

Impacto do balanceamento e regularização na segmentação semântica de imagens histopatológicas

Eduardo Henrique Souza Paraíso¹, Alexei Manso Correa Machado^{1,2}

¹Programa de Pós-graduação em Informática - PUC Minas, Brasil

²Departamento de Anatomia e Imagem - UFMG, Brasil

eduardo.paraiso@sga.pucminas.br, alexeimcmachado@gmail.com

Abstract. *This study investigates the impact of class balancing and regularization on improving diagnostic agreement in histological images. For example, U-Net models applied to the Prostate Cancer Grade Assessment dataset show that class balancing, combined with traditional loss functions, increases image agreement by up to 6 percentage points. Combining balancing with Focal Loss boosts classification agreement by an average of 13 percentage points compared to using imbalanced datasets with traditional loss functions. A case study on the analysis of prostate Gleason patterns 3 and 4 illustrates the importance of this discussion to clinical decisions and the prognosis of prostate cancer patients.*

Resumo. *Este estudo investiga o impacto do balanceamento de classes e da regularização na melhoria da concordância de diagnósticos baseados em imagens histológicas. Modelos U-Net aplicados na análise de biópsias de câncer de próstata, por exemplo, mostram que o balanceamento de classes, combinado com funções de perda tradicionais, aumenta a concordância das imagens em até 6 pontos percentuais. A combinação de balanceamento com Focal Loss eleva o acordo de classificação em média 13 pontos percentuais, em comparação com o uso de datasets desbalanceados com funções de perda tradicionais. Um estudo de caso sobre a diferenciação entre os padrões Gleason 3 e 4 ilustra a utilidade das técnicas em decisões clínicas e no prognóstico de pacientes com câncer de próstata.*

1. Introdução

Uma das aplicações mais relevantes da inteligência artificial, em especial do aprendizado profundo (AP), está relacionada ao auxílio para o diagnóstico médico através de imagens [Raciti et al., 2020]. Arquiteturas de redes neurais convolucionais (CNN) como a U-Net, por exemplo, podem ser aplicadas na segmentação de imagens histológicas obtidas de biópsias para o diagnóstico de diversos tipos de câncer.

O adenocarcinoma prostático (AcP) é o câncer mais comum entre os homens, representando 10,2% dos diagnósticos de câncer masculino no Brasil, com 72.000 novos casos projetados para 2023–2025 [INCA, 2023]. O diagnóstico é baseado na biópsia prostática e na escala de padrões de Gleason (PG) [Gleason e Mellinger, 1974], que avalia a diferenciação das células tumorais em uma escala de 1 a 5. No entanto, distinguir entre os padrões 3 e 4 é desafiador devido a diferenças morfológicas sutis, levando a discrepâncias diagnósticas de 30–53%, o que impacta as decisões de tratamento [Ozkan et al., 2016].

Este artigo investiga, por meio de um estudo ablativo [Meyes et al., 2019], o impacto do balanceamento de classes e da regularização em modelos de AP aplicados à segmentação semântica de imagens histológicas. Como estudo de caso, são utilizadas imagens de biópsias da próstata, com o objetivo de aumentar a precisão e a concordância entre os PGs 3 e 4, melhorando assim as chances de cura e eficácia do tratamento. Todo o trabalho foi realizado exclusivamente pelo primeiro autor, sob orientação do segundo.

De acordo com Bulten et al. [2022], as redes neurais convolucionais (CNNs) superam patologistas em precisão, sensibilidade e especificidade, na análise de biópsias de câncer de próstata. No estudo de Silva-Rodríguez et al. [2020], foi alcançado um índice de Kappa Ponderado Quadrático (KPQ) de 77% no diagnóstico de AcP utilizando o conjunto de dados SICAPv2. Por sua vez, Ikromjanov et al. [2022] obtiveram F1-scores de 78% para PG3 e 67% para PG4 no conjunto PANDA, utilizando *patches* de 256×256 pixels sem técnicas adicionais de pré-processamento, sugerindo potencial para melhorias com abordagens mais avançadas.

A pesquisa de Guerrero et al. [2024] investigou métodos de aumento de dados para mitigar o desbalanceamento em conjuntos histopatológicos, explorando técnicas tanto no nível do classificador quanto dos dados. Em um estudo complementar, Falahkheirkhah et al. [2023] aplicaram redes adversárias generativas para sintetizar imagens histológicas realistas, contribuindo para análises médicas e para o enriquecimento da diversidade de dados disponíveis. Além disso, Hancer et al. [2023] enfrentaram o problema do desbalanceamento de classes ao empregarem o modelo U-Net na segmentação de núcleos em imagens histopatológicas. De forma semelhante, os estudos de Haghofer et al. [2023] e Chen [2023] destacaram o desempenho superior do U-Net em segmentação de imagens médicas, abrangendo células e núcleos, evidenciando sua eficácia na análise histológica.

Este estudo se relaciona diretamente com os trabalhos de Guerrero et al. [2024], que utilizaram *Mask R-CNN* com aumento de dados baseado na técnica “*copy-paste*”, e de Chen [2023], que aplicaram U-Net em imagens prostáticas. No entanto, diferencia-se ao adotar uma metodologia ablativa para avaliar o impacto do balanceamento de classes e da regularização, proporcionando uma compreensão mais aprofundada desses fatores na segmentação de tecidos e na classificação dos PGs 3 e 4.

2. Referencial Teórico

A arquitetura U-Net, proposta por Ronneberger et al. [2015], é amplamente reconhecida por sua eficiência e robustez em tarefas de segmentação semântica, especialmente no contexto biomédico. Com uma estrutura em forma de “U”, permite que o modelo realize segmentações precisas, mesmo quando há uma quantidade limitada de dados, sendo, portanto, altamente eficaz em aplicações tanto médicas quanto em visão computacional.

A função de perda desempenha um papel crucial na otimização de modelos de segmentação semântica, pois é responsável por medir a discrepância entre a saída do modelo e os rótulos reais. A função de perda mais comumente utilizada é uma combinação de entropia cruzada (CEL), que avalia a semelhança entre a máscara segmentada prevista e a verdadeira, com termos adicionais de regularização que visam prevenir o *overfitting* e melhorar a generalização do modelo. O *Focal Loss* (FL), introduzido por Lin et al. [2018], tem sido explorado como uma alternativa à CEL, especialmente em cenários com desequilíbrio de classes. O FL modifica a CEL ao incorporar um termo modulador, $(1 - \hat{y}_i)^\gamma$,

onde o parâmetro $\gamma > 0$ reduz o impacto das instâncias facilmente classificadas, permitindo ao modelo focar mais em exemplos difíceis. Além disso, um fator de balanceamento opcional, α_i , pode ser utilizado para ajustar ainda mais o impacto das classes desbalanceadas.

O KPQ é uma métrica estatística que avalia a concordância entre avaliadores, ponderando discrepâncias mais graves com um fator quadrático. Isso o torna útil para métodos diagnósticos com variáveis ordinais, destacando desacordos importantes e refletindo melhor a concordância geral, especialmente quando a diferença entre categorias é crucial para o diagnóstico.

3. Materiais e Métodos

O conjunto PANDA contém 10.616 imagens de alta resolução coradas com Hematoxilina e Eosina (H&E), adquiridas entre 2012 e 2017 com microscopia óptica. As imagens foram rotuladas em duas categorias: tecido benigno e canceroso, e estroma e PG2 a PG5.

Para analisar o impacto do balanceamento de dados e regularizações no desempenho dos modelos, foram treinados 24 modelos U-Net utilizando dados balanceados e desbalanceados, com diferentes abordagens de normalização de pixels e funções de perda (Figura 1) e validação cruzada com 10 dobras.

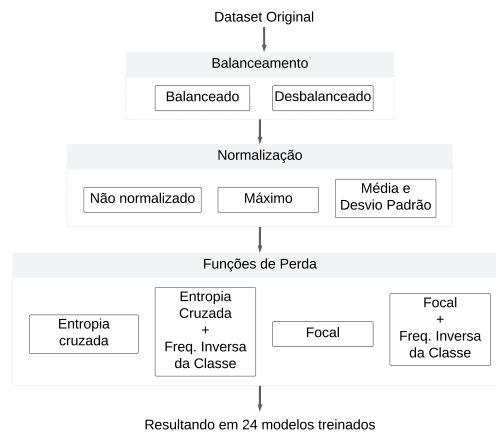


Figura 1. O esquema ablativo proposto neste estudo compreende 24 modelos distintos, cada um resultante da combinação de três etapas diferentes: Balanceamento, Normalização e Aplicação de Função de Perda.

Foram selecionadas 5.160 imagens com anotações detalhadas das glândulas prostáticas, distribuídas em 330 lâminas digitais para treinamento e 80 para teste, com amostragem estratificada. *Patches* de imagem de 224×224 pixels com 10% de sobreposição foram gerados, descartando-se o canal alfa e os canais azul e verde das máscaras. No total, 9.442 *patches* foram gerados para treinamento e 2.174 para teste. *Patches* com mais de 10% de fundo foram excluídos.

O balanceamento incluiu a seleção de imagens com PG3 e PG4, remoção de *patches* com mais de 80% de estroma e aumento artificial de *patches* das classes minoritárias. Após o balanceamento, o conjunto de treinamento final contou com 6.700 *patches*, com diferença de classe reduzida para menos de 30%. As variações de CEL e FL foram ajus-

tadas com pesos baseados na frequência inversa das classes para mitigar o viés, e todos os modelos seguiram a arquitetura U-Net proposta por Ronneberger et al. [2015].

4. Resultados e Discussão

O FL apresenta maior estabilidade na validação cruzada em comparação ao CEL, com redução mais consistente da perda, destacando sua eficácia na regularização e no manejo de dados desequilibrados, conforme mostram as Figuras 2a e 2b.

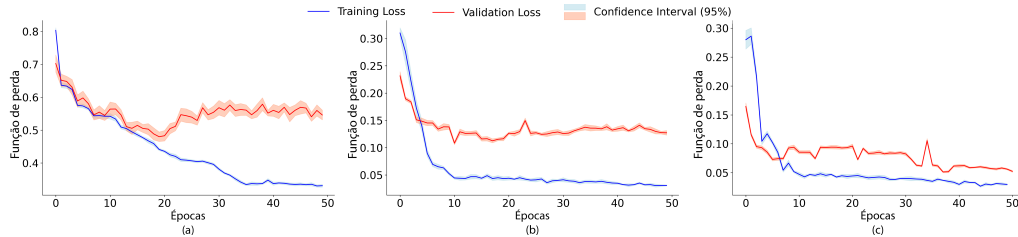


Figura 2. Funções de perda resultantes da validação cruzada para um conjunto de dados desbalanceado usando CEL (a), para o mesmo conjunto com FL (b) e para um conjunto balanceado utilizando FL (c)

A combinação de FL com balanceamento de dados (Figura 2c) reduz a diferença entre as perdas de treinamento e validação, melhorando a generalização do modelo e sua precisão em novos dados, além de diminuir o risco de *overfitting*.

Tabela 1. Validação cruzada e seus respectivos intervalos de confiança de 95% para treinamento em dataset desbalanceado.

Normalização	Função	Padrão Gleason 3			Padrão Gleason 4		
		Sensibilidade	Especificidade	F1-Score	Sensibilidade	Especificidade	F1-Score
Não normalizado	CEL	0,66 ± 0,02	0,97 ± 0,03	0,61 ± 0,02	0,37 ± 0,02	0,95 ± 0,01	0,46 ± 0,03
	CEL+FIC	0,58 ± 0,06	0,82 ± 0,08	0,57 ± 0,04	0,30 ± 0,07	0,88 ± 0,02	0,35 ± 0,07
	FL	0,72 ± 0,02	0,95 ± 0,01	0,66 ± 0,03	0,40 ± 0,01	0,97 ± 0,01	0,52 ± 0,02
	FL+ICF	0,64 ± 0,04	0,90 ± 0,04	0,59 ± 0,03	0,34 ± 0,03	0,93 ± 0,03	0,40 ± 0,04
Máximo	CEL	0,73 ± 0,01	0,95 ± 0,02	0,63 ± 0,02	0,39 ± 0,02	0,96 ± 0,03	0,46 ± 0,01
	CEL+FIC	0,57 ± 0,04	0,94 ± 0,03	0,52 ± 0,05	0,37 ± 0,02	0,91 ± 0,08	0,39 ± 0,02
	FL	0,76 ± 0,03	0,96 ± 0,01	0,68 ± 0,01	0,41 ± 0,03	0,95 ± 0,02	0,50 ± 0,01
	FL+ICF	0,65 ± 0,04	0,90 ± 0,02	0,58 ± 0,03	0,37 ± 0,03	0,88 ± 0,03	0,42 ± 0,03
Média/Desv. Padrão	CEL	0,69 ± 0,02	0,95 ± 0,03	0,64 ± 0,02	0,34 ± 0,03	0,97 ± 0,02	0,47 ± 0,02
	CEL+FIC	0,59 ± 0,04	0,90 ± 0,05	0,57 ± 0,02	0,40 ± 0,05	0,88 ± 0,09	0,40 ± 0,03
	FL	0,78 ± 0,02	0,95 ± 0,03	0,69 ± 0,01	0,43 ± 0,01	0,98 ± 0,02	0,53 ± 0,01
	FL+ICF	0,63 ± 0,03	0,87 ± 0,02	0,59 ± 0,01	0,35 ± 0,03	0,91 ± 0,01	0,41 ± 0,02

A Tabela 1 mostra que o Focal Loss melhora o equilíbrio entre sensibilidade e especificidade na classificação de PGs 3 e 4. A Figura 2b destaca sua estabilidade e leve vantagem no F1-score, embora a sobreposição dos intervalos de confiança impeça conclusões sobre a melhor normalização.

A Tabela 2 mostra que o balanceamento de dados com Focal Loss melhora o F1-score, com ganhos médios de 8 pontos percentuais para PG3 e 14 pontos para PG4. No entanto, a melhor normalização não é conclusiva devido à sobreposição dos intervalos de confiança.

A análise por KPQ mostra que o FL supera a CEL, com ganhos médios de 7 pontos em conjuntos desbalanceados e 6 em balanceados, totalizando 13 pontos ao se comparar CEL em datasets desbalanceados com FL em balanceados, o que está em concordância com o apresentado por Silva-Rodríguez et al. [2020].

Tabela 2. Validação cruzada e seus respectivos intervalos de confiança de 95% para treinamento em dataset balanceado

Normalização	Função	Padrão Gleason 3			Padrão Gleason 4		
		Sensibilidade	Especificidade	F1-Score	Sensibilidade	Especificidade	F1-Score
Não normalizado	CEL	0,77 ± 0,04	0,94 ± 0,02	0,71 ± 0,04	0,81 ± 0,02	0,95 ± 0,02	0,61 ± 0,04
	CEL+FIC	0,65 ± 0,06	0,84 ± 0,3	0,60 ± 0,03	0,60 ± 0,05	0,80 ± 0,04	0,48 ± 0,09
	FL	0,80 ± 0,01	0,94 ± 0,04	0,73 ± 0,02	0,80 ± 0,03	0,95 ± 0,04	0,66 ± 0,02
	FL+FIC	0,72 ± 0,03	0,96 ± 0,02	0,66 ± 0,03	0,80 ± 0,02	0,90 ± 0,03	0,51 ± 0,06
Maximum	CEL	0,76 ± 0,01	0,95 ± 0,03	0,66 ± 0,03	0,82 ± 0,02	0,92 ± 0,01	0,60 ± 0,01
	CEL+FIC	0,69 ± 0,03	0,85 ± 0,02	0,61 ± 0,05	0,60 ± 0,07	0,85 ± 0,02	0,58 ± 0,04
	FL	0,78 ± 0,03	0,96 ± 0,02	0,75 ± 0,03	0,82 ± 0,03	0,96 ± 0,02	0,66 ± 0,03
	FL+ICF	0,73 ± 0,02	0,93 ± 0,04	0,66 ± 0,04	0,73 ± 0,03	0,97 ± 0,01	0,60 ± 0,03
Média/Desv. Padrão	CEL	0,78 ± 0,02	0,91 ± 0,02	0,73 ± 0,02	0,81 ± 0,04	0,95 ± 0,02	0,59 ± 0,02
	CEL+FIC	0,70 ± 0,03	0,86 ± 0,01	0,59 ± 0,04	0,61 ± 0,02	0,84 ± 0,04	0,54 ± 0,05
	FL	0,85 ± 0,02	0,97 ± 0,02	0,77 ± 0,01	0,81 ± 0,01	0,97 ± 0,01	0,65 ± 0,02
	FL+ICF	0,75 ± 0,04	0,92 ± 0,03	0,66 ± 0,02	0,77 ± 0,01	0,96 ± 0,01	0,59 ± 0,04

Trabalhos vencedores no Kaggle alcançaram cerca de 90% de concordância removendo rótulos com discrepâncias, mas descartando casos desafiadores, como a distinção entre PG3 e PG4. Isso influencia os altos índices obtidos, mantendo os resultados deste estudo competitivos e destacando a importância do balanceamento e da regularização.

Tabela 3. Validação cruzada e seus respectivos intervalos de confiança de 95% para a métrica KPQ

		CEL	CEL+FIC	FL	FL+FIC
Desbalanceado	Não normalizado	0,57 ± 0,05	0,20 ± 0,14	0,65 ± 0,03	0,34 ± 0,04
	Máximo	0,61 ± 0,03	0,23 ± 0,06	0,67 ± 0,02	0,40 ± 0,02
	Média/Desv. Padrão	0,55 ± 0,02	0,21 ± 0,03	0,64 ± 0,02	0,51 ± 0,03
Balanceado	Não normalizado	0,66 ± 0,02	0,25 ± 0,09	0,64 ± 0,06	0,60 ± 0,05
	Máximo	0,65 ± 0,01	0,27 ± 0,07	0,70 ± 0,02	0,66 ± 0,02
	Média/Desv. Padrão	0,62 ± 0,03	0,28 ± 0,02	0,73 ± 0,01	0,64 ± 0,02

A normalização de pixels não melhorou a concordância, evidenciando que seu impacto é contextual (Tabela 3). Os pesos da FIC dificultaram a convergência, e o uso da frequência inversa das classes, combinado às funções de perda, piorou a classificação de PGs 3 e 4, agravando o desequilíbrio e prejudicando a generalização.

5. Conclusão

Este estudo destaca a importância do balanceamento de imagens para a precisão no diagnóstico através de imagens histológicas de câncer de próstata e como estratégia de regularização no treinamento de modelos. Aconselha-se o uso cauteloso de pesos em funções de perda, pois sua má aplicação pode desestabilizar o modelo. Resultados competitivos foram alcançados com pré-processamento mínimo, evidenciando o papel do balanceamento e da regularização.

Para trabalhos futuros, recomenda-se reduzir ruídos nas anotações e abordar distorções histológicas para melhorar as previsões. O uso de *ensembles* pode aprimorar a classificação entre PGs 3 e 4. Além disso, a exploração de novas arquiteturas é essencial para avançar na análise do adenocarcinoma prostático. Finalmente, a extensão da análise para diferentes bases de dados e outros tipos de câncer irá corroborar as vantagens do balanceamento e da regularização em modelos de aprendizado profundo.

Agradecimentos — EHS Paraíso agradece à Pontifícia Universidade Católica de Minas Gerais – PUC-Minas e à Coordenação de Aperfeiçoamento de Pessoal de

Nível Superior - CAPES - PUC/MG, Bolsa PDPG 88887.708960/2022-00 - PUC/MG - INFORMÁTICA e Código Financeiro 001). Conselho Nacional de Desenvolvimento Científico e Tecnológico do Brasil (CNPq – Código:311573/2022-3). AMC Machado agradece à FAPEMIG pelo auxílio financeiro através dos projetos APQ-02753-24 e APQ-06556-24.

Referências

- Bulten, W., Kartasalo, K., Chen, P. C., et al. (2022). Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nat Med*, 28(1):154–163.
- Chen, Z. (2023). Medical Image Segmentation Based on U-Net. *J. Phys.: Conf. Ser.*, 2547(1):012010.
- Falahkheirkhah, K., Tiwari, S., Yeh, K., et al. (2023). Deepfake Histologic Images for Enhancing Digital Pathology. *Laboratory Investigation*, 103(1):100006.
- Gleason, D. F. e Mellinger, G. T. (1974). Prediction of Prognosis for Prostatic Adenocarcinoma by Combined Histological Grading and Clinical Staging. *Journal of Urology*, 111(1):58–64.
- Guerrero, E. D., Lina, R., Lina, R., et al. (2024). A Data Augmentation Methodology to Reduce the Class Imbalance in Histopathology Images. *J Digit Imaging. Inform. med.*
- Haghofer, A., Fuchs-Baumgartinger, A., Lipnik, K., et al. (2023). Histological classification of canine and feline lymphoma using a modular approach based on deep learning and advanced image processing. *Sci Rep*, 13:19436.
- Hancer, E., Traoré, M., Samet, R., et al. (2023). An imbalance-aware nuclei segmentation methodology for H&E stained histopathology images. *Biomedical Signal Processing and Control*, 83:104720.
- Ikromjanov, K., Bhattacharjee, S., Hwang, Y.-B., et al. (2022). Whole Slide Image Analysis and Detection of Prostate Cancer using Vision Transformers. In *2022 ICAIIC*, p. 399–402, Jeju Island, Korea, Republic of.
- INCA, I. N. D. C. (2023). *Estimativa 2023: incidência de câncer no Brasil*. Instituto Nacional De Câncer, Rio de Janeiro, RJ.
- Lin, T.-Y., Goyal, P., Girshick, R., et al. (2018). Focal Loss for Dense Object Detection.
- Meyes, R., Lu, M., de Puiseau, C. W., et al. (2019). Ablation Studies in Artificial Neural Networks.
- Ozkan, T. A., Eruyar, A., Cebeci, O., et al. (2016). Interobserver variability in Gleason histological grading of prostate cancer. *Scandinavian Journal of Urology*, 50(6):420–424.
- Raciti, P., Sue, J., e Rodrigo, C. (2020). Novel artificial intelligence system increases the detection of prostate cancer in whole slide images of core needle biopsies. *Modern Pathology*, 33(10):2058–2066.
- Ronneberger, O., Fischer, P., e Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI 2015*, volume 9351, p. 234–241.
- Silva-Rodríguez, J., Colomer, A., Sales, M. A., et al. (2020). Going deeper through the Gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Computer Methods and Programs in Biomedicine*, 195:105637.