A Comparative Analysis of Machine Learning Models for Acoustic Classification: A Reproducible Modular Pipeline Across Three Domains

Filipe Gomes Ferreira[1], Denival Biotto Filho[1, 2], Luan Vitor Pereira Rocha[1]

[1]Faculdade de Tecnologia do Estado de São Paulo (FATEC)
Rio Claro, SP, Brazil

[2]Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP)
Piracicaba, SP, Brazil

filipegomesfrr@gmail.com, denivaldenival@gmail.com,
luanv_rocha@hotmail.com

Abstract. Acoustic classification appears in domains as different as predictive maintenance, bioacoustic monitoring, and urban sound analysis, yet most studies evaluate a single model on a single dataset. This paper presents a reproducible modular pipeline that extracts MFCCs, Mel spectrograms, and spectral and temporal descriptors, and uses it to compare six models (Decision Trees, Random Forests, SVM, MLP, CNN, and YAMNet transfer learning) under one protocol across three public-data domains: bird species (Xeno-canto), urban sounds (UrbanSound8K), and industrial valves (MIMII). The CNN over Mel spectrograms led in every domain (up to 99.6%), classical models with enriched features stayed within a few points at much lower cost, and transfer learning was not uniformly best, which we discuss together with an error asymmetry relevant to fault detection.

Resumo. A classificação acústica aparece em domínios distintos, da manutenção preditiva ao monitoramento bioacústico e à análise de sons urbanos, mas a maioria dos estudos avalia um único modelo sobre uma única base. Este artigo apresenta um pipeline modular e reprodutível que extrai MFCCs, espectrogramas Mel e descritores espectrais e temporais, e o usa para comparar seis modelos (Árvores de Decisão, Random Forests, SVM, MLP, CNN e aprendizado por transferência com YAMNet) sob um mesmo protocolo em três domínios com dados públicos: aves (Xeno-canto), sons urbanos (UrbanSound8K) e válvulas industriais (MIMII). A CNN sobre espectrogramas Mel liderou em todos os domínios (até 99,6%), os modelos clássicos enriquecidos ficaram a poucos pontos a um custo bem menor, e o aprendizado por transferência não foi uniformemente superior, o que discutimos junto a uma assimetria de erro relevante para detecção de falhas.

1. Introdução

O som produzido por sistemas físicos carrega informação sobre o estado de funcionamento desses sistemas. Motores, compressores, ventiladores e válvulas emitem assinaturas acústicas que se alteram quando seus componentes apresentam desgaste ou falha. De modo análogo, vocalizações de aves permitem identificar espécies em ambientes naturais, e o ambiente sonoro urbano pode ser decomposto em eventos reconhecíveis, como sirenes, buzinas e britadeiras [Salamon et al. 2014; Stowell 2022]. Em todos esses casos, a tarefa subjacente é a mesma: dada uma gravação de áudio, atribuir a ela uma classe a partir de um conjunto pré-definido.

Apesar dessa unidade conceitual, a literatura costuma tratar cada aplicação de forma isolada. Estudos de manutenção preditiva costumam empregar uma única arquitetura, com frequência uma Rede Neural Convolucional, sobre uma única base de máquinas [Purohit et al. 2019]. Trabalhos de bioacústica concentram-se em arquiteturas específicas para canto de aves [Stowell 2022]. Essa fragmentação dificulta comparar, de forma justa, a engenharia de atributos e os paradigmas de aprendizado, e alimenta a ideia de que cada cenário exige uma solução construída do zero. Falta, em particular, evidência empírica sobre quão bem um mesmo conjunto de modelos generaliza entre domínios acústicos distintos.

Este trabalho investiga essa lacuna a partir de uma pergunta prática: até que ponto um único pipeline modular, com uma camada comum de extração de atributos e um conjunto intercambiável de classificadores, resolve problemas de classificação acústica em domínios diferentes. Para respondê-la, o trabalho oferece três contribuições. A primeira é um pipeline modular e reprodutível, com atenção explícita a fontes comuns de erro metodológico, como o vazamento de dados na separação entre treino e teste. A segunda é um estudo comparativo de seis modelos, a saber Árvores de Decisão (DT), Random Forests (RF), Máquinas de Vetores de Suporte (SVM), Perceptrons Multicamadas (MLP), Redes Neurais Convolucionais (CNN) e aprendizado por transferência com YAMNet, avaliados sob um mesmo protocolo em três domínios com dados públicos. A terceira é uma discussão dos compromissos entre acurácia, custo computacional e confiabilidade, incluindo uma assimetria de erro observada na detecção de falhas que tem implicação direta para aplicações de manutenção.

A motivação aplicada decorre da escassez de soluções acessíveis para diagnóstico acústico em equipamentos de uso cotidiano, como válvulas, eletrodomésticos e ferramentas elétricas, cujo atendimento técnico ainda depende fortemente da inspeção auditiva presencial [Mobley 2002; Ferreira et al. 2021]. Um pipeline genérico, instanciável com bases distintas, reduz o custo de prototipagem dessas soluções e fornece um ponto de partida reprodutível para investigações futuras. O projeto que originou este estudo foi acolhido pela pré-incubadora da Faculdade de Tecnologia do Estado de São Paulo (FATEC) de Rio Claro, no contexto de aplicações de diagnóstico sonoro.

O restante do artigo está organizado da seguinte forma. A Seção 2 discute trabalhos relacionados. A Seção 3 detalha o pipeline de extração de atributos. A Seção 4

descreve os seis modelos. A Seção 5 apresenta o protocolo experimental e os resultados nos três domínios. A Seção 6 discute limitações e implicações. A Seção 7 conclui o trabalho.

2. Trabalhos Relacionados

A literatura em classificação acústica pode ser organizada em três frentes: a representação do sinal, o paradigma de aprendizado e as aplicações que motivaram a pesquisa.

2.1. Representação de sinais acústicos

Os Mel-Frequency Cepstral Coefficients (MFCCs) permanecem como representação padrão desde os trabalhos seminais de modelagem musical [Logan 2000]. Sua robustez decorre da combinação de escala perceptiva (Mel), compressão logarítmica e descorrelação por transformada discreta de cosseno, o que torna o vetor adequado a classificadores lineares e baseados em kernel. Trabalhos posteriores complementam os MFCCs com derivadas temporais (deltas de primeira e segunda ordem) e descritores estatísticos de ordem superior, observando ganhos em bases pequenas [Eyben et al. 2016].

Espectrogramas Mel são preferidos com arquiteturas convolucionais, por preservarem a estrutura tempo-frequência sobre a qual operadores convolucionais aprendem padrões locais [Choi et al. 2017]. Descritores temporais, como zero-crossing rate (ZCR) e energia RMS, e espectrais, como centroide, rolloff, bandwidth e contraste, são combinados com MFCCs para enriquecer a representação em sinais não estacionários [McFee et al. 2015].

2.2. Modelos clássicos e modelos profundos

A comparação entre modelos clássicos e arquiteturas profundas em tarefas acústicas revela um padrão recorrente. Em bases reduzidas, classificadores como SVM e Random Forests competem com modelos profundos quando a engenharia de atributos é cuidadosa, ao passo que, em bases de larga escala, arquiteturas convolucionais tendem a dominar [Hershey et al. 2017]. Em sons urbanos, Salamon et al. (2014) introduzem a base UrbanSound8K e mostram que classificadores clássicos sobre atributos espectrais constituem uma linha de base forte, depois superada por redes convolucionais com aumento de dados [Salamon e Bello 2017]. Em manutenção industrial, Purohit et al. (2019) demonstram desempenho elevado de CNNs na detecção de anomalias da base MIMII, observando que modelos clássicos bem ajustados continuam competitivos a um custo menor.

2.3. Aprendizado por transferência em áudio

O aprendizado por transferência tem sido usado para mitigar a escassez de dados rotulados. Hershey et al. (2017) introduzem o YAMNet, uma CNN baseada em MobileNet treinada na base AudioSet, com mais de 5.000 horas de áudio em 521 classes. Tsalera et al. (2021) mostram que embeddings do YAMNet combinados a classificadores simples atingem desempenho comparável ou superior a modelos treinados do zero em

diversas tarefas downstream. Apesar disso, há poucos estudos que avaliem um mesmo conjunto de modelos sobre múltiplos domínios com um protocolo unificado, condição necessária para discutir a generalidade dessas soluções, lacuna que este trabalho procura preencher.

3. Pipeline de Extração de Atributos Acústicos

O pipeline proposto tem quatro estágios sequenciais, mostrados na Figura 1: pré-processamento, extração de atributos, agregação por estatísticas descritivas e composição do vetor de entrada. A implementação foi feita em Python sobre a biblioteca Librosa [McFee et al. 2015]. Cada estágio expõe uma interface uniforme, o que permite ativar ou desativar descritores e classificadores sem alterações estruturais, propriedade que viabiliza o uso do pipeline como ambiente experimental controlado para o estudo das Seções 4 e 5.

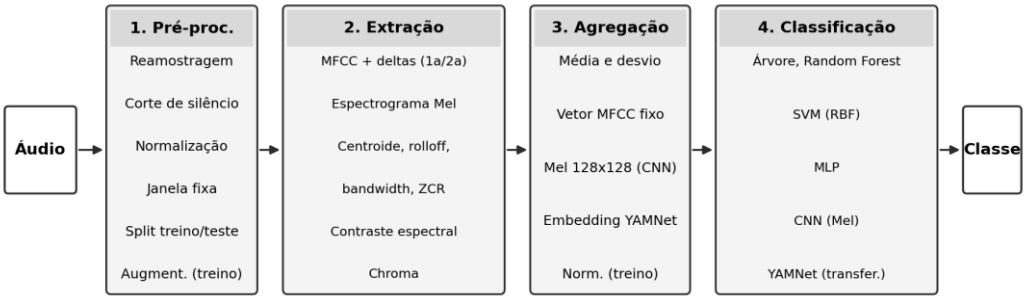


Figura 1. Visão geral do pipeline modular de classificação acústica.

3.1. Pré-processamento e separação treino/teste

Os áudios são padronizados em uma taxa de amostragem única por reamostragem com filtro polifásico e convertidos para mono. Trechos de silêncio prolongado são truncados por um limiar de energia adaptativo, calculado como fração do RMS médio do sinal, e a duração é limitada a uma janela fixa, de modo que o tamanho do vetor de atributos independa da duração original. A amplitude é normalizada por sinal.

A ordem das operações recebe atenção especial. A separação entre treino e teste é feita no nível dos arquivos, antes de qualquer aumento de dados ou ajuste de normalização. Esse cuidado evita o vazamento de informação que ocorre quando versões aumentadas de um mesmo áudio aparecem ao mesmo tempo no treino e no teste, situação que infla artificialmente a acurácia. O aumento de dados, quando ativado, é aplicado apenas ao treino e somente para bases pequenas, por adição de ruído, alteração de tom, variação de velocidade e ganho aleatório. Esse ponto é reportado de forma explícita porque é uma causa frequente de resultados superestimados em trabalhos de classificação de áudio.

3.2. Extração de atributos

Para os modelos clássicos e o MLP, o pipeline extrai um conjunto rico de atributos por gravação. A base é formada por 40 MFCCs (ou 20, em bases de taxa de amostragem mais baixa), acompanhados de suas derivadas de primeira e segunda ordem. A esses somam-se descritores espectrais e temporais complementares:

- Centroide, rolloff e bandwidth espectrais, que descrevem o centro de massa, a borda superior e a largura efetiva do espectro.
- Zero-Crossing Rate (ZCR), associado a conteúdo percussivo e a ruído de banda larga.
- Contraste espectral, que mede a diferença entre picos e vales em sub-bandas de frequência.
- Cromagrama, que captura informação harmônica e tonal.

Para a CNN, o pipeline gera espectrogramas Mel com 128 bandas, em escala logarítmica de decibéis, redimensionados para 128 por 128. Para o YAMNet, o áudio reamostrado para 16 kHz alimenta o modelo, que produz embeddings de quadro posteriormente agregados.

3.3. Agregação e normalização

A maioria dos atributos é gerada como série temporal de tamanho variável. Para os modelos clássicos, que exigem vetores fixos, cada série é substituída por sua média e desvio padrão ao longo do tempo. Para o YAMNet, os embeddings de quadro são agregados por concatenação de média e desvio padrão, produzindo um vetor de 2048 dimensões por gravação.

A normalização é o segundo ponto sensível a vazamento. Para os atributos MFCC e para os embeddings, um normalizador padrão é ajustado apenas no treino e depois aplicado ao teste. Para a CNN, adota-se normalização global por escore-z, com média e desvio do treino, em vez da normalização por imagem usada em versões iniciais, que provocava o colapso do treinamento. Em experimentos preliminares, substituir o uso isolado das médias dos MFCCs por essa agregação mais rica elevou de forma expressiva a acurácia dos classificadores clássicos.

3.4. Vetor de entrada

Ao final, cada modelo recebe a representação mais adequada ao seu paradigma. Os modelos clássicos (DT, RF, SVM) e o MLP recebem o vetor MFCC enriquecido e normalizado. A CNN recebe o espectrograma Mel em forma matricial. O YAMNet alimenta uma rede densa a partir de seus embeddings. Essa organização permite comparar, sob o mesmo protocolo, desde árvores rasas até aprendizado por transferência.

4. Modelos Avaliados

Foram avaliados seis modelos em três grupos por paradigma. Em todos os casos, o desbalanceamento entre classes foi tratado por ponderação proporcional à frequência de

cada classe, e os modelos foram treinados e avaliados sobre a mesma partição de treino e teste, garantindo comparação justa.

4.1. Modelos clássicos

A Árvore de Decisão (DT) usou critério de Gini, profundidade máxima 12, mínimo de 10 amostras por folha e poda por complexidade de custo. O Random Forest (RF) usou 300 árvores com seleção aleatória de atributos por divisão e ponderação balanceada por subamostra. A Máquina de Vetores de Suporte (SVM) usou kernel RBF, C igual a 10 e gamma em escala, com ponderação balanceada de classes.

4.2. Perceptron multicamadas

O Perceptron Multicamadas (MLP) teve três camadas ocultas de 512, 256 e 128 unidades, com ativação ReLU, normalização em lote, dropout entre 0,3 e 0,4 e regularização L2. O treinamento usou Adam, entropia cruzada categórica esparsa, parada antecipada por perda de validação e redução adaptativa da taxa de aprendizado.

4.3. Modelos baseados em aprendizado profundo

A Rede Neural Convolucional (CNN) recebe o espectrograma Mel e tem quatro blocos convolucionais com 32, 64, 128 e 256 filtros, cada bloco com duas convoluções de kernel 3 por 3, normalização em lote e max-pooling, seguidos de agrupamento global por média e de uma camada densa de 256 unidades antes do softmax.

O YAMNet [Hershey et al. 2017] é usado como extrator de embeddings: o áudio reamostrado para 16 kHz é encaminhado ao modelo, cujos vetores de quadro são agregados por média e desvio padrão. Sobre esses embeddings treina-se uma rede densa de duas camadas ocultas (512 e 256 unidades) com normalização em lote e dropout, em um cenário clássico de aprendizado por transferência no qual apenas a cabeça de classificação é ajustada.

5. Experimentos e Resultados

Os experimentos avaliam como os seis modelos se comportam quando o pipeline é instanciado em domínios com complexidades e números de classes distintos. Os três cenários são descritos a seguir.

5.1. Conjuntos de dados

Todos os dados de áudio provêm de bases públicas amplamente usadas pela comunidade, o que favorece a reprodutibilidade:

- Sons urbanos. Subconjunto da base UrbanSound8K [Salamon et al. 2014], com 10 classes de eventos urbanos: ar condicionado, buzina de carro, crianças brincando, latido de cachorro, britadeira, motor em marcha lenta, música de rua, furadeira, sirene e tiro de arma de fogo.

- Cantos de aves. Conjunto bioacústico montado a partir de gravações da plataforma colaborativa Xeno-canto [Xeno-canto Foundation 2026], com duas espécies de ocorrência comum no Brasil, o Bem-te-vi e o Tico-tico.
- Válvulas industriais. Subconjunto da base MIMII [Purohit et al. 2019], de sons de máquinas industriais, com gravações de válvulas em duas condições, regular e falha.

Em todos os cenários aplicou-se divisão estratificada de 80% para treino e 20% para teste, com a separação no nível dos arquivos, antes de qualquer aumento de dados. Um subconjunto de 15% do treino foi reservado para validação durante o ajuste das redes neurais.

5.2. Métricas

Foram reportados a acurácia e o F1-score ponderado no teste, além do gap entre acurácia de treino e teste como indicador de sobreajuste. O F1-score ponderado foi adotado como métrica principal por sua robustez frente a eventual desbalanceamento. Como os cenários têm números de classes diferentes, a linha de chance varia: 10% nos sons urbanos e 50% nas aves e nas válvulas.

5.3. Comparação entre modelos

A Tabela 1 resume a acurácia e o F1-score ponderado dos seis modelos nos três cenários, e a Figura 2 apresenta as acurácias de forma comparativa. Em todos os domínios, a CNN sobre espectrogramas Mel obteve o melhor resultado, seguida de perto pelos modelos baseados em MFCC enriquecido e pelo YAMNet.

Tabela 1. Acurácia (Ac.) e F1-score ponderado (F1), em %, no conjunto de teste, nos três domínios.

Modelo	Urb. Ac.	Urb. F1	Aves Ac.	Aves F1	Vál. Ac.	Vál. F1
Árvore	67.5	67.4	88.3	88.3	92.3	92.5
Random Forest	91.9	91.9	93.3	93.3	96.0	95.5
SVM	94.0	94.0	94.2	94.2	97.3	97.1
MLP	95.0	95.0	92.5	92.5	98.7	98.6
CNN	96.9	96.9	97.5	97.5	99.6	99.6
YAMNet	93.4	93.4	87.5	87.5	95.5	95.2

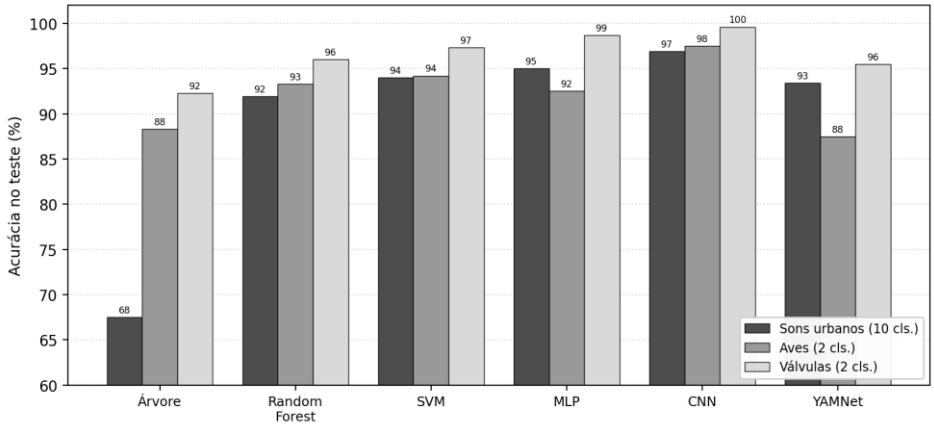


Figura 2. Acurácia no teste dos seis modelos nos três domínios.

Nos sons urbanos, que têm 10 classes e portanto a tarefa mais difícil sob a ótica da chance, a CNN atingiu 96,9% de acurácia, com o MLP em 95,0% e o SVM em 94,0%. O YAMNet alcançou 93,4%, próximo dos modelos clássicos enriquecidos, ainda que tenha sido projetado para sons ambientais semelhantes aos da base. A Árvore isolada ficou bem atrás, em 67,5%, como esperado para um modelo de baixa capacidade diante de muitas classes.

Nas aves, a CNN novamente liderou, com 97,5%, seguida por SVM e Random Forest. O YAMNet teve o desempenho mais baixo do cenário, 87,5%, com a maior parte do erro concentrada na espécie Tico-tico, como detalha a Figura 3. Esse resultado sugere que, em uma base pequena e específica, os embeddings genéricos do YAMNet capturam menos discriminação do que atributos espectrais ajustados ao problema.

Nas válvulas, todos os modelos superaram 92% de acurácia, refletindo a boa separabilidade entre os estados regular e de falha. A CNN praticamente saturou a tarefa, com 99,6%, e o MLP atingiu 98,7%. Um ponto de atenção, contudo, aparece na análise por classe e motiva a subseção seguinte.

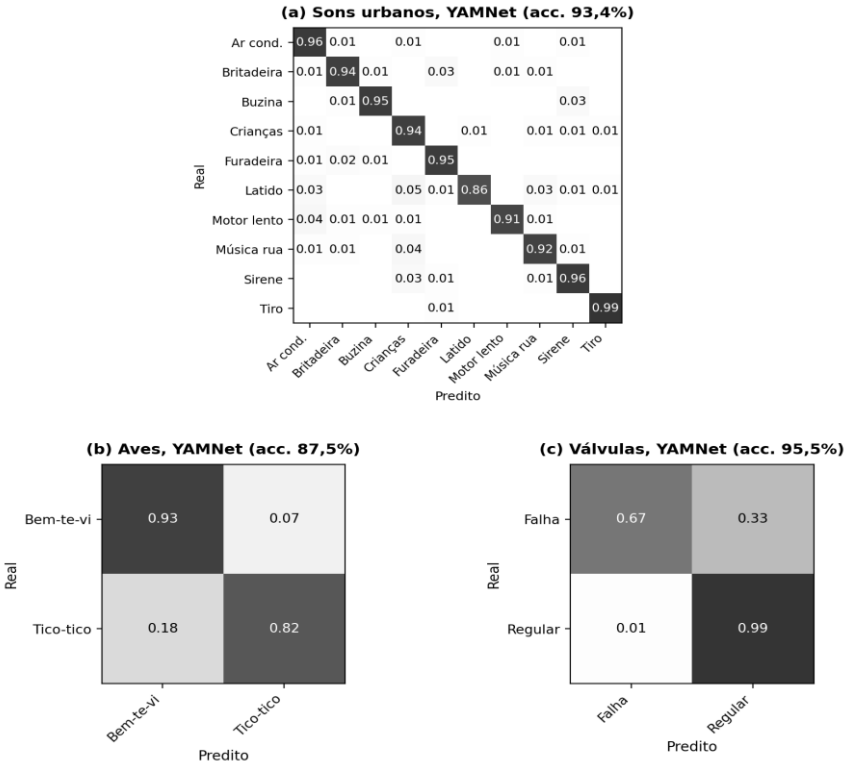
5.4. Aprendizado por transferência e assimetria de erro

A Figura 3 apresenta as matrizes de confusão do YAMNet nos três domínios. Nos sons urbanos, o modelo distribui bem os acertos entre as 10 classes, com a maior confusão entre latido de cachorro e crianças brincando. Nas aves, fica clara a dificuldade com o Tico-tico, classificado corretamente em apenas 82% dos casos. Nas válvulas, embora a acurácia global seja alta, o modelo acerta somente 67% das amostras de falha, confundindo um terço delas com o estado regular.

Essa assimetria é o achado mais relevante para aplicações de manutenção. Em diagnóstico de falhas, deixar de detectar um defeito, o falso negativo, costuma ser muito mais grave do que um alarme falso. Um modelo com acurácia global de 95,5%, mas que detecta apenas dois terços das falhas reais, pode ser inadequado para esse uso, ainda que pareça excelente sob a métrica agregada. O resultado reforça a importância de avaliar o

recall da classe crítica, e não apenas a acurácia, ponto alinhado às discussões sobre IA confiável e centrada na aplicação.

Figura 3. Matrizes de confusão do YAMNet nos três domínios. As células indicam a fração de cada classe real (linha) atribuída a cada classe predita (coluna). Destaque para a baixa taxa de acerto da classe falha em (c) e da espécie Tico-tico em (b).



5.5. Sobreajuste e custo

Além da acurácia, o gap entre treino e teste foi monitorado como indicador de sobreajuste. A Tabela 2 reúne esse gap para quatro modelos representativos. A CNN apresentou os menores valores, com apenas 0,4 ponto percentual nas válvulas, sinal de boa generalização. O YAMNet registrou o maior gap nas aves, próximo de 12 pontos, coerente com o sobreajuste em uma base pequena.

Tabela 2. Gap entre acurácia de treino e teste (pontos percentuais) por domínio.

Modelo	Urbano	Aves	Válvulas
SVM	5.5	5.8	2.7
MLP	4.2	7.5	1.4
CNN	2.6	2.5	0.4
YAMNet	5.8	11.8	3.9

Do ponto de vista de custo, os modelos clássicos sobre atributos MFCC têm inferência muito mais leve do que a CNN e o YAMNet, pois operam sobre um vetor compacto em vez de uma matriz ou de uma rede pré-treinada de grande porte. Em cenários embarcados,

com inferência em dispositivos de baixa potência, essa diferença favorece SVM e Random Forest, que mantiveram acurácia acima de 92% em todos os cenários de duas classes.

6. Discussão

6.1. Compromissos na escolha do modelo

Três padrões emergem da análise. Primeiro, a CNN sobre espectrogramas Mel foi a mais acurada nos três domínios e a que menos sobreajustou, o que a torna a escolha preferencial quando há dados e capacidade de processamento. Segundo, os modelos clássicos com atributos enriquecidos ficaram muito próximos da CNN nos cenários de duas classes, oferecendo um compromisso atraente entre desempenho e custo. Terceiro, o aprendizado por transferência com YAMNet teve desempenho irregular, forte nos sons urbanos, intermediário nas válvulas e fraco nas aves, o que indica que sua vantagem depende da proximidade entre o domínio de pré-treinamento e o domínio alvo.

6.2. Limitações

Algumas limitações merecem registro. Os cenários de aves e válvulas têm apenas duas classes, o que simplifica a tarefa frente a uma operação real com muitas espécies ou múltiplos modos de falha. As bases de aves e válvulas usadas são de tamanho reduzido, o que ajuda a explicar o sobreajuste do YAMNet. Não foi avaliada a condição de domínio cruzado, em que o modelo é treinado em uma distribuição de gravações e testado em outra, comum em manutenção quando a base provém de poucos equipamentos. Por fim, o estudo se concentra em atributos tradicionais e em uma CNN treinada do zero, deixando arquiteturas recentes baseadas em Transformers, como o Audio Spectrogram Transformer [Gong et al. 2021], para trabalhos futuros.

6.3. Implicações práticas

Os resultados sugerem uma estratégia em camadas para sistemas de diagnóstico acústico. Em dispositivos embarcados de baixo custo, classificadores clássicos com atributos enriquecidos atendem à maioria das demandas com baixa latência. Em servidores ou nuvem, a CNN oferece a maior margem de segurança preditiva. A modularidade do pipeline facilita essa coexistência, pois o mesmo conjunto de atributos alimenta tanto um SVM compacto quanto uma rede convolucional. No caso de manutenção, a assimetria de erro observada no YAMNet indica que a métrica de avaliação deve priorizar o recall da classe de falha, e que a escolha do modelo não deve se basear apenas na acurácia agregada.

7. Conclusão

Este trabalho apresentou um pipeline modular e reprodutível de classificação acústica, usado como ambiente controlado para comparar seis modelos de aprendizado de máquina em três domínios com dados públicos: sons urbanos (UrbanSound8K), cantos de aves (Xeno-canto) e válvulas industriais (MIMII). A CNN sobre espectrogramas Mel obteve a

melhor acurácia em todos os cenários, com 99,6% nas válvulas, 97,5% nas aves e 96,9% nos sons urbanos, e baixo sobreajuste. Os modelos clássicos enriquecidos ficaram poucos pontos atrás, a um custo muito menor, e o aprendizado por transferência com YAMNet mostrou desempenho dependente da afinidade entre domínios. A análise por classe revelou uma assimetria de erro relevante para manutenção, em que falhas reais foram subdetectadas apesar da alta acurácia global.

Trabalhos futuros incluem a avaliação em cenários de domínio cruzado, a ampliação do número de classes nos domínios industrial e bioacústico, a incorporação de arquiteturas baseadas em Transformers e a construção de uma base proprietária de sons reais de equipamentos eletromecânicos, voltada a manutenção preditiva.

Referências

Choi, K., Fazekas, G., Sandler, M. and Cho, K. (2017) "Convolutional Recurrent Neural Networks for Music Classification", In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), p. 2392-2396.

Eyben, F., Scherer, K. R., Schuller, B. W., et al. (2016) "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing", IEEE Transactions on Affective Computing, v. 7, n. 2, p. 190-202.

Ferreira, L. L., Oliveira, A., Teixeira, N., Bulut, B., Landeck, J., Morgado, N. and Sousa, O. (2021) "Predictive Maintenance of home appliances: Focus on Washing Machines", In: Proceedings of the School of Engineering of the Polytechnical Institute of Porto / University of Coimbra.

Gong, Y., Chung, Y.-A. and Glass, J. (2021) "AST: Audio Spectrogram Transformer", In: Proceedings of Interspeech.

Hershey, S., Chaudhuri, S., Ellis, D. P. W., et al. (2017) "CNN Architectures for Large-Scale Audio Classification", In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), p. 131-135.

Logan, B. (2000) "Mel Frequency Cepstral Coefficients for Music Modeling", In: International Symposium on Music Information Retrieval.

McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., et al. (2015) "librosa: Audio and music signal analysis in Python", In: Proceedings of the 14th Python in Science Conference, p. 18-25.

Mobley, R. K. (2002) "An Introduction to Predictive Maintenance", 2nd edition, Butterworth-Heinemann.

Purohit, H., Tanabe, R., Ichige, K., Endo, T., Nikaido, Y., Suefusa, K. and Kawaguchi, Y. (2019) "MIMII Dataset: Sound dataset for malfunctioning industrial machine investigation and inspection", In: Proceedings of the 4th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE).

Salamon, J. and Bello, J. P. (2017) "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification", IEEE Signal Processing Letters, v. 24, n. 3, p. 279-283.

Salamon, J., Jacoby, C. and Bello, J. P. (2014) "A Dataset and Taxonomy for Urban Sound Research", In: Proceedings of the 22nd ACM International Conference on Multimedia, p. 1041-1044.

Stowell, D. (2022) "Computational bioacoustics with deep learning: a review and roadmap", PeerJ, v. 10, e13152.

Tsalera, E., Papadakis, A. and Samarakou, M. (2021) "Comparison of pre-trained CNNs for audio classification using transfer learning", Journal of Sensor and Actuator Networks, v. 10, n. 4, p. 72.

Xeno-canto Foundation. (2026) "Xeno-canto: Sharing bird sounds from around the world", https://xeno-canto.org/, Janeiro.