Revista Eletrônica de Iniciação Científica em Computação, 2025, 23:1
ISSN: 3085-8461 • doi: 10.5753/reic.2025.6197

Artigo de Pesquisa/Research Paper

Impacto do balanceamento e regularização na segmentação semântica de imagens histopatológicas

Impact of Balancing and Regularization on the Semantic Segmentation of Histopathological Images

Eduardo Paraíso [Pontifícia Universidade Católica de Minas Gerais | *eduardo.paraiso@sga.pucminas.br*]
Alexei Machado [Pontifícia Universidade Católica de Minas Gerais, Universidade Federal de Minas Gerais |*alexeimcmachado@gmail.com*]

Programa de Pós-graduação em Informática, Pontifícia Universidade Católica de Minas Gerais, Av. Dom José Gaspar, 500, Coração Eucarístico, Belo Horizonte, MG, 30535-901, Brasil.

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.

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.

Palavras-chave: Segmentação de imagens, Aprendizado Profundo, Balanceamento de Classes, Regularização, Câncer de Próstata

Keywords: Image Segmentation, Deep Learning, Class Balancing, Regularization, Prostate Cancer

Recebido/Received: 01 July 2025 • **Aceito/Accepted:** 03 November 2025 • **Publicado/Published:** 28 November 2025

1 Introdução

O adenocarcinoma de próstata (AcP) é o tipo de câncer mais comum entre os homens em todo o mundo, representando 10,2% dos diagnósticos de câncer masculino no Brasil, com uma projeção de 72 mil novos casos para o período de 2023 a 2025 [INCA, 2023]. O diagnóstico padrão-ouro é baseado na biópsia da próstata e no sistema de escores de Gleason [Gleason and Mellinger, 1974], que avalia a diferenciação celular do tumor em uma escala de 1 a 5 (Figura 1), sendo os padrões (PG) 3 e 4 indicativos de malignidade moderada e alta, respectivamente. As sutis diferenças morfológicas entre estes padrões torna sua distinção desafiadora, resultando em discrepâncias que variam de 30% a 53% [Ozkan *et al*., 2016]. Essas imprecisões impactam diretamente o manejo clínico do paciente, como a indicação de prostatectomia, procedimento que pode acarretar efeitos colaterais severos. Portanto, uma diferenciação precisa entre os padrões é essencial para evitar tratamentos excessivamente agressivos e melhorar os desfechos clínicos dos pacientes.

Uma das aplicações mais relevantes da inteligência artificial, em especial do aprendizado profundo (DL), 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 proposta por Ronneberger *et al*. [2015], por exemplo, podem ser aplicadas na segmentação de imagens histológicas obtidas de biópsias para o diagnóstico de diversos tipos de câncer. No entanto, a distinção entre os PGs 3 e 4 ainda representa um desafio, devido à baixa concordância entre os resultados, o que limita o uso clínico dessas abordagens de DL. Além disso, o desbalanceamento de classes nos dados de treinamento introduz vieses, comprometendo a eficácia dos modelos na detecção de classes minoritárias [Dablain *et al*., 2024].

Este artigo investiga, por meio de um estudo ablativo [Meyes *et al*., 2019], o impacto do balanceamento de classes em modelos de DL aplicados à segmentação semântica de imagens histológicas da próstata, bem como o efeito da regularização na prevenção de overfitting ao se empregar uma função de perda projetada para lidar com o desbalanceamento em problemas de classificação. O estudo é direcionado especificamente à avaliação de modelos baseados na arquitetura U-Net. Nosso objetivo final é aumentar a acurácia e as métricas de concordância nos PGs 3 e 4, a fim de ampliar as chances de cura e a eficácia do tratamento dos pacientes.

A estrutura deste trabalho está organizada da seguinte forma: a Seção 2 revisa estudos sobre segmentação semântica em conjuntos de dados de imagens da próstata, destacando o uso e a importância do balanceamento nesses conjuntos. A Seção 3 explora os conceitos fundamentais necessários para uma compreensão mais aprofundada deste estudo. A Seção

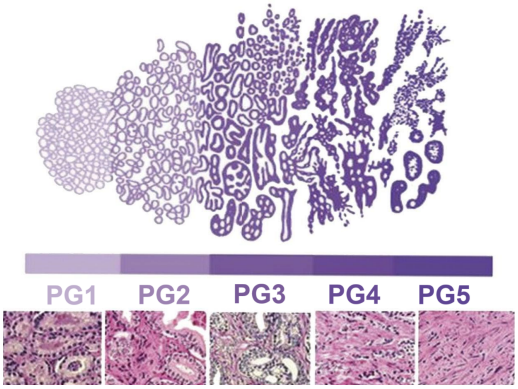

Figura 1. Escala dos Padrões de Gleason: PG1: Células regulares, uniformes e pequenas. PG2: Células uniformes, agrupadas de forma frouxa, com bordas irregulares. PG3: Células muito pequenas, uniformes, de formato angular ou alongado. PG4: Muitas células fundidas formando grandes massas amorfas. PG5: Grandes massas celulares com invasão de órgãos e tecidos vizinhos, apresentando diferenciação glandular mínima. Adaptado de: University of Pittsburgh Medical Center (UPMC) Cancer Centers, Pittsburgh, EUA.

4 apresenta a metodologia adotada. A Seção 5 discute os resultados obtidos durante os experimentos. Por fim, a Seção 6 apresenta as considerações finais e propõe direções para trabalhos futuros.

2 Trabalhos Relacionados

De acordo com Bulten *et al*. [2022], as CNNs são capazes de superar 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 de melhorias com abordagens mais modernas.

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.

3 Referencial Teórico

Selecionar métricas adequadas é fundamental para avaliar com precisão o desempenho do modelo no contexto específico. As funções de perda desempenham papel-chave ao orientar o treinamento, permitindo que o modelo distinga os padrões de forma eficaz. Isso assegura uma segmentação precisa e clinicamente significativa dos PGs, conduzindo a diagnósticos mais acurados e tratamentos mais adequados.

3.1 Funções de Perda

A função de perda é crucial para otimizar modelos de segmentação semântica, garantindo que a saída da rede seja adequadamente comparada aos rótulos reais. A função de perda mais comum combina a cross-entropy (CEL), que avalia a similaridade entre a máscara segmentada prevista e a máscara real, com termos de regularização para evitar overfitting. A função CEL é definida como:

$$CEL = -\sum_{i=1}^{N} y_i \log(\hat{y}_i), \tag{1}$$

onde N é o número total de classes, y_i representa um vetor com a classe verdadeira e $\hat{y}_i$ representa a probabilidade da classe prevista. Essa função incentiva o aprendizado preciso das características discriminativas de cada classe [Rączkowska *et al*., 2019].

A Focal Loss (FL), proposta por Lin *et al*. [2018], foi explorada como alternativa à cross-entropy para lidar com o desbalanceamento de classes em problemas de classificação. Ela adiciona um termo modulador, $(1 - \hat{y}_i)^\gamma$, à CEL, onde $\gamma > 0$ reduz a perda para exemplos bem classificados. Um fator de balanceamento opcional, α_i, também pode ser usado para tratar desequilíbrios entre classes:

$$FL = -\alpha_i (1 - \hat{y}_i)^\gamma \log \hat{y}_i. \tag{2}$$

Essa abordagem é especialmente útil em conjuntos de dados com classes minoritárias, melhorando o desempenho da rede, conforme demonstrado por Nguyen *et al*. [2024].

3.2 Métricas

O desempenho dos modelos foi avaliado com base em um conjunto de métricas conhecidas:

1. **Sensibilidade** (recall): indica a proporção de verdadeiros positivos em relação ao total de casos positivos, considerando também os falsos negativos [Powers, 2015].
2. **Especificidade**: representa a proporção de observações verdadeiramente negativas no conjunto de dados [Monaghan *et al*., 2021], refletindo a capacidade do modelo de evitar falsos positivos.
3. **F1-Score**: métrica fundamental para avaliação de modelos de classificação, especialmente em cenários de desbalanceamento de classes. É a média harmônica entre precisão e recall, sendo útil quando é necessário equilibrar esses dois aspectos, sobretudo quando um tipo de erro (falsos positivos ou falsos negativos) tem maior impacto [Hicks *et al*., 2022].
4. **Kappa Ponderado Quadrático (KPQ)**: medida estatística que avalia o acordo entre avaliadores considerando

pesos para discrepâncias, de acordo com a distância entre categorias. A diferença entre classes é ponderada por um fator quadrático, sendo o peso para a célula na linha i e coluna j da matriz dado por

$$P(i,j) = \frac{(i-j)^2}{(N-1)^2}, \tag{3}$$

onde N é o número total de categorias.
O KPQ é calculado comparando a matriz de confusão ponderada observada com a matriz de expectativa ponderada:

$$KPQ = 1 - \frac{\sum P(i,j)O(i,j)}{\sum P(i,j)E(i,j)}, \tag{4}$$

onde $O(i,j)$ é a frequência observada de concordância na categoria e $E(i,j)$ é a frequência esperada de concordância.
A ponderação quadrática atribui pesos maiores para discrepâncias mais distantes na escala ordinal. Aplicando esses pesos, o KPQ dá maior importância a desacordos graves, resultando em valores menores que o Kappa simples. Assim, o KPQ é útil para avaliar a reprodutibilidade de métodos diagnósticos com variáveis ordinais [Silva *et al.*, 2016]. Contudo, o KPQ avalia o acordo geral entre as classificações, fornecendo uma visão agregada do nível geral de concordância entre todas as classes.

4 Materiais e Métodos

O conjunto de dados PANDA foi desenvolvido em conjunto pelo grupo de patologia computacional do Departamento de Epidemiologia Médica e Bioestatística do Instituto Karolinska (DEMBIK) e pelo Centro Médico da Universidade Radboud (RUMC)[Bulten *et al.*, 2022]. O conjunto é composto por biópsias por agulha grossa realizadas entre os anos de 2012 e 2017. Devido à natureza subjetiva dos PGs, ocorrem divergências classificatórias, como apontado por Corte [2023], que destaca a presença de ruído significativo nos rótulos das imagens, decorrente de registros inconclusivos, erros de anotação, imprecisões e discrepâncias entre patologistas.

O conjunto é composto por 10.616 imagens de alta resolução coradas com hematoxilina e eosina (H&E), armazenadas no formato TIFF (*Tagged Image File Format*). Essas imagens foram obtidas por meio de microscopia óptica, onde foram utilizadas para este trabalho a ampliação da lente objetiva de 20x. Uma característica essencial das *Whole Slide Images* (WSIs) é sua capacidade de fornecer múltiplos níveis de ampliação (ver Figura 2a), nos quais a imagem original é subdividida em diversas resoluções.

As amostras fornecidas pelo DEMBIK foram rotuladas por regiões (ver Figura 2b) como fundo, tecido benigno e canceroso. Em contraste, o RUMC realizou uma classificação mais detalhada (ver Figura 2b), rotulando individualmente a citoarquitetura como fundo, estroma, PG2, PG3, PG4 e PG5.

Com o objetivo de investigar o impacto do balanceamento e da regularização por meio de uma abordagem ablativa, os modelos U-Net foram treinados utilizando combinações de conjuntos de imagens desbalanceados e balanceados, juntamente com diferentes métodos de normalização de pixels e funções de perda. Isso resultou em um total de 24 modelos

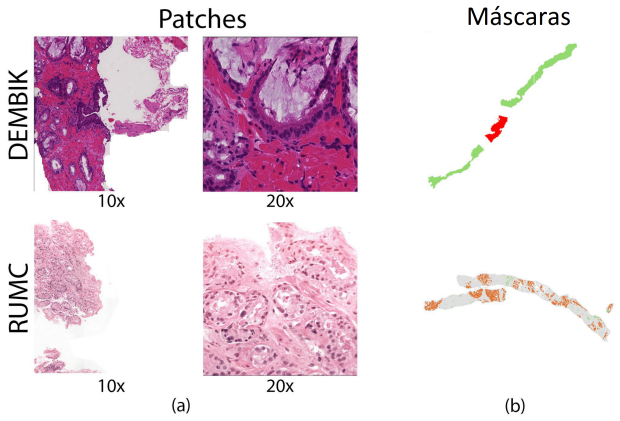

Figura 2. (a) Múltiplos níveis de ampliação fornecidos pela estrutura piramidal de uma WSI. (b) Máscara completa de segmentação de uma WSI.

treinados. Todos os modelos foram avaliados utilizando validação cruzada com 10 folds, em uma GPU T4 com 28 GB de RAM, sendo calculado um intervalo de confiança de 95%. A abordagem ablativa é útil para compreender de forma mais profunda como diferentes componentes do processo de treinamento influenciam o desempenho final do modelo. Neste caso, a ablação do balanceamento de dados permite avaliar como o desbalanceamento entre as classes afeta as métricas de segmentação, especialmente entre os padrões PG3 e PG4. A escolha da arquitetura U-Net fundamenta-se em seu desempenho consolidado na segmentação de imagens médicas, na capacidade de capturar informações contextuais em múltiplas escalas e na facilidade de implementação associada a menores demandas computacionais, aspecto particularmente relevante diante das restrições de hardware deste estudo.

4.1 Seleção e Pré-processamento das Imagens

Foram selecionadas 5.160 imagens provenientes do RUMC, devido às anotações individualizadas das glândulas (Figura 3a). Para o treinamento, foram escolhidas 330 WSIs, e outras 80 WSIs para teste, por meio de amostragem aleatória estratificada (Figura 3b). A estratificação foi baseada no Escore de Gleason, sistema de classificação histológica composto por dois escores numéricos entre 1 e 5, que representam os dois padrões tumorais predominantes no tecido. A adoção da abordagem ablativa envolve uma ampla gama de combinações; por isso, restrições de tempo e limitações de hardware justificaram o desenho metodológico desta investigação.

4.2 Geração de Patches

O uso de patches é fundamental para o treinamento de modelos em imagens histológicas, pois possibilita a diversificação e a captura de detalhes localizados, aprimorando a capacidade do modelo de reconhecer características complexas e sutilezas [Dablain *et al.*, 2024].

Durante a geração dos patches, o canal alfa foi excluído, uma vez que a transparência é irrelevante para tarefas de segmentação [Alsayat *et al.*, 2023], enquanto os canais azul e verde, provenientes das máscaras, foram omitidos, pois os dados de classificação de pixels estão armazenados apenas no canal vermelho. Os patches foram criados (Figura 3c) com dimensões de 224×224×3 para as imagens e 224×224×1 para as

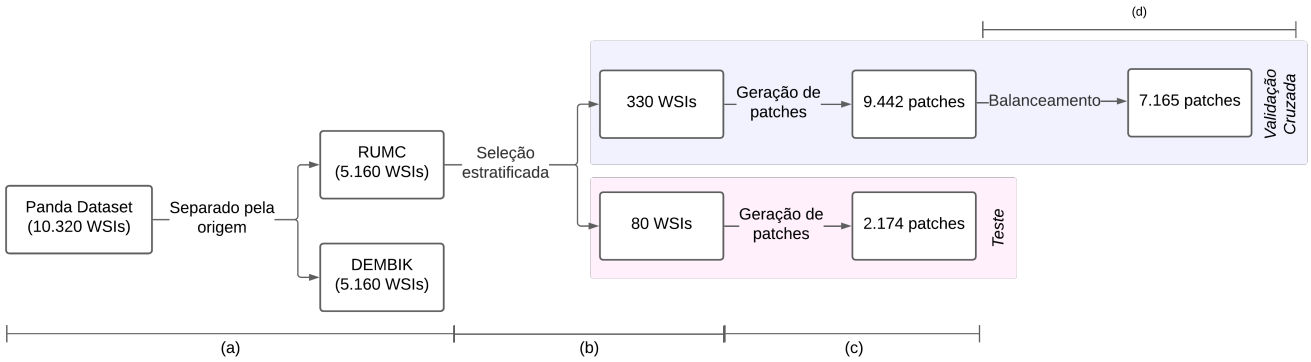

Figura 3. (a) Separação do conjunto de imagens de acordo com sua origem. (b) Os conjuntos de treinamento e teste são criados por meio de seleção aleatória estratificada a partir do conjunto de dados do RUMC. (c) Conjunto selecionado de patches com pelo menos 90% de área relevante para a classificação. (d) O conjunto final de patches resultou do balanceamento entre classes.

máscaras. Esse tamanho equilibra a eficiência computacional com a capacidade do DL em lidar com dados de alta dimensionalidade [Ciga *et al*., 2021] e assegura compatibilidade com arquiteturas amplamente utilizadas, como as treinadas no ImageNet [Russakovsky and et al., 2015].

Foi implementada uma abordagem de janela deslizante (*sliding window*) com sobreposição de 10% do tamanho do patch, permitindo a geração de patches com limites sobrepostos. Entre os patches gerados, a seleção das imagens mais representativas baseou-se na minimização dos pixels rotulados como fundo. Patches com uma proporção de fundo superior a 10% da área total da imagem foram excluídos do conjunto de dados, garantindo maior concentração de pixels relevantes para a análise histopatológica.

Após o processamento descrito, foram obtidos 9.442 patches para o conjunto de treinamento e 2.174 patches para o conjunto de teste (Figura 3c).

Devido à natureza das imagens, foi calculada a proporção de cada rótulo, constatando-se uma diferença de 68% entre a classe majoritária (estroma) e a classe minoritária (PG3). O balanceamento foi realizado em quatro etapas:

1. Seleção de imagens contendo PG3 ou PG4;
2. Remoção de patches com composição de pixels classificados como estroma superior a 80%;
3. Seleção de patches compostos por mais de 50% de PG3 ou PG4 para aumento artificial;
4. Aumento artificial na proporção de quatro novas imagens para cada patch original de PG3 e uma nova imagem para cada patch original de PG4.

As seguintes transformações foram utilizadas para o data augmentation:

a. Ajustes aleatórios de contraste e brilho;
b. Rotação limitada a 35º;
c. Reflexão horizontal e/ou vertical.

Após a etapa de balanceamento do conjunto de treinamento (ver Figura 3d), obteve-se um conjunto final de 7.165 patches, com uma diferença de desbalanceamento entre as classes majoritária e minoritária inferior a 30% (Figura 4).

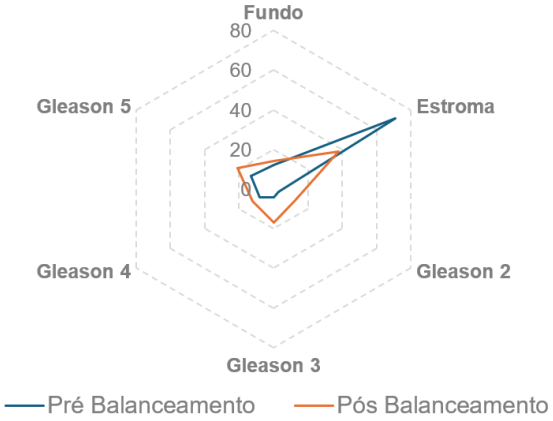

Figura 4. Distribuição das classes no conjunto de dados original (azul) e no conjunto balanceado (laranja)

4.3 Normalização

Os conjuntos de imagens balanceados e desbalanceados foram inicialmente analisados sem qualquer normalização prévia. Em seguida, foram aplicadas duas técnicas diferentes de normalização: normalização pelo valor máximo de pixel e normalização pela média e desvio padrão do conjunto de treinamento. Cada um desses conjuntos resultantes foi utilizado para treinar 24 modelos distintos, empregando diferentes funções de perda, conforme ilustrado na Figura 5.

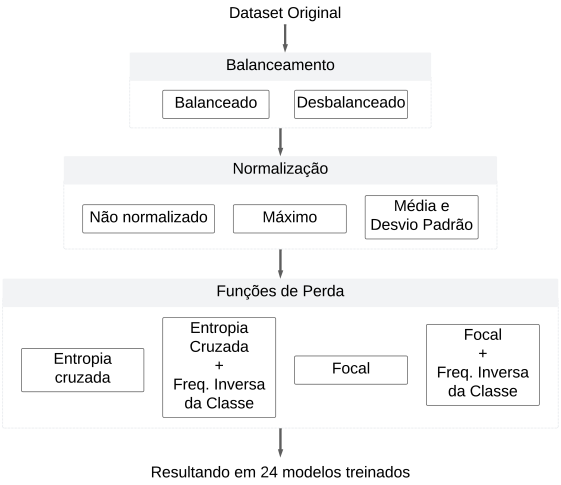

Figura 5. O esquema ablativo proposto neste estudo compreende 24 modelos distintos, cada um resultante da combinação de três etapas diferentes: Balanceamento, Normalização e Aplicação de Função de Perda.

4.4 Função de Perda

Além da Cross Entropy Loss (CEL), este estudo utilizou a Focal Loss (FL), desenvolvida para lidar com cenários de desbalanceamento extremo entre classes. Adicionalmente, foram aplicadas variações da CEL e da FL que incorporam pesos baseados na frequência inversa das classes. Esse ajuste busca minimizar possíveis vieses e facilitar um aprendizado mais equilibrado pelo modelo, promovendo melhor generalização e desempenho, especialmente para classes sub-representadas. Todos os modelos foram implementados com base na arquitetura padrão U-Net [Ronneberger *et al.*, 2015].

5 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 6a e 6b.

A combinação de FL com balanceamento de dados (Figura 7) 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*.

A Tabela 1 mostra que o Focal Loss melhora o equilíbrio entre sensibilidade e especificidade na classificação de PGs 3 e 4. A Figura 6b 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].

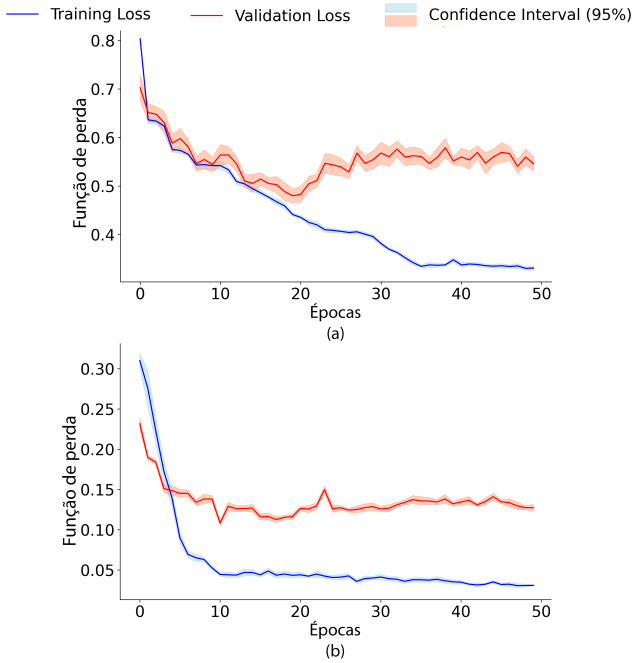

Figura 6. Funções de perda resultantes da validação cruzada para um conjunto de dados desbalanceado usando CEL (a), para o mesmo conjunto com FL (b)

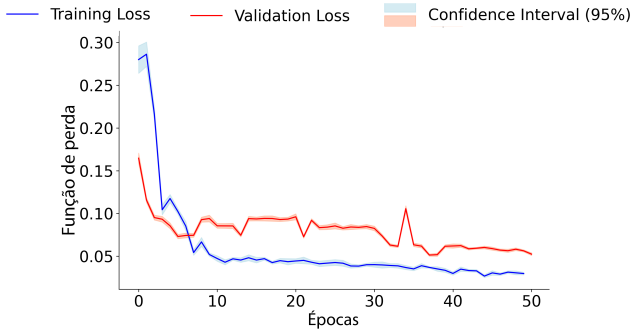

Figura 7. Funções de perda resultantes da validação cruzada para um conjunto de dados balanceado com uso simultâneo de Focal Loss (c)

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.

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.

6 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.

Declarações complementares

Agradecimentos

Eduardo Paraíso agradece à Pontifícia Universidade Católica de Minas Gerais – PUC-Minas e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico do Brasil (CNPq – Código: 132509/2024-5). Alexei Machado agradece à FAPEMIG pelo auxílio financeiro através dos projetos APQ-02753-24 e APQ-06556-24.

Contribuições dos autores

Este trabalho foi desenvolvido como parte dos requisitos para a conclusão do curso de graduação. As contribuições dos autores foram as seguintes: EP foi responsável por toda a parte prática e escrita do trabalho. AM contribuiu oferecendo orientação acadêmica e revisando o texto final.

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+FIC	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+FIC	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+FIC	0,63 ± 0,03	0,87 ± 0,02	0,59 ± 0,01	0,35 ± 0,03	0,91 ± 0,01	0,41 ± 0,02

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+FIC	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+FIC	0,75 ± 0,04	0,92 ± 0,03	0,66 ± 0,02	0,77 ± 0,01	0,96 ± 0,01	0,59 ± 0,04

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

Disponibilidade de dados e materiais

Os conjuntos de dados analisados pode ser acessados na plataforma Kaggle em `https://www.kaggle.com/competitions/prostate-cancer-grade-assessment`. Os códigos e resultados gerados durante o estudo atual estão disponíveis no repositório do trabalho podendo ser acessado em: `https://github.com/eduardoparaiso/SBCAS_25`".

Referências

Alsayat, A., Elmezain, M., Alanazi, S., Alruily, M., Mostafa, A. M., and Said, W. (2023). Multi-Layer Preprocessing and U-Net with Residual Attention Block for Retinal Blood Vessel Segmentation. *Diagnostics*, 13(21):3364. DOI: 10.3390/diagnostics13213364.

Bulten, W., Kartasalo, K., Chen, P. C., Ström, P., Pinckaers, H., and Nagpal, K. (2022). Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nat Med*, 28(1):154–163. DOI: 10.1038/s41591-021-01620-2.

Chen, Z. (2023). Medical Image Segmentation Based on U-Net. *J. Phys.: Conf. Ser.*, 2547(1):012010. DOI: 10.1088/1742-6596/2547/1/012010.

Ciga, O., Xu, T., Nofech-Mozes, S., Noy, S., Lu, F.-I., and Martel, A. L. (2021). Overcoming the limitations of patch-based learning to detect cancer in whole slide images. *Sci Rep*, 11(1):8894. DOI: 10.1038/s41598-021-88494-z.

Corte, D. D. (2023). Towards a Clinically Useful AI Tool for Prostate Cancer Detection: Recommendations from a PANDA Dataset Analysis. *JCRMHS*, 5(3). DOI: 10.55920/JCRMHS.2023.05.001216.

Dablain, D., Krawczyk, B., and Chawla, N. (2024). Towards a holistic view of bias in machine learning: bridging algorithmic fairness and imbalanced learning. *Discov Data*, 2(1):4. DOI: 10.1007/s44248-024-00007-1.

Falahkheirkhah, K., Tiwari, S., Yeh, K., Gupta, S., Herrera-Hernandez, L., McCarthy, M. R., Jimenez, R. E., Cheville, J. C., and Bhargava, R. (2023). Deepfake Histologic Images for Enhancing Digital Pathology. *Laboratory Investigation*, 103(1):100006. DOI: 10.1016/j.labinv.2022.100006.

Gleason, D. F. and Mellinger, G. T. (1974). Prediction of Prognosis for Prostatic Adenocarcinoma by Combined Histological Grading and Clinical Staging. *Journal of Urology*, 111(1):58–64. DOI: 10.1016/s0022-5347(17)59889-4.

Guerrero, E. D., Lina, R., Lina, R., Bocklitz, T., Popp, J., and Oliveira, J. L. (2024). A Data Augmentation Methodology

to Reduce the Class Imbalance in Histopathology Images. *J Digit Imaging. Inform. med.*. DOI: 10.1007/s10278-024-01018-9.

Haghofer, A., Fuchs-Baumgartinger, A., Lipnik, K., Klopfleisch, R., Aubreville, M., Scharinger, J., Weissenböck, H., Winkler, S. M., and Bertram, C. A. (2023). Histological classification of canine and feline lymphoma using a modular approach based on deep learning and advanced image processing. *Sci Rep*, 13:19436. DOI: 10.1038/s41598-023-46607-w.

Hancer, E., Traoré, M., Samet, R., Yıldırım, Z., and Nemati, N. (2023). An imbalance-aware nuclei segmentation methodology for H&E stained histopathology images. *Biomedical Signal Processing and Control*, 83:104720. DOI: 10.1016/j.bspc.2023.104720.

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., and Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, 12(1):5979.

Ikromjanov, K., Bhattacharjee, S., Hwang, Y.-B., Sumon, R. I., Kim, H.-C., and Choi, H.-K. (2022). Whole Slide Image Analysis and Detection of Prostate Cancer using Vision Transformers. In *2022 ICAIIC*, pages 399–402, Jeju Island, Korea, Republic of. DOI: 10.1109/ICAIIC54071.2022.9722635.

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., He, K., and Dollár, P. (2018). Focal Loss for Dense Object Detection.

Meyes, R., Lu, M., de Puiseau, C. W., and Meisen, T. (2019). Ablation Studies in Artificial Neural Networks.

Monaghan, T. F., Rahman, S. N., Agudelo, C. W., Wein, A. J., Lazar, J. M., Everaert, K., and Dmochowski, R. R. (2021). Foundational Statistical Principles in Medical Research: Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value. *Medicina*, 57(5):503.

Nguyen, T. T. U., Nguyen, A.-T., Kim, H., Jung, Y. J., Park, W., and Kim, Kyoung Min, e. a. (2024). Deep-learning model for evaluating histopathology of acute renal tubular injury. *Sci Rep*, 14(1):9010. DOI: 10.1038/s41598-024-58506-9.

Ozkan, T. A., Eruyar, A., Cebeci, O., Memik, O., Ozcan, L., and Kuskonmaz, I. (2016). Interobserver variability in Gleason histological grading of prostate cancer. *Scandinavian Journal of Urology*, 50(6):420–424. DOI: 10.1080/21681805.2016.1206619.

Powers, D. M. W. (2015). Evaluation Evaluation a Monte Carlo study. DOI: 10.48550/arXiv.1504.00854.

Raciti, P., Sue, J., Ceballos, R., Godrich, R., Kunz, J. D., Kapur, S., Reuter, V., Grady, L., Kanan, C., Klimstra, D. S., and Fuchs, T. J. (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. DOI: 10.1038/s41379-020-0551-y.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI 2015*, volume 9351, pages 234–241. DOI: 10.1007/978-3-319-24574-4_28.

Russakovsky, O. and et al. (2015). ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vis*, 115(3):211–252. DOI: 10.48550/arXiv.1409.0575.

Rączkowska, A., Możejko, M., Zambonelli, J., and Szczurek, E. (2019). ARA: accurate, reliable and active histopathological image classification framework with Bayesian deep learning. *Sci Rep*, 9(1):14347. DOI: 10.1038/s41598-019-50587-1.

Silva, A. F. D., Velo, M. M. D. A. C., and Pereira, A. C. (2016). Importância da reprodutibilidade dos métodos para diagnóstico em odontologia. *Rev. da Fac. de Odontologia, UPF*, 21(1). DOI: 10.5335/rfo.v21i1.4433.

Silva-Rodríguez, J., Colomer, A., Sales, M. A., Molina, R., and Naranjo, V. (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. DOI: 10.1016/j.cmpb.2020.105637.