Método Automático para classificação de Câncer de Próstata em WSIs utilizando EfficientNet, Ensemble e Modelagem de Perda Ordinal

Woshington V. S. Rodrigues[1] , Antonio O. C. Filho[2] , Armando L. Borges[2] , Jose D. Araújo[2] , João O. B. Diniz[3]

[1] Instituto Federal do Piauí (IFPI) - Picos, PI - Brasil

[2]Universidade Federal do Piauí (UFPI) - Picos, PI - Brasil

[3]Instituto Federal do Maranhão (IFMA) - São Luís, MA - Brasil

woshington@ifpi.edu.br

Abstract. *Manual histopathological evaluation of prostate cancer in WSIs is subject to interobserver and intraobserver variability. To mitigate this problem, this work proposes the automatic classification of ISUP grade groups using Convolutional Neural Networks. The methodology used the PANDA dataset, applying entropy-based noise reduction and patch extraction. The EfficientNet models (B0, B3, and B7) were trained with the Ordinal Focal Loss function to preserve disease progression and mitigate class imbalance. Finally, the individual predictions were combined into an ensemble. The simple averaging strategy achieved the best performance, with a Quadratic Weighted Kappa of 0.879 and an accuracy of 0.698. The approach preserved the ordinal integrity of the grades, rendering extreme errors nearly nonexistent and demonstrating its clinical viability.*

Resumo. *A avaliação histopatológica manual do câncer de próstata em WSIs está sujeita a variabilidades interobservador e intraobservador. Para mitigar esse problema, este trabalho propõe a classificação automática dos grupos de graus ISUP utilizando Redes Neurais Convolucionais. A metodologia utilizou o dataset PANDA , aplicando a redução de ruído por entropia e a extração de patches. Os modelos EfficientNet (B0, B3 e B7) foram treinados com a função Ordinal Focal Loss para preservar a progressão da doença e mitigar o desbalanceamento entre as classes. Por fim, as predições individuais foram combinadas em um ensemble. A estratégia de média simples obteve o melhor desempenho, com Kappa Quadrático Ponderado de 0,879 e acurácia de 69,8%. A abordagem preservou a integridade ordinal dos graus, reduzindo quase a zero os erros extremos e demonstrando sua viabilidade clínica.*

1. Introdução

O câncer de próstata é uma das neoplasias malignas mais comuns em homens, representando um desafio significativo para a saúde pública [Siegel et al. 2024]. Estatísticas coletadas pela *International Agency for Research on Cancer* indicam que, em 2022, essa neoplasia foi a quarta mais comum no mundo, representando 7,3% de todos os casos de câncer [Bray et al. 2024]. Neste contexto, a graduação tumoral desempenha um papel

central na definição do manejo clínico e do prognóstico. O sistema de *Gleason*, baseado na soma dos dois padrões histológicos predominantes, evoluiu para o sistema de graduação proposto pela *International Society of Urological Pathology* (ISUP), organizado em cinco grupos, de 1 a 5. Este sistema fornece uma estratificação prognóstica aprimorada e é amplamente utilizado na prática clínica [Epstein et al. 2016].

A avaliação histopatológica tradicional de *Whole-Slide Images* (WSIs) depende da análise manual de patologistas, sendo um processo repetitivo e suscetível à fadiga, o que pode gerar variabilidades inter e intraobservador nos resultados [Bulten et al. 2020]. Nesse contexto, técnicas de Inteligência Artificial (IA) têm sido amplamente exploradas para automatizar a análise histopatológica e auxiliar no diagnóstico do câncer de próstata, visando reduzir essas limitações [Albahri et al. 2022].

Neste contexto, as Redes Neurais Convolucionais (RNCs) têm representado o *estado-da-arte* na análise de imagens médicas [Campanella et al. 2019], sendo capazes de extrair características relevantes e combinar padrões visuais complexos para identificar determinadas condições patológicas [Araújo et al. 2021]. Tal capacidade torna essa abordagem particularmente promissora para a análise automatizada dos padrões de Gleason e dos grupos de grau ISUP. Com isso, a abordagem contribui para a padronização do diagnóstico do câncer de próstata e para a redução das variabilidades intra e interobservador, inerentes à avaliação histopatológica convencional. Assim, além de reduzir o tempo de diagnóstico, a abordagem viabilizaria um manejo clínico mais preciso e reprodutível.

Este trabalho propõe um pipeline estruturado para a classificação automática dos grupos de grau ISUP em imagens histopatológicas de próstata. As principais contribuições são: (i) uma função de perda híbrida que combina penalização ordinal e perda focal, considerando a natureza hierárquica do problema e o desbalanceamento da base de dados; (ii) um método automatizado de remoção de ruído baseado na incerteza estimada pela média das predições dos modelos EfficientNet (B0, B1, B2 e B3); e (iii) a avaliação comparativa das arquiteturas EfficientNet B0, B3 e B7 integradas em ensemble probabilístico.

2. Trabalhos Relacionados

Nesta seção, revisam-se trabalhos recentes que aplicam métodos computacionais à análise de imagens histopatológicas para o diagnóstico do câncer de próstata, destacando-se as bases de dados, as abordagens, os resultados e as limitações.

[Xiang et al. 2023] utilizaram o conjunto interno do PANDA e propuseram o GCN-MIL, baseado em Rede Neural de Grafos com *Robust Training* e ResNet50 como extrator, alcançando κ_{quad} de 93,10%. Entre as limitações, empregaram a *Cross-Entropy* tradicional, desconsiderando a natureza ordinal do ISUP, bem como o uso de uma arquitetura clássica que pode restringir a representação morfológica. [Kosoko et al. 2024] avaliaram CNNs clássicas com Grad-CAM no SICAPv2, obtendo AUC-ROC de 85,0% e métricas médias próximas de 75%, porém sem validação clínica externa e com limitação a arquiteturas tradicionais.

[Afifi et al. 2024] adotaram uma classificação binária com InceptionV3, ResNet50 e InceptionResNetV2 em imagens do *Radboud UMC* e do *Karolinska Institute*, alcançando 91,80% de acurácia; entretanto, a formulação binária ignora a graduação histológica e a avaliação restrita à acurácia é limitada em cenários desbalanceados. De forma

semelhante, [Alici-Karaca and Akay 2024] empregaram EfficientNet-B4 com atenção no DiagSet, reportando acurácias de 96,18% (binária), 94,86% (4 classes) e 93,32% (8 classes), mas utilizaram predominantemente fragmentos de WSI e apenas a acurácia como métrica.

De modo geral, os estudos apresentam limitações comuns: uso exclusivo de acurácia [Afifi et al. 2024, Alici-Karaca and Akay 2024], adoção de CNNs clássicas como extratores [Xiang et al. 2023, Kosoko et al. 2024] e formulações binárias que desconsideram a natureza ordinal da doença [Afifi et al. 2024, Alici-Karaca and Akay 2024]. Para superar tais lacunas, o presente trabalho propõe uma abordagem que trata rótulos ruidosos, incorpora estratégias de Redes Neurais Profundas e *Ensemble*, modela explicitamente a ordinalidade do ISUP por meio de uma função de perda hierárquica e utiliza um conjunto mais robusto de métricas de avaliação.

3. Materiais e Método

Esta seção apresenta os procedimentos metodológicos adotados no desenvolvimento e na avaliação da abordagem proposta. O método, ilustrado na Figura 1, foi estruturado em quatro etapas principais: (i) materiais, (ii) pré-processamento, (iii) extração de características e treinamento dos modelos e (iv) *ensemble*. Cada etapa é descrita detalhadamente nas seções subsequentes.

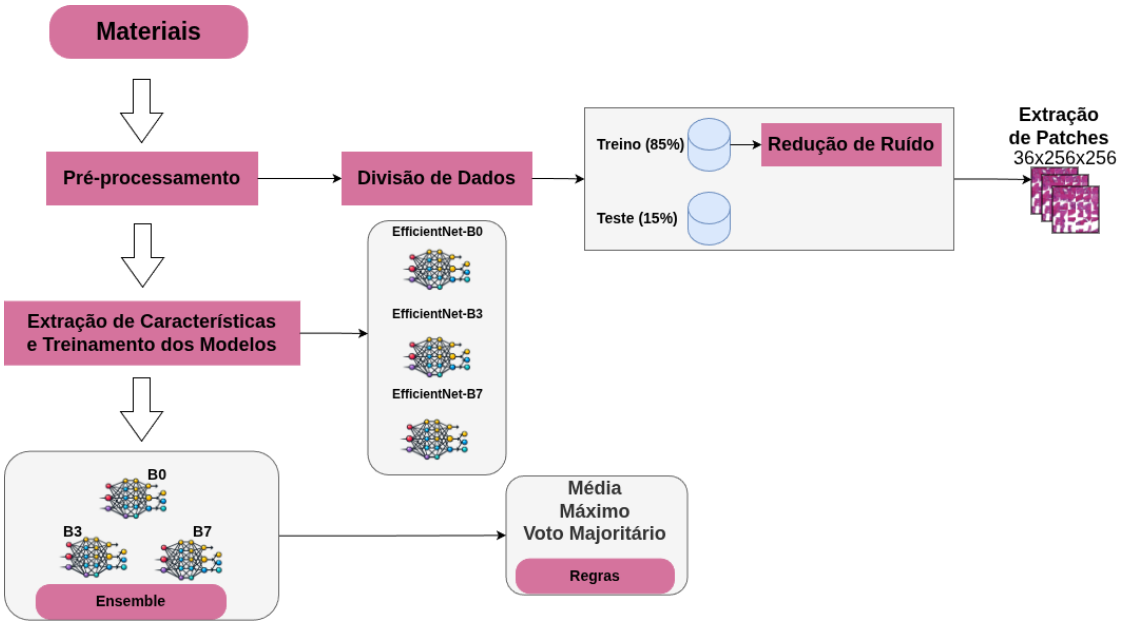


Figura 1. Método Proposto.
Os ícones dos modelos gerados por IA.

3.1. Materiais

Neste trabalho, utilizou-se a base de dados PANDA (*Prostate Cancer Grade Assessment*), composta por aproximadamente 10.616 imagens histopatológicas completas (WSIs) de biópsias prostáticas [PANDA Challenge 2020]. O conjunto de dados foi disponibilizado no contexto do desafio PANDA, promovido na plataforma Kaggle, e é amplamente utilizado em pesquisas de classificação automática do câncer de próstata.

As imagens estão anotadas com escore e grau de Gleason, fornecendo rótulos clínicos para a graduação tumoral. Neste trabalho, utilizou-se exclusivamente a versão pública da base de dados, sem acesso ao conjunto privado do desafio. Por serem provenientes de dois centros médicos (Radboud e Karolinska), os dados apresentam variabilidade institucional, o que os aproxima da prática clínica real. Além disso, a base contém ruídos inerentes ao processo de anotação, como variabilidade interobservador e artefatos visuais (por exemplo, marcas de caneta), o que aumenta a complexidade da tarefa. A Figura 2 ilustra esse tipo de interferência. Observa-se ainda uma distribuição irregular entre as seis classes (ISUP 0–5), com predominância de ISUP 0 e 1, conforme apresentado na Tabela 1.

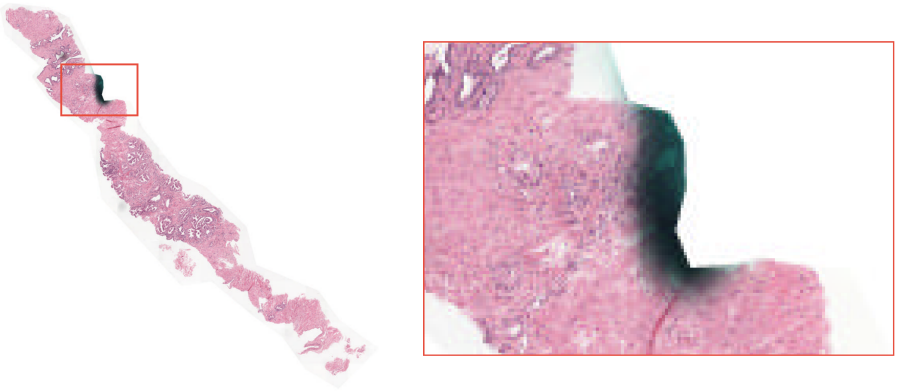


Figura 2. Exemplo de artefato visual em WSI do PANDA.

Grau ISUP	Quantidade de WSIs	Proporção (%)
ISUP 0 (Benigno)	2.892	27,24%
ISUP 1	2.666	25,11%
ISUP 2	1.343	12,65%
ISUP 3	1.242	11,70%
ISUP 4	1.249	11,77%
ISUP 5	1.224	11,53%
Total	**10.616**	**100,00%**

Tabela 1. Distribuição das amostras por grau ISUP do PANDA.

3.2. Pré-processamento

Na fase de pré-processamento, os dados foram inicialmente divididos em conjuntos de treinamento e teste, garantindo a separação adequada entre as etapas de ajuste e de avaliação do modelo. Em seguida, aplicou-se uma técnica de redução de ruído baseada em entropia e, por fim, realizou-se a extração de *patches* a partir da densidade de informação de cada lâmina.

Para a etapa de redução de ruído, utilizaram-se quatro arquiteturas baseadas no *backbone* EfficientNet (B0, B1, B2 e B3) para calcular a entropia de Shannon, a fim de quantificar a incerteza das predições [Ali et al. 2023]. A entropia média dos quatro modelos foi calculada para cada amostra do conjunto de treinamento, e o percentil superior de

10% foi selecionado, formando um banco de dados de amostras potencialmente ruidosas. A partir desse conjunto, avaliaram-se quatro cenários: manutenção integral dos dados e remoção de 100%, 50% e 20% das imagens identificadas como ruidosas. O melhor desempenho foi obtido com a remoção de 20% dessas imagens (180 amostras), indicando que muitas amostras, apesar de apresentarem alta incerteza, contribuem com informações relevantes para o problema no contexto do banco de dados.

Posteriormente, foram extraídos *patches* apenas de regiões teciduais relevantes das WSIs, reduzindo o custo computacional e evitando o processamento de áreas sem informação histológica.

Foram avaliadas três configurações de extração de *patches*: 5×5 (25), 6×6 (36) e 7×7 (49). Conforme a Figura 3, a configuração 5×5 apresenta maior densidade de tecido, porém menor cobertura espacial, podendo perder regiões relevantes; já a 7×7 amplia a cobertura, mas inclui muitas áreas vazias e aumenta significativamente a dimensionalidade e o custo computacional. Assim, optou-se pela configuração intermediária 6×6 (36 *patches*), que equilibra a preservação da informação tecidual e a eficiência computacional.

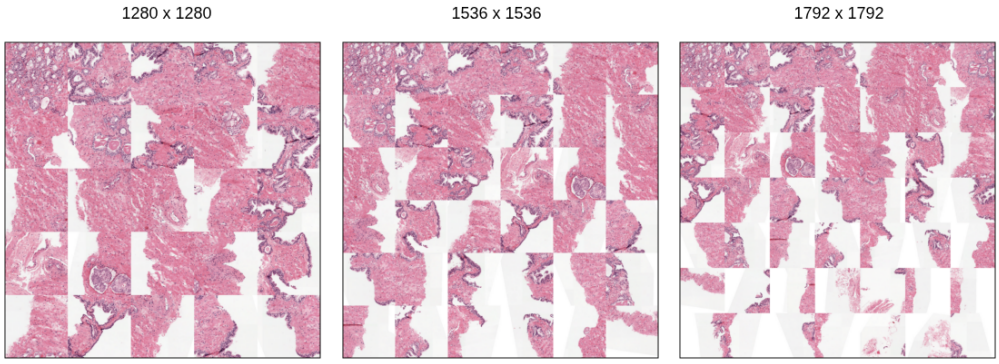


Figura 3. Ilustração da extração de *patches* em diferentes tamanhos: 5x5 (à esquerda), 6x6 (no centro) e 7x7 (à direita).

3.3. Extração de características e treinamento dos modelos

Nesta seção, apresenta-se a abordagem proposta para o treinamento dos classificadores individuais, incluindo a definição dos hiperparâmetros, das estratégias de regularização e da função de perda. Após o pré-processamento, iniciaram-se o treinamento e a avaliação dos classificadores. A EfficientNet-B0 foi utilizada como modelo-base para ajustar os hiperparâmetros. Com a configuração definida, os experimentos foram estendidos às EfficientNet-B3 e B7, permitindo comparar diferentes níveis de complexidade da mesma família e analisar o impacto da maior capacidade representacional no desempenho e na generalização.

Inicialmente, adotou-se a *Binary Cross-Entropy* (BCE) como função de perda *baseline*, em virtude de sua ampla utilização em cenários de classificação multirrótulo. A partir dessa configuração inicial, foram conduzidos experimentos exploratórios para analisar a estabilidade do treinamento e o impacto dos hiperparâmetros na capacidade de generalização do modelo. Foram avaliadas variações na taxa de aprendizado, no número de camadas descongeladas durante o *fine-tuning* e nas estratégias de regularização por

meio de *Dropout*. Para a seleção da melhor combinação de hiperparâmetros, empregou-se *grid search* no modelo EfficientNet-B0, utilizando um subconjunto correspondente a 20% do conjunto de treinamento, reservado para validação interna, o que permitiu identificar a configuração com melhor desempenho antes do treinamento final.

Apesar da avaliação dessas configurações no treinamento, observou-se um desempenho estagnado que poderia estar associado não apenas à escolha dos hiperparâmetros, mas também à formulação da própria função de perda. Considerando que a graduação do câncer de próstata é ordinal, em que os erros entre classes adjacentes são menos severos do que entre classes mais distantes, o problema foi reformulado como uma regressão ordinal.

Nessa formulação, a predição do grau ISUP não é tratada como um problema multiclasse com categorias independentes, e sim como uma sequência de classificações binárias cumulativas. Para um grau ISUP $k \in \{0, \ldots, 5\}$, o rótulo é codificado como um vetor binário cumulativo, no qual os k primeiros elementos assumem valor 1 e os demais, 0; por exemplo, o grau 3 é representado por $[1, 1, 1, 0, 0]$. A perda ordinal é então aplicada de forma independente a cada limiar do vetor, o que faz com que o erro total corresponda à soma das discrepâncias cumuladas. Essa modelagem impõe uma penalização proporcional à distância entre as classes, de modo que erros mais distantes acarretam maior custo, preservando a estrutura hierárquica do problema.

Para mitigar o desbalanceamento dos dados (Seção 3.1), a perda ordinal foi combinada à *Focal Loss* [Lin et al. 2017, Dina et al. 2023, Chen 2026]. Esta função modula a entropia cruzada para reduzir a influência de instâncias classificadas com alta confiança, priorizando amostras de maior incerteza diagnóstica. No presente estudo, configurou-se o fator de foco $\gamma = 2,0$ para atenuar a contribuição de exemplos fáceis e o peso $\alpha = 0,25$ para equilibrar a importância relativa entre as classes. Essa estratégia direciona a otimização para os casos mais complexos e sub-representados, fundamentais para a distinção precisa entre os graus ISUP.

3.4. Ensemble

Para maximizar a capacidade preditiva e a robustez do sistema de diagnóstico proposto, a etapa final do método consistiu na combinação das predições geradas pelos classificadores individuais (EfficientNet-B0, EfficientNet-B3 e EfficientNet-B7). A estratégia de *ensemble* foi adotada para explorar a diversidade estrutural e a complementaridade das capacidades representacionais dessas arquiteturas, reduzindo o impacto de vieses específicos e de erros individuais. Com o intuito de identificar a estratégia de fusão mais eficaz, foram implementadas e avaliadas três regras de combinação.

Adotaram-se três estratégias clássicas de *ensemble* [Zhou 2025]: a votação majoritária, em que a decisão é a classe mais frequente entre os modelos; a regra do máximo, que atribui a saída à classe com a maior probabilidade individual, priorizando a predição de maior confiança; e a média das probabilidades, que utiliza a média aritmética das estimativas probabilísticas para uma agregação mais estável.

As predições consolidadas por meio dessas abordagens foram comparadas sistematicamente aos resultados dos modelos individuais. As métricas de desempenho resultantes desse processo são apresentadas e discutidas na Seção 4, permitindo analisar o

impacto da técnica de combinação sobre a estabilidade e a capacidade de generalização do sistema.

3.5. Métricas de avaliação

Foram empregadas as métricas de acurácia e Kappa Quadrático Ponderado (κ_{quad}), sendo a primeira responsável por medir a proporção de predições corretas e a segunda por avaliar a concordância entre rótulos preditos e reais com penalização proporcional à magnitude do erro [Cohen 1968]. A variabilidade e os intervalos de confiança foram estimados por *bootstrap* não paramétrico, com 1.000 reamostragens, a partir das quais se calcularam a média, o desvio padrão e o intervalo de confiança de 90% (percentis 5% e 95%).

4. Resultados

Esta seção apresenta o ambiente de implementação, a seleção do melhor pipeline de configurações, o impacto da função de perda híbrida Ordinal-Focal e os resultados obtidos com as estratégias de *ensemble*.

4.1. Ambiente de Implementação

Os experimentos foram conduzidos em uma estação Linux com GPU RTX 3060, CPU AMD Ryzen 5 5600X e 32 GB de RAM DDR4. Além disso, utilizaram-se PyTorch, CUDA e OpenSlide para a leitura de WSIs e a extração de *patches*.

4.2. Etapa inicial - Configurando hiperparâmetros

A etapa inicial de treinamento, conduzida conforme descrito na Seção 3, permitiu definir a melhor configuração de hiperparâmetros do modelo. Os valores avaliados e a configuração selecionada são apresentados na Tabela 2.

Parâmetro	**Valores avaliados**	**Valor selecionado**
Taxa de aprendizado	3×10^{-4}, 3×10^{-3}, 1×10^{-4}	3×10^{-4}
Número de épocas	50	50
Camadas descongeladas (fine-tuning)	100, 150	150
Dropout	0.4, 0.5, 0.6	0.4
Tamanho do *batch*	2	2
Função de perda	BCE	BCE
Otimizador	Adam	Adam

Tabela 2. Hiperparâmetros avaliados

4.3. Impacto da Remoção de Ruído e Formulações de Perda

A Tabela 3 apresenta a evolução incremental do desempenho da *EfficientNet-B0*. As estratégias foram aplicadas cumulativamente, de modo que cada configuração incorpora as melhorias da etapa anterior.

Inicialmente, a remoção de ruído foi aplicada ao modelo *baseline*, elevando o κ_{quad} de $0{,}826$ para $0{,}833$. Esse resultado indica que a correção de inconsistências no conjunto de dados contribuiu para maior estabilidade no treinamento. Na sequência, mantendo a

remoção de ruído, a substituição da BCE pela perda ordinal resultou em um salto para $0{,}851$. Esse avanço demonstra que a modelagem explícita da relação de ordem entre as classes melhora a qualidade das predições. Essa configuração também apresentou o menor desvio padrão ($0{,}009$), indicando maior estabilidade. Por fim, a estratégia híbrida foi aplicada à formulação ordinal, resultando no melhor desempenho observado ($\kappa_{\text{quad}} = 0{,}856$). A combinação permitiu tratar simultaneamente a estrutura ordinal das classes e o desbalanceamento do conjunto de dados. No total, a sequência de refinamentos resultou em um ganho absoluto de 3% em relação ao modelo *baseline*.

	Kappa Quadrático			**Acurácia**		
Configuração	**Resultado**	**Std**	**CI 95%**	**Resultado**	**Std**	**CI 95%**
BCE (*baseline*)	0,826	0,012	[0,806; 0,846]	0,592	0,012	[0,572; 0,612]
BCE + Remoção de Ruído	0,833	0,013	[0,812; 0,853]	0,638	0,116	[0,619; 0,657]
Perda Ordinal	0,851	**0,009**	[0,833; 0,869]	0,608	0,126	[0,587, 0,628]
Ordinal + Focal (Híbrida)	**0,856**	0,011	[0,833; 0,876]	**0,669**	0, 011	[0,635; 0,683]

Tabela 3. Impacto da formulação da função de perda no desempenho da EfficientNet-B0

4.4. Comparação entre Arquiteturas EfficientNet

Após a definição da melhor combinação de hiperparâmetros e da função de perda utilizando a EfficientNet-B0, a mesma configuração experimental foi aplicada às variantes EfficientNet-B3 e EfficientNet-B7, que representam modelos de complexidade intermediária e elevada, respectivamente. O objetivo foi avaliar se o refinamento metodológico desenvolvido para o modelo base produziria ganhos adicionais quando empregado em arquiteturas com maior capacidade representacional dentro da mesma família.

Os resultados são apresentados na Tabela 4. Observa-se que a EfficientNet-B0 apresentou o melhor desempenho tanto no kappa quadrático quanto na acurácia. As arquiteturas mais profundas (B3 e B7) não apresentaram ganho estatisticamente relevante, havendo, inclusive, leve degradação nos valores médios das métricas avaliadas. Tal comportamento pode estar associado ao tamanho da base de dados ou à maior propensão ao sobreajuste em modelos com um número maior de parâmetros. Assim, a EfficientNet-B0 demonstrou ser a que apresenta o melhor equilíbrio entre desempenho preditivo, estabilidade estatística e custo computacional.

	Kappa Quadrático			**Acurácia**		
Arquitetura	**Resultado**	**Std**	**CI 95%**	**Resultado**	**Std**	**CI 95%**
EfficientNet-B0	**0,857**	0,011	[0,833; 0,876]	**0,669**	0,119	[0,635; 0,683]
EfficientNet-B3	0,846	0,010	[0,829; 0,861]	0,659	0,125	[0,554; 0,594]
EfficientNet-B7	0,844	0,011	[0,821; 0,865]	0,602	0,112	[0,578; 0,628]

Tabela 4. Comparação de desempenho entre as variantes da EfficientNet.

No contexto clínico, os modelos demonstraram desempenho globalmente satisfatório, com predominância de erros entre classes adjacentes — comportamento esperado em problemas de classificação ordinal, como o sistema ISUP. Observa-se que as maiores taxas de confusão concentram-se nas classes intermediárias (ISUP 2, ISUP 3 e

ISUP 4), conforme evidenciado na Tabela 5, sugerindo que os modelos conseguem, em grande parte, preservar a proximidade semântica entre os graus de agressividade tumoral. Entretanto, permanecem limitações relevantes do ponto de vista clínico. Destaca-se a baixa precisão na identificação da classe ISUP 5, a mais agressiva e de maior impacto prognóstico. Mesmo o modelo de melhor desempenho apresentou taxa de acerto inferior a 50% para essa classe, indicando dificuldade na discriminação dos casos mais severos. Esse achado sugere a necessidade de estratégias adicionais, como o uso de *ensemble*.

Modelo	Classe Real	Prev. ISUP 0	Prev. ISUP 1	Prev. ISUP 2	Prev. ISUP 3	Prev. ISUP 4	Prev. ISUP 5
B0	ISUP 0	**0,857**	0,088	0,014	0,021	0,021	0,000
	ISUP 1	0,095	**0,693**	0,188	0,022	0,003	0,000
	ISUP 2	0,015	0,230	**0,485**	0,230	0,040	0,000
	ISUP 3	0,027	0,032	0,146	**0,492**	0,270	0,032
	ISUP 4	0,037	0,016	0,048	0,134	**0,652**	0,112
	ISUP 5	0,033	0,011	0,016	0,065	0,380	**0,495**
B3	ISUP 0	**0,818**	0,152	0,009	0,012	0,007	0,002
	ISUP 1	0,075	**0,708**	0,200	0,010	0,007	0,000
	ISUP 2	0,015	0,195	**0,530**	0,210	0,050	0,000
	ISUP 3	0,038	0,043	0,205	**0,432**	0,216	0,065
	ISUP 4	0,037	0,053	0,059	0,134	**0,572**	0,144
	ISUP 5	0,049	0,005	0,027	0,076	0,201	**0,641**
B7	ISUP 0	**0,788**	0,173	0,023	0,007	0,007	0,002
	ISUP 1	0,140	**0,640**	0,203	0,018	0,000	0,000
	ISUP 2	0,015	0,215	**0,525**	0,195	0,050	0,000
	ISUP 3	0,022	0,059	0,178	**0,357**	0,297	0,086
	ISUP 4	0,043	0,053	0,059	0,241	**0,433**	0,171
	ISUP 5	0,033	0,022	0,011	0,082	0,266	**0,587**

Tabela 5. Previsões por Modelo e Classe Real (ISUP)

4.5. Desempenho das Estratégias de Ensemble

Como última etapa dos experimentos, buscou-se explorar o comportamento complementar dos classificadores (B0, B3 e B7) por meio de uma estratégia de *ensemble*, com o objetivo principal de melhorar a concordância geral do modelo, ao mesmo tempo em que se reduzem os erros clinicamente mais graves, especialmente aqueles que podem subestimar a agressividade tumoral.

Modelo / Estratégia	Kappa	Std	IC 95%	Acurácia	Std	IC 95%
B0 - *Baseline*	0,826	0,012	[0,806; 0,846]	0,592	0,012	[0,572; 0,612]
B0 - Ordinal Focal	0,856	0,011	[0,833; 0,876]	0,660	0,012	[0,635; 0,683]
Ensemble - Média	**0,879**	0,010	[0,858; 0,898]	**0,698**	0,011	[0,676; 0,721]
Ensemble - Votação	0,859	0,011	[0,835; 0,882]	0,692	0,011	[0,668; 0,715]
Ensemble - Máximo	0,839	0,011	[0,815; 0,859]	0,582	0,012	[0,559; 0,605]

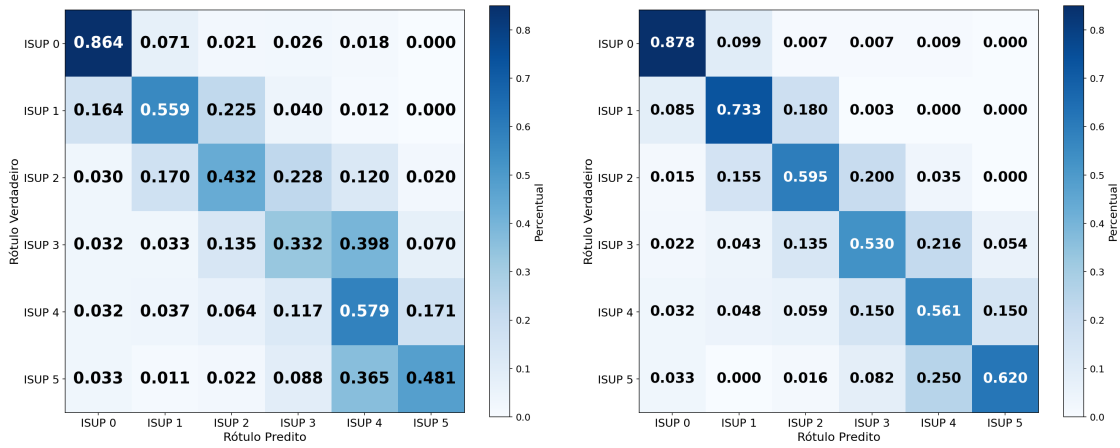
Tabela 6. Comparação entre os resultados individuais e estratégias de Ensemble.

4.6. Discussão

Com base nos experimentos realizados, a função de perda híbrida demonstrou impacto direto na robustez do treinamento, elevando a acurácia do *baseline* de 0,592 para 0,660 (+7 p.p.) e o κ_{quad} de 0,826 para 0,856 (+3 p.p.), conforme apresentado na Tabela 6 da Seção 4.

O melhor desempenho global foi obtido pelo *ensemble* formado por EfficientNet-B0, EfficientNet-B3 e EfficientNet-B7 (média simples), alcançando $\kappa_{\text{quad}} = 0,879$ e acurácia de 0,698, superando o *baseline* em 5,3% e 10,6%, respectivamente.

A superioridade do *ensemble* torna-se ainda mais evidente na comparação das matrizes de confusão (Figuras 4a e 4b). Observa-se maior concentração de acertos na diagonal principal em todas as classes, com redução consistente de erros graves. Erros com distância superior a duas classes (por exemplo, ISUP 0 predito como ISUP ≥ 3) foram reduzidos de 3,70% para 2,20%, o que representa uma queda absoluta de 1,5 p.p. e uma redução relativa de aproximadamente 40%, reforçando o ganho clínico da abordagem proposta.



(a) Matriz de Confusão - *Baseline*

(b) Matriz de Confusão - *Ensemble* (Média Simples)

Figura 4. Comparação entre o *baseline* e o melhor *ensemble*.

A análise comparativa com trabalhos anteriores (Tabela 7) evidencia que, embora alguns estudos reportem métricas absolutas elevadas, frequentemente o fazem em cenários simplificados, com validações menos desafiadoras ou com métricas isoladas. Por exemplo, [Xiang et al. 2023] obteve $\kappa = 0,931$ em validação interna, porém 0,801 em validação externa; em contraste, nosso modelo alcança $\kappa = 0,879$ diretamente nos dados públicos ruidosos do PANDA, indicando maior robustez prática. De forma semelhante, [Afifi et al. 2024] reportou acurácia de 91,80%, mas $\kappa = 0,608$, o que sugere discrepâncias ordinais relevantes.

Diferentemente dessas abordagens, o modelo proposto realiza a graduação completa dos seis níveis do ISUP (0-5), incorpora explicitamente a natureza ordinal da doença por meio de uma função de perda sensível à hierarquia e reduz erros extremos, proporcionando um equilíbrio superior entre desempenho estatístico, coerência clínica e capacidade de generalização.

5. Conclusão

Este trabalho apresentou uma metodologia consistente para a classificação automática dos graus ISUP em imagens histopatológicas de próstata, abordando desafios relevantes

Trabalho	Banco de Dados	Abordagem / Modelo	Tarefa	Acurácia	Kappa
[Xiang et al. 2023]	PANDA	ResNet50 + GCN	ISUP (6 níveis)	-	0.931 (Int.) / 0.801 (Ext.)
[Alici-Karaca and Akay 2024]	DiagSet	EfficientNet-B4 + ECA	Binária / 4 classes	0,961 / 0,948	-
[Kosoko et al. 2024]	SICAPv2	VGG19	Escore de Gleason	0,780	-
[Afifi et al. 2024]	PANDA	InceptionResNetV2	Escore de Gleason	0,918	0.608
Abordagem Proposta	**PANDA**	**Ensemble + Ordinal Focal Loss**	**ISUP (6 níveis)**	**0,698**	**0.879**

Tabela 7. Comparação do modelo proposto com trabalhos recentes da literatura.

como o desbalanceamento de classes, o ruído nos rótulos e a natureza ordinal da progressão tumoral. A adoção de uma função de perda híbrida, combinando regressão ordinal e Focal Loss, elevou o índice κ_{quad} de 0,826 para 0,856 e a acurácia de 59,2% para 66,0%, sugerindo que a modelagem explícita da hierarquia entre os graus contribui para maior coerência nas predições. A filtragem de amostras de alta entropia proporcionou maior estabilidade ao treinamento, e o melhor desempenho foi obtido por meio de um ensemble baseado na média das probabilidades dos modelos EfficientNet-B0, B3 e B7, alcançando κ_{quad} de 0,879 e acurácia de 69,8%. Embora esses resultados sejam promissores, a acurácia, isoladamente, é uma métrica limitada nesse contexto, por não considerar a gravidade dos erros. O coeficiente κ_{quad}, por sua vez, mostrou-se mais alinhado à prática clínica, ao penalizar discrepâncias entre classes distantes. O desempenho obtido indica uma redução de erros extremos, aspecto particularmente relevante para o suporte à decisão diagnóstica. Ainda assim, a análise dos erros revela que, embora predominem confusões entre classes adjacentes, persistem desafios nas classes intermediárias, que apresentam maior ambiguidade e maior relevância clínica. Do ponto de vista aplicado, os resultados sugerem potencial para uso como ferramenta de apoio, especialmente na padronização das avaliações. No entanto, sua utilização prática requer validações adicionais em bases externas e em diferentes cenários clínicos.

Como limitação, destaca-se o uso de um único banco de dados, ainda que multicêntrico e heterogêneo. Como trabalhos futuros, propõe-se investigar arquiteturas baseadas em *Transformers*, incluindo *Vision Transformers* e modelos híbridos *CNN-Transformer*, bem como explorar diferentes espaços de cores, com foco na melhoria do desempenho em classes intermediárias e no fortalecimento da generalização do modelo.

Referências

Afifi, A., Alrahmawy, M., and El-Baz, A. (2024). Prostate cancer detection using deep learning models on histopathological image slides: An experimental analysis. *International Journal of Advanced Computer Science and Applications*, 15(1):xxx–xxx.

Albahri, O. S., Albahri, A. S., Mohammed, K. I., Zaidan, A., Zaidan, B. B., Hashim, M., Salman, O. H., Alaa, M., and Alsalem, M. A. (2022). Expert system research for prostate cancer: A literature review. *Computers in Biology and Medicine*, 145:105487.

Ali, I., Khan, A., Khan, M., Ahmad, R., and Ullah, I. (2023). Shannon entropy in artificial intelligence and its applications based on information theory. *Entropy*, 25(2):364.

Alici-Karaca, D. and Akay, B. (2024). An efficient deep learning model for prostate cancer diagnosis. *IEEE Access*, 12:150776–150790.

Araújo, F. H., Silva, R. R., Medeiros, F. N., Neto, J. F. R., Oliveira, P. H. C., Bianchi, A. G. C., and Ushizima, D. (2021). Active contours for overlapping cervical cell segmentation. *International Journal of Biomedical Engineering and Technology*, 35(1):70–92.

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., and Jemal, A. (2024). Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74(3):229–263.

Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., van der Laak, J., Hulsbergen-van de Kaa, C., and Litjens, G. (2020). Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *The Lancet Oncology*, 21(2):233–241.

Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Silva, V. W. K., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., and Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(9):1301–1309.

Chen, X. (2026). A transfer learning-based deep focal multiclass network for psychological emotion recognition in community-correction populations. *Alexandria Engineering Journal*, 135:235–242.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

Dina, A. S., Siddique, A., and Manivannan, D. (2023). A deep learning approach for intrusion detection in internet of things using focal loss function. *Internet of Things*, 22:100699.

Epstein, J. I., Egevad, L., Amin, M. B., Delahunt, B., Srigley, J. R., and Humphrey, P. A. (2016). The 2014 International Society of Urological Pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma: Definition of grading patterns and proposal for a new grading system. *The American Journal of Surgical Pathology*, 40(2):244–252.

Kosoko, I., Garg, A., Jain, S., and Hewage, P. (2024). Leveraging deep learning and explainable ai for diagnosis of prostate cancer. In *Proceedings of the International Conference on Applied Artificial Intelligence in Medical Imaging*, Bolton, United Kingdom.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988.

PANDA Challenge (2020). Prostate cANcer grade assessment challenge. `https://www.kaggle.com/c/prostate-cancer-grade-assessment`. Accessed: 2024-01-01.

Siegel, R. L., Giaquinto, A. N., and Jemal, A. (2024). Cancer statistics, 2024. *CA: A Cancer Journal for Clinicians*, 74(1):12–49.

Xiang, J., Wang, X., Wang, X., Zhang, J., Yang, S., Yang, W., Han, X., and Liu, Y. (2023). Automatic diagnosis and grading of prostate cancer with weakly supervised learning on whole slide images. *Computers in Biology and Medicine*, 152:106340.

Zhou, Z.-H. (2025). *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC.