kanani, "Music genre classification using transfer learning on log-based mel spectrogram," in *Proc. 5th International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2021, pp. 1101–1107.

[23] K. Palanisamy, D. Singhania, and A. Yao, "Rethinking cnn models for audio classification," 2020. [Online]. Available: https://arxiv.org/abs/2007.11154

[24] E. Tsalera, A. Papadakis, and M. Samarakou, "Comparison of pre-trained cnns for audio classification using transfer learning," *Journal of Sensor and Actuator Networks*, vol. 10, p. 72, Dec. 2021.

[25] S. Shin, J. Kim, Y. Yu, S. Lee, and K. Lee, "Self-supervised transfer learning from natural images for sound classification," *Applied Sciences*, vol. 11, p. 3043, Mar. 2021.

[26] M. Dong, "Convolutional neural network achieves human-level accuracy in music genre classification," *CoRR*, vol. abs/1802.09697, 2018. [Online]. Available: http://arxiv.org/abs/1802.09697

[27] S. Sugianto and S. Suyanto, "Voting-based music genre classification using melspectogram and convolutional neural network," in *Proc. 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Yogyakarta, Indonesia, 2019, pp. 330–333.