AVALIAÇÃO DO EFEITO DO BALANCEAMENTO DE CLASSES NO DESEMPENHO DE UMA *RESNET-50* PARA CLASSIFICAÇÃO DE SONOGRAMAS DE *Myotis lavali* E *Lasiurus blossevillii*

Davi Palma Fonseca[1] - davi03palma@gmail.com
Arthur Mendes Fernandes[1] - arthurmendes960@gmail.com
Artur Pereira Neto[1] - artur.neto@ifnmg.edu.br
Angelina de Meiras-Ottoni[1] - angelina.m.ottoni@gmail.com
Allysson Steve Mota Lacerda[1] - steve.lacerda@unimontes.br
Luiz Alberto Dolabela Falcão[1] - luizdolabelafalcao@gmail.com
[1]Universidade Estadual de Montes Claros, UNIMONTES – Montes Claros, MG, Brazil

Resumo.

O desbalanceamento de classes constitui um dos principais desafios no treinamento de classificadores supervisionados, podendo favorecer a classe majoritária e comprometer a capacidade do modelo de reconhecer exemplos menos representados. Este trabalho investiga o efeito desse desbalanceamento no desempenho de uma ResNet-50 pré-treinada, com camadas convolucionais congeladas, aplicada à classificação binária de sonogramas de Myotis lavali e Lasiurus blossevillii. A partir de 446 sonogramas brutos coletados no Parque Estadual da Lapa Grande (Montes Claros, MG), foram construídos dois conjuntos de dados com os mesmos hiperparâmetros (resize de 256×224 pixels, lote 16 e taxa de aprendizado de 10^{-4}): um balanceado por subamostragem da classe majoritária (5.012 imagens, 2.506 por classe) e outro desbalanceado, preservando todos os recortes disponíveis (8.924 imagens, proporção 71,9%/28,1%). O modelo balanceado alcançou F1-Score macro de 0,8310 e acurácia de 83,11%, enquanto o desbalanceado obteve F1-Score macro de 0,8190 e acurácia de 86,25%, com recall de apenas 0,662 para a espécie minoritária. Embora o desbalanceamento produza maior acurácia global, essa métrica, isoladamente, pode ocultar perdas na classe minoritária. Ainda que o delineamento não permita atribuir essa diferença exclusivamente ao balanceamento, verificou-se que a subamostragem da classe majoritária contribuiu para um desempenho mais homogêneo entre as espécies.

Keywords: *Quirópteros, Bioacústica, Redes neurais convolucionais, Balanceamento de classes, Aprendizado por transferência*

1. INTRODUÇÃO

A identificação automática de espécies a partir de chamados de ecolocalização vem se consolidando como ferramenta importante para o monitoramento acústico da biodiversidade, sobretudo quando o volume de gravações inviabiliza a análise manual. Redes convolucionais treinadas por transferência de aprendizado apresentam resultados promissores na classificação de representações espectrais de áudio, tanto para aves (Salamon et al., 2017) quanto para morcegos (Vogelbacher et al., 2023), embora o desempenho desses modelos dependa também das características do conjunto de treinamento, entre elas a distribuição das amostras entre classes.

No monitoramento de morcegos, a identificação automática é especialmente relevante devido à elevada variabilidade intraespecífica dos chamados e à sobreposição acústica entre espécies filogeneticamente próximas, que dificultam a identificação manual (Schwab et al., 2022). Além disso, o monitoramento contínuo em unidades de conservação produz grandes volumes de gravações, tornando os classificadores automáticos um complemento importante a métodos tradicionais, como as redes de neblina (Fróes, 2025). No Brasil, são reconhecidas 186 espécies de morcegos em nove famílias (Garbino et al., 2024), entre elas *Myotis lavali* e *Lasiurus blossevillii*, cujos padrões acústicos permanecem pouco explorados em classificação automática.

Diante disso, este trabalho investiga o efeito do balanceamento de classes sobre o desempenho de uma *ResNet-50* pré-treinada (He et al., 2016), comparando um conjunto balanceado por subamostragem da classe majoritária e um conjunto desbalanceado que preserva todos os recortes disponíveis, sob os mesmos hiperparâmetros de treinamento, analisando o impacto do balanceamento tanto nas métricas agregadas quanto no desempenho individual de cada classe, aspecto frequentemente mascarado pela acurácia global.

2. METODOLOGIA

2.1 ÁREA DE ESTUDO E COLETA DE DADOS

As gravações acústicas utilizadas neste estudo foram obtidas no Parque Estadual da Lapa Grande, em Montes Claros (MG), com gravadores ultrassônicos *Song Meter SM4BAT Mini* (Wildlife Acoustics), conforme Fróes (2025). As campanhas de campo, a aquisição dos registros e a identificação das espécies foram conduzidas pelo Laboratório de Ecologia Evolutiva (LEE) da Unimontes, que disponibilizou 446 sonogramas rotulados de *Myotis lavali* e *Lasiurus blossevillii*, base do conjunto de dados dos experimentos.

2.2 PRÉ-PROCESSAMENTO E PREPARAÇÃO DO CONJUNTO DE IMAGENS

Como um mesmo sonograma podia conter múltiplas vocalizações, a equipe do Laboratório de Inteligência Computacional Aplicada (LICA) extraiu recortes individuais, correspondentes a cada vocalização, para padronizar as imagens de entrada do classificador. Essa etapa resultou em 6.418 recortes de *Myotis lavali* e 2.506 de *Lasiurus blossevillii*, refletindo a maior frequência de registros da primeira espécie, base para os cenários experimentais descritos a seguir.

2.3 COMPOSIÇÃO DOS CONJUNTOS EXPERIMENTAIS

A partir do banco de recortes descrito na Seção 2.2, foram construídos dois conjuntos de dados diferindo apenas na distribuição das amostras entre as classes.

– **Experimento balanceado (E1):** foi realizada subamostragem aleatória da classe majoritária (*M. lavali*) até que ambas as espécies passassem a possuir 2.506 imagens, totalizando 5.012 recortes.

– **Experimento desbalanceado (E2):** foram preservados todos os recortes disponíveis, resultando em 8.924 imagens, sendo 6.418 pertencentes à classe *M. lavali* (71,9%) e 2.506 à classe *L. blossevillii* (28,1%).

Em ambos, os conjuntos foram particionados em treino e validação (70%/30%, estratificado por classe): E1 ficou com 3.508 imagens de treino e 1.504 de validação; E2, com 6.247 de treino e 2.677 de validação. Todas as imagens foram redimensionadas para 256×224 pixels antes do treinamento.

2.4 ARQUITETURA DA REDE E CONFIGURAÇÃO DO TREINAMENTO

Os dois experimentos empregaram os mesmos hiperparâmetros, correspondentes ao ponto central de um planejamento experimental mais amplo do grupo (Tabela 1): uma *ResNet-50* (He et al., 2016) pré-treinada com pesos do ImageNet (Russakovsky et al., 2015), camadas convolucionais congeladas durante todo o treinamento e ajuste apenas da camada final (`Linear(2048→128)`, ReLU, `Linear(128→2)`), por 3.000 épocas, com entropia cruzada e otimizador Adam (Kingma e Ba, 2015). Implementação em Python, com PyTorch e pesos pré-treinados do *torchvision*.

Tabela 1: Hiperparâmetros de treinamento, comuns aos dois experimentos.

Hiperparâmetro	Valor
Resize (R)	256×224 px
Batch size (B)	16
Learning rate (L)	1×10^{-4}
Otimizador	Adam
Função de perda	Entropia cruzada
Épocas	3.000
Camadas convolucionais	Congeladas (*frozen*)

Fonte: Elaborado pelos autores.

Todos os hiperparâmetros de treinamento foram mantidos constantes entre os dois experimentos, de forma que a única variável experimental avaliada fosse a composição do conjunto de treinamento.

2.5 CRITÉRIO DE SELEÇÃO DO MODELO E MÉTRICAS DE AVALIAÇÃO

Em cada experimento, o modelo selecionado para comparação foi aquele correspondente à menor perda observada no conjunto de validação daquele próprio experimento. Como os conjuntos de validação apresentam diferentes distribuições entre as classes, os valores absolutos

de perda e de acurácia refletem, em parte, essa diferença de composição. Assim, a comparação entre os experimentos foi baseada principalmente no F1-Score macro, por atribuir o mesmo peso às duas classes e fornecer uma avaliação mais apropriada em cenários com desbalanceamento.

3. RESULTADOS E DISCUSSÃO

3.1 DESEMPENHO GLOBAL NA ÉPOCA SELECIONADA

A Tabela 2 apresenta as métricas obtidas pelos modelos selecionados em cada experimento, correspondentes à menor perda de validação. O experimento E2 (desbalanceado) apresentou menor perda de validação (0,3254) e maior acurácia global (86,25%) em comparação ao experimento E1 (balanceado), que obteve perda de 0,3932 e acurácia de 83,11%. Entretanto, essa superioridade não se manteve quando a avaliação foi realizada por meio do F1-Score macro, que apresentou valor ligeiramente superior para o experimento balanceado (0,8310 contra 0,8190).

Esse comportamento evidencia que métricas agregadas podem conduzir a interpretações distintas do desempenho do classificador dependendo da distribuição das classes no conjunto de validação. No experimento E2, cerca de 72% das amostras pertencem à classe *M. lavali*, de modo que um desempenho elevado nessa classe exerce influência significativa sobre a acurácia global. Em contrapartida, o F1-Score macro atribui a mesma importância às duas espécies avaliadas, permitindo uma comparação mais equilibrada entre os experimentos.

Tabela 2: Comparação das métricas globais na época de menor perda de validação de cada experimento.

Métrica (validação)	**Balanceado (ép. 829)**	**Desbalanceado (ép. 265)**
Perda	0,3932	0,3254
Acurácia	83,11%	86,25%
Precisão macro	0,8319	0,8454
Recall macro	0,8311	0,8015
F1-Score macro	**0,8310**	**0,8190**

Fonte: Elaborado pelos autores.

3.2 DESEMPENHO POR CLASSE

A análise das métricas por classe (Tabela 3) permite compreender o comportamento observado nas métricas globais. Enquanto o experimento desbalanceado apresentou desempenho superior para *M. lavali*, espécie majoritária no conjunto de treinamento, o desempenho para *L. blossevillii* foi significativamente inferior.

O efeito mais evidente ocorreu na métrica de recall da classe minoritária, que passou de 0,807 no experimento balanceado para 0,662 no experimento desbalanceado, correspondendo a uma redução de aproximadamente 14,5 pontos percentuais. Em termos práticos, isso significa que o modelo treinado com o conjunto desbalanceado deixou de identificar corretamente uma parcela substancialmente maior das vocalizações pertencentes à espécie menos representada.

Por outro lado, o recall de *M. lavali* aumentou de 0,855 para 0,941, indicando que o modelo passou a privilegiar a identificação da classe majoritária. Esse comportamento é compatível com o efeito esperado em cenários de desbalanceamento, nos quais a função de perda tende a ser mais influenciada pela classe com maior número de exemplos durante o treinamento.

Tabela 3: Métricas de validação por classe, na época selecionada de cada experimento.

Experimento / Classe	**Precisão**	**Recall**	**F1-Score**	**Suporte**
Balanceado – *M. lavali*	0,816	0,855	0,835	752
Balanceado – *L. blossevillii*	0,848	0,807	0,827	752
Desbalanceado – *M. lavali*	0,877	0,941	0,908	1.925
Desbalanceado – *L. blossevillii*	0,814	0,662	0,730	752

Fonte: Elaborado pelos autores.

3.3 CURVAS DE TREINAMENTO

A Figura 1 mostra a curva de perda do experimento balanceado: a perda de validação caiu junto com a de treino nas primeiras épocas, estabilizou-se e atingiu seu ponto mínimo na época 829, voltando a subir depois disso, o que caracteriza sobreajuste (*overfitting*) progressivo a partir desse ponto. A Figura 2 mostra o mesmo comportamento no experimento desbalanceado, porém com sobreajuste começando bem antes, por volta da época 265, o que é consistente com um conjunto de validação proporcionalmente menor e mais dominado pela classe majoritária.

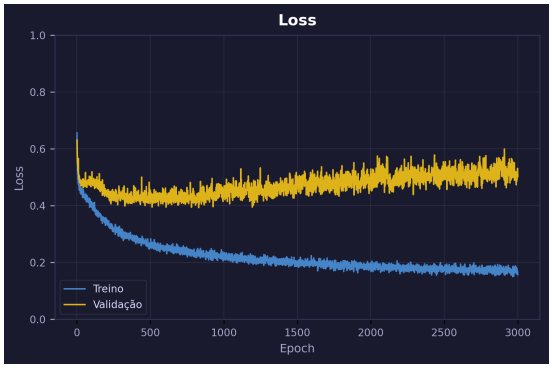

Figura 1: Perda de treino/validação – balanceado (E1).

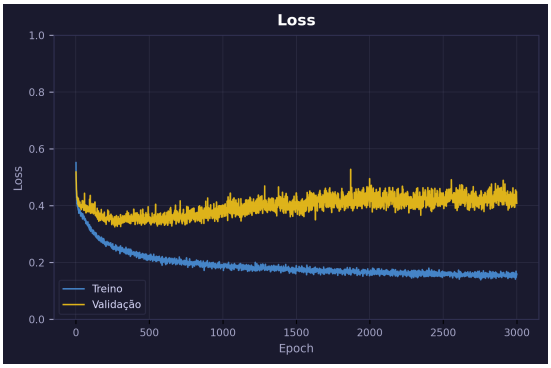

Figura 2: Perda de treino/validação – desbalanceado (E2).

Fonte: Elaborado pelos autores.

3.4 DISCUSSÃO

O balanceamento por subamostragem não melhorou a acurácia global nem a perda de validação, mas produziu um classificador mais homogêneo entre as duas espécies, sobretudo pelo aumento do *recall* da classe minoritária, reforçando a importância de métricas por classe sob distribuição desigual. Essa leitura deve considerar o delineamento: como o balanceamento também reduziu o volume total de treino, os experimentos diferem tanto na distribuição das classes quanto no volume de dados, o que impede atribuir a diferença exclusivamente ao balanceamento. Ainda assim, é um indicativo consistente de que reduzir o desbalanceamento favorece um desempenho mais equilibrado entre as espécies.

Em comparação com (Schwab et al., 2022), que reportou acurácia de 96,13% com uma *ResNet-50* para 18 espécies europeias, os modelos deste trabalho tiveram desempenho inferior; a comparação deve ser vista com cautela, pois os estudos diferem no volume de dados, no número de espécies e na estratégia de treinamento. Parte da diferença também pode vir de se manter as camadas convolucionais da *ResNet-50* congeladas: reduz custo computacional e

preserva os pesos do pré-treinamento no ImageNet, mas limita a adaptação ao domínio dos sonogramas (Zualkernan et al., 2020; Vogelbacher et al., 2023).

Assim, a acurácia global isolada pode passar uma impressão otimista sobre classificadores treinados com dados desbalanceados; em monitoramento de espécies pouco frequentes, *recall* por classe e F1-Score macro avaliam melhor a capacidade do modelo de identificar todas as classes.

4. CONSIDERAÇÕES FINAIS

Este trabalho investigou o efeito do balanceamento de classes sobre o desempenho de uma *ResNet-50* pré-treinada na classificação de vocalizações de *Myotis lavali* e *Lasiurus blossevillii*, comparando dois cenários a partir do mesmo banco de dados: um balanceado por subamostragem da classe majoritária e outro desbalanceado, com todos os recortes disponíveis, resultando também em volumes de dados distintos.

O desbalanceado produziu maior acurácia global, enquanto o balanceado teve desempenho mais homogêneo entre as espécies (maior F1-Score macro e maior *recall* da classe minoritária), reforçando que métricas agregadas exigem cautela sob distribuição desigual. Como o balanceamento alterou também o volume de dados, e como foi usada apenas uma divisão treino-validação sem teste independente ou repetição com sementes diferentes, os resultados são um indicativo consistente, mas não isolam completamente o efeito do balanceamento.

Como perspectivas futuras, pretende-se repetir a comparação mantendo constante o volume de dados entre os cenários, avaliar outras técnicas de balanceamento, explorar *fine-tuning* parcial da *ResNet-50* e incorporar parada antecipada ao treinamento.

Agradecimentos

Os autores agradecem à FAPEMIG pelo apoio financeiro ao projeto de pesquisa.

REFERENCES

Fróes, L. B. R. (2025), *"Pelos ares e pelos ecos: explorando a riqueza de morcegos em ambientes em regeneração"*, Dissertação (Mestrado em Biodiversidade e Uso dos Recursos Naturais), Universidade Estadual de Montes Claros, Montes Claros.

Garbino, G. S. T. et al. (2024), *"Updated checklist of bats (Mammalia: Chiroptera) from Brazil"*, Zoologia, v. 41, e23073.

He, K. et al. (2016), *"Deep residual learning for image recognition"*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, p. 770–778.

Kingma, D.P.; Ba, J. (2015), *"Adam: A Method for Stochastic Optimization"*, Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego.

Russakovsky, O. et al. (2015), *"ImageNet large scale visual recognition challenge"*, International Journal of Computer Vision, v. 115, n. 3, p. 211–252.

Salamon, J. et al. (2017), *"Fusing shallow and deep learning for bioacoustic bird species classification"*, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), p. 141–145.

Schwab, E. et al. (2022), *"Automated bat call classification using deep convolutional neural networks"*, Bioacoustics, v. 32, n. 1, p. 1–16.

Vogelbacher, M. et al. (2023), *"Deep learning for recognizing bat species and bat behavior in audio recordings"*, Proceedings of the IEEE Swiss Conference on Data Science, p. 50–57.

Zualkernan, I. et al. (2020), *"A tiny CNN architecture for identifying bat species from echolocation calls"*, Proceedings of the IEEE/ITU International Conference on Artificial Intelligence for Good, p. 81–86.