Classificação Automática de Tecidos do Microambiente Tumoral no Câncer Gástrico com *Ensemble* por *Stacking*

François F. R. Barbosa[1,2], **Ivan S. Silva**[1], **Rodrigo M. S. Veras**[1], **Rodrigo N. Borges**[1], **Isaac S. S. Ramos**[1], **Mariana Recamonde-Mendoza**[3]

[1] Departamento de Computação, Universidade Federal do Piauí (UFPI)
CEP: 64.049-550 – Teresina – PI, Brasil.
[2] Instituto Federal de Educação, Ciência e Tecnologia do Maranhão (IFMA)
CEP: 65604-500 – Caxias – MA, Brasil.
[3] Instituto de Informática, Universidade Federal do Rio Grande do Sul (UFRGS)
CEP: 91501-970 – Porto Alegre – RS, Brasil.

francois.barbosa@ifma.edu.br

Abstract. *Gastric cancer remains a major cause of cancer-related mortality, and characterization of the tumor microenvironment (TME) is essential for prognosis and therapeutic planning. However, the morphological heterogeneity of gastric tissues still poses challenges for automated classification, despite advances in deep learning (DL). This study proposes a stacking-based ensemble of DL models for multiclass classification of TME tissues in H&E-stained histopathological images using the HMU-GC-HE-30K dataset. Eight pretrained architectures were employed as base models, and their predictions were combined by an MLP meta-model capable of refining decision boundaries by learning a nonlinear combination of the base-model outputs. Data augmentation and transfer learning were adopted to mitigate dataset limitations and improve generalization. The results indicate that the tuned_stacking ensemble statistically outperformed the baseline, achieving an F1-score of 81.3%, surpassing individual models and suggesting increased robustness in distinguishing complex tissue classes. Overall, the proposed method shows potential to support decision-making in digital pathology.*

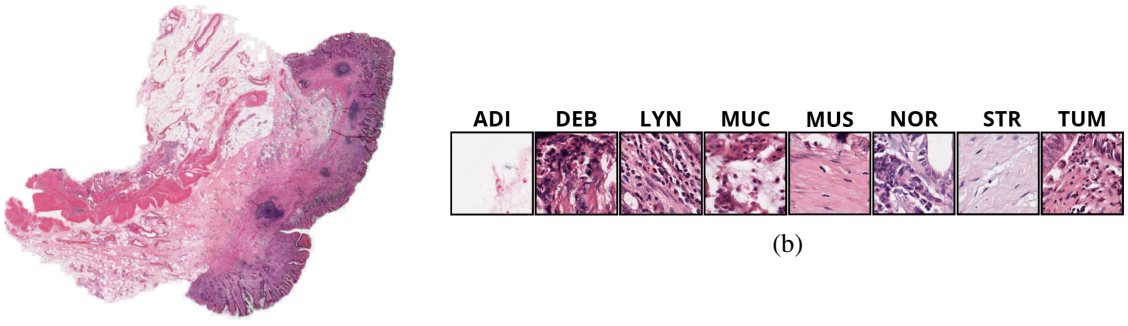
Resumo. *O câncer gástrico permanece como uma importante causa de mortalidade relacionada ao câncer, e a caracterização do Microambiente Tumoral (Tumor Microenvironment – TME) é essencial para o prognóstico e o planejamento terapêutico. No entanto, a heterogeneidade morfológica dos tecidos gástricos ainda impõe desafios à classificação automatizada, mesmo diante dos avanços em Deep Learning (DL). Este estudo propõe um ensemble por empilhamento (stacking) de modelos de DL para a classificação multiclasse de tecidos do TME em imagens histopatológicas H&E, utilizando o dataset HMU-GC-HE-30K. Oito arquiteturas pré-treinadas foram empregadas como modelos base e suas predições foram combinadas por um meta-modelo MLP, capaz de aprimorar as fronteiras de decisãoao aprender uma combinação não linear das predições dos modelos base. Técnicas de aumento de dados e aprendizado por transferência foram adotadas para mitigar limitações do conjunto de dados e aumentar a capacidade de generalização. Os resultados indicam*

que o stacking_ajustado superou estatisticamente o baseline alcançando 81,3% de F1-score, superando os modelos individuais e sugerindo maior robustez na distinção de classes teciduais complexas. Assim, o método demonstra potencial para apoiar a decisão em patologia digital.

1. Introdução

O entendimento contemporâneo do câncer ultrapassa a visão de uma doença exclusivamente genética, passando a ser compreendido como um ecossistema complexo e heterogêneo [de Visser and Joyce 2023]. Nesse ecossistema, encontram-se uma variedade de células não cancerígenas que interagem com as tumorais, formando o que é chamado de microambiente tumoral (TME – *tumor microenvironment*) (Figura 1-a). Esse microambiente desempenha um papel importante no desenvolvimento da doença, assim como na resposta terapêutica, tornando a identificação e a distinção dos diferentes componentes do TME (Figura 1-b) particularmente relevantes no contexto do câncer gástrico (CG), em que padrões morfológicos do microambiente podem refletir características clínicas e biológicas da doença [Lou et al. 2025].

Tradicionalmente, o diagnóstico do CG a partir de lâminas histopatológicas coradas por *Hematoxilina e Eosina* (H&E) depende, tradicionalmente, da inspeção visual do patologista. Embora consolidado, esse método é trabalhoso, demorado e subjetivo, pois depende da experiência do patologista. Esses fatores comprometem a acurácia diagnóstica, especialmente em cenários com alta demanda e disponibilidade limitada de especialistas [Wang et al. 2024].



(a) (b)

Figura 1. TME no câncer gástrico: (a) lâmina destacando a heterogeneidade morfológica; (b) *patches* rotulados desses tecidos. Adaptado de [Lou et al. 2025].

Segundo o Relatório Global de Estatísticas do Câncer de 2022, do Observatório Global de Câncer (*Global Cancer Observatory* – GCO) [Ferlay et al. 2024], o CG ocupa a quinta posição em incidência no mundo, representando 5,96% de todos os casos de câncer. Além disso, figura como a quinta principal causa de morte por câncer (Figura **??**), com mais de 660 mil óbitos anuais.

Nesse contexto, algoritmos de Aprendizado de Máquina (*ML - Machine Learning*), especialmente aqueles baseados em Redes Neurais Convolucionais (*CNN - Convolutional Neural Network*), têm se mostrado uma alternativa de expressivo crescimento

em pesquisas oncológicas, obtendo excelentes resultados na análise de imagens digitalizadas e no suporte ao diagnóstico clínico [Mandal et al. 2025]. Nesse contexto, a capacidade dessas redes em extrair, das imagens, características morfológicas relevantes supera limitações subjetivas da análise humana, contribuindo para maior acurácia diagnóstica.

A utilização de empilhamento de modelos (*stacking*) para potencializar o desempenho das CNNs é uma estratégia de *ensemble* baseada na combinação de predições de múltiplos modelos para produzir uma decisão final por meio de um meta-modelo (*meta-learner*), que aprende automaticamente a melhor forma de ponderar os erros e acertos de cada componente. Como diferentes arquiteturas tendem a capturar padrões morfológicos complementares, o *stacking* é particularmente promissor em tarefas de classificação multiclasse dos componentes do TME ([Naimi and Balzer 2018, Kablan et al. 2023]).

Frente ao crescente interesse da patologia moderna em caracterizar os componentes do TME, este estudo propõe o desenvolvimento de um *ensemble* baseado em empilhamento de modelos de DL para a classificação multiclasse de imagens histológicas do TME do câncer gástrico, visando distinguir as diferentes estruturas morfológicas que compõem esse microambiente.

A Seção 2 deste trabalho apresenta estudos com escopo voltado para a classificação de componentes do TME. Na Seção 3 é detalhada a metodologia empregada, juntamente com as condições experimentais. Os resultados são exibidos e discutidos na Seção 4 e por fim, a Seção 5 contém as conclusões obtidas no estudo, bem como direcionamento de pesquisas futuras.

2. Trabalhos Relacionados

Na literatura existente, muitos autores se dedicaram a explorar o uso de algoritmos de IA para classificar, segmentar ou contabilizar componentes do TME, para as mais diversas finalidades, desde propor um novo biomarcador até a predição de um evento de risco, como o óbito do paciente. No entanto, a classificação precisa desses componentes é essencial para garantir que as informações serão extraídas dos componentes corretos.

No contexto de biomarcadores derivados de histopatologia, Kather *et al.* [Kather et al. 2019] apresentaram o Deep Stroma Score, um escore baseado em DL voltado à quantificação de componentes do TME e à predição de sobrevida em CRC. Para isso, os autores empregaram quatro *datasets* corados em H&E: dois destinados à classificação dos componentes teciduais (com nove classes) e dois utilizados na etapa de cálculo do escore. O treinamento do classificador foi conduzido no NCT-CRC-HE-100K, composto por 100.000 recortes, enquanto a validação externa foi realizada no CRC-VAL-HE-7K (7.180 imagens). Em seguida, a CNN treinada foi aplicada para decomposição tecidual em coortes independentes do *The Cancer Genome Atlas* (TCGA), com 862 lâminas de 500 pacientes com dados clínicos completos e do *Darmkrebs: Chancen der Verhütung durch Screening* (DACHS), com 409 lâminas de 409 pacientes. Na etapa de classificação, o método alcançou acurácia global de 94,0%.

Rong *et al.* [Rong et al. 2022] apresentaram um sistema para segmentação e classificação de núcleos celulares no TME, avaliando-o em múltiplas coleções com diferentes órgãos e cenários clínicos. Foram utilizados quatro conjuntos de dados: (i) NLST (adenocarcinoma pulmonar), com 127 patches de 500×500 pixels rotulados em seis tipos

de núcleos; (ii) um conjunto hepático com 76 lâminas (tecidos humano e murino, com ou sem carcinoma), distribuído em quatro classes; (iii) NuCLS, com 1.744 patches de câncer de mama provenientes do TCGA, também em quatro classes; e (iv) TCGA-BRCA, empregado na análise de sobrevivência, contendo 1.061 lâminas com dados clínicos associados. Na tarefa de classificação, os autores reportaram, para o conjunto hepático, acurácia de 84,4% (sem divulgação de F1-score). No NLST, foram obtidos 71,1% de acurácia e 74,1% de F1-score. O melhor desempenho foi observado no NuCLS, onde um *ensemble* de modelos alcançou 86,6% de acurácia e 89,4% de F1-score.

Mikhailov et al. [Mikhailov et al. 2022] investigaram o TME no CG com ênfase em duas frentes: (i) classificação de tecidos da parede gástrica e (ii) inferência da profundidade de invasão tumoral. Para favorecer o desempenho dos classificadores, os autores definiram cinco classes teciduais no processo de anotação e rotulagem. As imagens pertencem ao conjunto PATH-DT-MSU, construído no departamento de patologia do Centro Médico-Científico da Universidade Estatal de Moscou, composto por 20 lâminas anotadas manualmente a partir de fragmentos de parede gástrica. A geração de amostras foi realizada via janela deslizante, resultando em 70.871 patches para treinamento e 14.462 para teste. O modelo preditivo alcançou acurácia global de 91,7% e apresentou desempenho particularmente elevado na identificação da muscular da mucosa, apontada como marcador-chave para a avaliação da invasão tumoral.

Um estudo semelhante ao anterior foi o de Lou *et al.* [Lou et al. 2025], em que os autores desenvolveram um extenso conjunto de dados de imagens histológicas do TME de CG, denominado HMU-GC-HE-30K. Trata-se do conjunto de dados adotado no presente trabalho e, até onde se tem conhecimento, do único estudo que o utiliza. As imagens foram obtidas a partir de 300 lâminas coradas com H&E, provenientes do *Cancer Hospital of Harbin Medical University*. As lâminas foram divididas em 31.096 *patches* de 224×224 *pixels*, e utilizou-se o classificador proposto por Kather *et al.* [Kather et al. 2019] para dividi-las em oito classes de tecidos. Para a classificação, os autores empregaram os modelos ViT e EfficientNet-B0, ambos com pesos pré-treinados. Um subconjunto correspondente a 20% dos dados foi separado para os testes, sendo o melhor resultado alcançado pelo EfficientNet-B0, alcançando 70% de acurácia global e 71,0% de *F1-score*.

No contexto de diagnóstico universal e leve em imagens de histopatologia, Su *et al.* [Su et al. 2025] propuseram o *framework* Pathology-NAS, que utiliza o conhecimento de Modelos de Linguagem de Grande Escala (Large Language Models - LLMs) para otimizar a busca de arquiteturas neurais (Neural Architecture Search - NAS) em diversos cenários clínicos. Na tarefa de classificação de componentes do TME em CG foi utilizado o *dataset* HMU-GC-HE-30K, alcançando acurácia de 63,15% sem apresentar informações sobre *F1-score*.

3. Materiais e Métodos

Foram avaliadas oito arquiteturas pré-treinadas sob duas configurações experimentais: (i) treinamento sem aumento e (ii) treinamento com aumento de dados (*data augmentation*). Em ambos os casos, adotou-se um protocolo de validação cruzada estratificada e manteve-se o mesmo cenário de ajuste de hiperparâmetros, a fim de assegurar comparabilidade entre as abordagens. Por fim, as predições gerados pelos modelos treinados serviram de artefatos para a construção de *ensemble* baseado em *stacking*.

3.1. Base de Imagens

A fim de conduzir os experimentos de classificação do TME em CG, utilizou-se a base HMU-GC-HE-30K [Lou et al. 2025], disponibilizada publicamente via *Figshare* [Lou et al. 2024]. O *dataset* foi construído a partir de 300 lâminas histológicas coradas com H&E, provenientes do Hospital de Câncer da Universidade Médica de Harbin (China). As lâminas foram particionadas em 31.096 *patches* não sobrepostos de 224×224 pixels, igualmente distribuídos em oito classes (ADI, DEB, MUC, MUS, LYM, STR, NOR e TUM) (Figura 1-b), totalizando 3.887 imagens por classe, o que dispensou a necessidade posterior de balanceamento.

3.2. Estratégias de treinamento

Inicialmente, 20% das imagens foram separadas, de forma estratificada, como conjunto de teste independente, permanecendo totalmente fora do processo de treinamento. Os 80% restantes foram então avaliados por validação cruzada *k-fold*, também estratificada. Para preservar a proporção aproximada de 70/10/20 (treino/validação/teste), adotou-se $k = 8$, de modo que, em cada iteração, fosse usado para validação o correspondente a 10% do conjunto completo.

Para a otimização dos hiperparâmetros, foi utilizado o *GridSearch*, adotando-se como intervalo de busca, configurações frequentemente reportadas na literatura. Foram avaliadas três taxas de aprendizagem ($1 \times 10^{-3}, 1 \times 10^{-4}$ e 9×10^{-5}), dois otimizadores (Adam e SGD), dois valores de *weight decay* (1×10^{-3} e 1×10^{-4}), três opções de *batch size* (32, 64 e 128) e quatro valores de *dropout* (0,2; 0,3; 0,4 e 0,5). Como melhor configuração, obteve-se taxa de aprendizagem de 9×10^{-5} otimizador Adam, *weight decay* de 1×10^{-4}, *batch size* igual a 64 e *dropout* de 0,5. Todos os modelos foram treinados de forma fim a fim, com uma parada antecipada configurada para interromper o treinamento após 10 épocas sem melhora na *loss* da validação.

Os experimentos foram executados em um ambiente computacional equipado com processador Intel i5 de 12ª geração, 16GB de memória RAM e uma placa gráfica NVIDIA GeForce RTX 3060 com 12GB de VRAM dedicada. Toda a implementação do pipeline de treinamento foi desenvolvida em linguagem Python (v. 3.11.7), utilizando-se o arcabouço (*framework*) PyTorch.

3.3. Experimentos

Considerando a limitação de dados em *datasets* de imagens do TME de CG, foi adotada a estratégia de transferência de aprendizado (*Transfer Learning* — TL), inicializando as redes com pesos pré-treinados no *ImageNet* [Deng et al. 2009]. A partir dessa base, os experimentos foram estruturados para avaliar arquiteturas representativas e, posteriormente, utilizar suas predições na construção dos comitês.

Foram avaliadas oito arquiteturas pré-treinadas frequentemente empregadas em estudos com DL em imagens do TME: EfficientNet-B0 (*efn*) e EfficientNet-B2 (*efn_b2*) [Tan and Le 2019], DenseNet121 (*dsn121*) e DenseNet201 (*dsn201*) [Huang et al. 2017], MobileNetV2 (*mbn*) [Howard et al. 2019], ResNet18 (*rsn18*) e ResNet101 (*rsn101*) [He et al. 2016], além do *Vision Transformer* (*vit*) [Dosovitskiy et al. 2021].

Os experimentos foram conduzidos sob duas configurações de treinamento:

- **Cenário 1 — Treinamento sem aumento de dados:** os modelos foram treinados apenas com as imagens originais, sem aplicação de transformações de *data augmentation*.
- **Cenário 2 — Treinamento com aumento de dados:** aplicou-se *data augmentation* durante o treinamento, como estratégia de regularização para aumentar a variabilidade das amostras e favorecer a generalização [Mumuni and Mumuni 2022].

O aumento de dados foi implementado usando a biblioteca Albumentations [Iglovikov et al. 2025], em esquema *on-the-fly* (tempo de execução), ampliando a diversidade sem gerar novas imagens armazenadas. As transformações empregadas foram: *SquareSymmetry* (rotações de 90°/180°/270° e espelhamentos), *Affine* (translação, rotação, zoom e cisalhamento), *RandomToneCurve* (ajustes de iluminação) e *RandomGridShuffle* (embaralhamento por grade). Todas foram configuradas com probabilidade de 0,3 por imagem e podem ser aplicadas em conjunto.

3.4. Métricas de avaliação

Para avaliar os resultados obtidos nos experimentos descritos na Seção 3, foram empregadas as métricas acurácia, precisão, *recall* e *F1-score*. As definições adotadas são apresentadas nas Equações (1-4). A métrica *F1-score* foi adotada como principal porpor sintetizar o equilíbrio entre precisão e *recall* na tarefa de classificação multiclasse dos componentes do TME.

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \quad (1)$$

$$\text{Precisão} = \frac{VP}{VP + FP} \quad (2)$$

$$\text{Recall} = \frac{VP}{VP + FN} \quad (3)$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precisão} \cdot \text{Recall}}{\text{Precisão} + \text{Recall}} \quad (4)$$

onde VP, VN, FP e FN representam, respectivamente, Verdadeiros Positivos, Verdadeiros Negativos, Falsos Positivos e Falsos Negativos.

3.5. Formação de Comitês

Um comitê de classificadores é uma estratégia bastante comum em problemas de classificação baseados em algoritmos de ML. Sua ideia central é combinar um conjunto finito de modelos preditivos (classificadores base) para classificar novas amostras, melhorando a acurácia e a capacidade de generalização em relação aos classificadores individuais [Dietterich 2000].

A eficácia de um *ensemble* está ligada a dois fatores principais: a diversidade entre os classificadores e a estratégia de combinação das predições individuais, de modo que o *ensemble* apresente desempenho superior ao de qualquer modelo individual [Guehria et al. 2023]. Dentre as estratégias de combinação, destacam-se a média simples, a média ponderada (por desempenho), a votação majoritária, bem como a estratégia de *stacking*. Neste último caso, as predições dos modelos base formam um conjunto de treino para um meta-modelo, que aprende pesos e interações, decidindo quando e em quais modelos confiar, conforme ilustrado na Figura 2

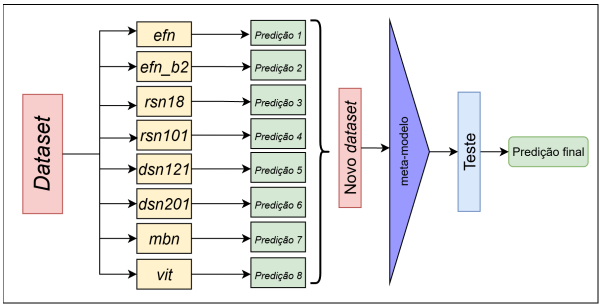


Figura 2. Fluxo do *stacking*: as predições dos modelos base compõem o novo conjunto, utilizado para treinar o meta-modelo para em seguida ter a predição final

O meta-modelo adotado no *stacking* foi uma *Multi-Layer Perceptron* (MLP), treinada com os novos dados formados pelas predições dos modelos base, com o objetivo de aprender uma regra de combinação não linear e capturar interações entre os erros e padrões complementares dos classificadores. Na literatura recente a MLP é frequentemente empregada em *stacking* sendo reportado ganhos ao substituir regras fixas como média e votação [Mohammed et al. 2021, Taifa et al. 2024].

4. Resultados e Discussões

Os resultados usados como *baseline* deste trabalho foram obtidos a partir da replicação dos experimentos reportados por Lou *et al.* [Lou et al. 2025]. Embora o artigo forneça os hiperparâmetros necessários à reprodutibilidade, as divisões dos *folds* não foram disponibilizadas.

A Tabela 1 apresenta os resultados do experimento *baseline*. Nota-se que o *F1-score* obtido foi de 75,4% com desvio padrão de 0,8 pontos percentuais. Esses resultados estão associados à forte ambiguidade existente entre alguns componentes, sobretudo aqueles que se confundem com classes tumorais (NOR, STR, TUM), evidenciando o que Mikhailov *et al.* [Mikhailov et al. 2022] já afirmavam sobre as dificuldades de classificação entre elas.

A Tabela 2 apresenta os resultados obtidos no cenário de treinamento sem a técnica de aumento de dados. Nesse contexto, o modelo *rsn101* apresentou os melhores resultados em todas as métricas, assim como alta estabilidade, refletida na baixa variação entre as execuções. Em seguida, *dsn121* e *efn_b2* mantiveram resultados próximos, também com baixa variância entre execuções.

Tabela 1. Baseline - Replicação do trabalho de Lou *et al.* [Lou et al. 2025].

Modelo	Acurácia	F1-*Score*	Precisão	*Recall*
efn_replicação	$73{,}6 \pm 0{,}8$	$73{,}6 \pm 0{,}8$	$75{,}4 \pm 0{,}8$	$73{,}6 \pm 0{,}8$

-Todos os valores estão expressos em porcentagem.

Buscando aumentar a variabilidade das amostras em tempo de treinamento, realizou-se o aumento de dados aos modelos base. Observou-se ganho de desempenho

em todas as arquiteturas, como é possível constatar na Tabela 3 na qual o teste estatístico compara, para cada modelo, as condições com e sem aumento de dados. Todos os *p-values* ficaram com valor inferior a 0,05, descartando, assim, uma melhora por aleatoriedade.

Como teste estatístico, utilizou-se o teste de Wilcoxon ([Demšar 2006]), aplicado ao *F1-score*, para comparar com os resultados do experimento *baseline*. Os principais achados mostram que os modelos *rsn101*, *dsn121*, *efn_b2* e *efn* apresentaram diferenças estatisticamente significativas em relação ao *baseline* $(p = 0,008)$, sugerindo um ganho consistente entre *folds*. Em contraste, *rsn18*, *dsn201*, *mbn* e *vit* não exibiram evidências de diferença ($p \geq 0{,}312$), indicando desempenho semelhante ao *baseline* dentro da variabilidade amostral. O *vit* se destaca negativamente por ser o único que piora em 5 *folds*.

Tabela 2. Resultados comparativos dos treinamentos sem aumento de dados em relação ao *baseline*.

Modelo	Acurácia	F1-*Score*	Precisão	*Recall*	*p*-value	Δ*F1-score*	Ganho (folds)
rsn101	**75,9 ± 0,4**	**75,9 ± 0,4**	**76,0 ± 0,4**	**75,9 ± 0,4**	0.008	+2,25	8/8
dsn121	75,2 ± 0,4	75,2 ± 0,4	75,4 ± 0,4	75,2 ± 0,4	0.008	+1,56	8/8
efn_b2	74,9 ± 0,3	74,8 ± 0,3	75,1 ± 0,3	74,9 ± 0,3	0.008	+1,23	8/8
efn	74,6 ± 0,4	74,6 ± 0,4	74,7 ± 0,4	74,6 ± 0,4	0.015	+0,96	7/8
rsn18	73,9 ± 0,3	73,9 ± 0,3	74,0 ± 0,3	73,9 ± 0,3	0.382	+0,28	5/8
dsn201	74,0 ± 0,8	73,9 ± 0,7	74,7 ± 0,6	74,0 ± 0,8	0.640	+0,29	5/8
mbn	73,8 ± 0,4	73,8 ± 0,5	74,0 ± 0,4	73,8 ± 0,4	0.843	+0,15	5/8
vit	73,3 ± 0,7	73,2 ± 0,8	74,8 ± 0,8	73,3 ± 0,7	0.312	-0,44	3/8

Acurácia, *F1-score*, Precisão e *Recall* estão expressos em porcentagem. O *p-value* refere-se ao teste de Wilcoxon pareado (F1-*Score*), Δ*F1-score* corresponde ao ganho médio comparado ao *baseline* e Ganho (folds) indica em quantos dos 8 *folds* houve melhora.

A Tabela 4 destaca da Tabela 3 as métricas obtidas com os modelos que, no experimento sem aumento de dados, não apresentaram melhora estatística em relação ao baseline.Ao considerar o treinamento com aumento de dados, observa-se que *rsn18*, *dsn201* e *vit* passaram a apresentar melhora estatisticamente significativa. Em contraste, *mbn* indicou apenas uma tendência de melhora, sem atingir o nível de significância de 0,05.

Com esses achados, foram construídos *ensembles* por empilhamento das predições do conjunto de validação dos modelos treinados anteriormente, organizados em quatro estratégias: i) empilhando apenas as predições dos experimentos sem aumento de dados (*stacking_simples*); ii) empilhando as predições dos experimentos com aumento de dados (*stacking_aumento*); iii) empilhando todas as predições (*stacking_total*); iv) empilhamento apenas das predições provenientes de modelos que apresentaram diferença estatisticamente significativa em relação ao *baseline*, com *p-value* inferiores a 0,05 (*stacking_ajustado*). A Tabela 5 mostra os resultados das métricas de cada estratégia, com destaque para o *stacking_ajustado*, com melhor *F1-score*. Esses achados corroboram com um dos princípios do *ensemble*, que é a escolha dos membros, mostrando que a escolha por melhora estatística apresenta bons resultados.

Tabela 3. Resultados comparativos com aumento de dados em relação ao mesmo modelo sem aumento de dados.

Modelo	Acurácia	F1-*Score*	Precisão	*Recall*	*p*-value	ΔF1	Ganho (folds)
rsn101	77,5 ± 0,2	77,5 ± 0,2	77,9 ± 0,2	77,5 ± 0,2	0.008	+2,56	8/8
dsn121	77,5 ± 0,4	77,5 ± 0,4	77,8 ± 0,4	77,5 ± 0,4	0.008	+1,66	8/8
efn_b2	76,6 ± 0,6	76,6 ± 0,6	76,8 ± 0,6	76,6 ± 0,6	0.008	+3,07	8/8
efn	77,1 ± 0,4	77,2 ± 0,4	77,4 ± 0,3	77,1 ± 0,4	0.008	+1,21	8/8
rsn18	76,4 ± 0,5	76,4 ± 0,5	76,8 ± 0,6	76,4 ± 0,5	0.008	+1,76	8/8
dsn201	76,9 ± 0,5	77,0 ± 0,4	77,3 ± 0,4	76,9 ± 0,5	0.008	+3.1	8/8
mbn	74,9 ± 0,7	75,0 ± 0,7	75,7 ± 0,6	74,9 ± 0,7	0.008	+1,35	7/8
vit	74,5 ± 0,5	74,5 ± 0,5	75,0 ± 0,5	74,5 ± 0,5	0.008	+1,76	8/8

Acurácia, *F1-score*, Precisão e *Recall* estão expressos em porcentagem.

Aplicou-se o teste de Friedman ([Demšar 2006]) para comparar o desempenho entre os quatro *ensembles*. O resultado apontou diferença global estatisticamente significativa entre eles, indicando evidência de que nem todos os métodos apresentam desempenho equivalente e, portanto, que há diferença entre pelo menos dois *ensembles* ($\chi^2 = 411.60$, $p = 6.81 \times 10^{-89}$). Nesse contexto, χ^2 quantifica o quanto os *rankings* observados entre os modelos se afastam do que seria esperado caso todos tivessem desempenho semelhante (quanto maior esse valor, maior a evidência de diferença global), enquanto p é a probabilidade de se obter uma discrepância igual ou maior que a observada, assumindo verdadeira a hipótese nula de igualdade entre todos os *ensembles*, de modo que um valor tão baixo de p sustenta a rejeição dessa hipótese e confirma a existência de diferença significativa em pelo menos uma comparação entre pares.

Tabela 4. Resultados comparativos dos modelos que não superaram significativamente o *baseline* apos aumento de dados.

Modelo	Acurácia	F1-*Score*	Precisão	*Recall*	*p*-value
rsn18	76,4 ± 0,5	76,4 ± 0,5	76,8 ± 0,6	76,4 ± 0,5	0.008
dsn201	76,9 ± 0,5	77,0 ± 0,4	77,3 ± 0,4	76,9 ± 0,5	0.008
mbn	74,9 ± 0,7	75,0 ± 0,7	75,7 ± 0,6	74,9 ± 0,7	0.054
vit	74,5 ± 0,5	74,5 ± 0,5	75,0 ± 0,5	74,5 ± 0,5	0.015

Acurácia, *F1-score*, Precisão e *Recall* estão expressos em porcentagem.

Tabela 5. Resultado das métricas dos ensemble em diferentes etratégias.

Modelo	Acurácia	F1-*Score*	Precisão	*Recall*
stacking_simples	80,4	80,3	80,5	80,4
stacking_aumento	80,3	80,2	80,3	80,3
stacking_total	80,6	80,5	80,8	80,6
stacking_ajustado	81,3	81,3	81,4	81,3

-Todos os valores estão expressos em porcentagem.

A Tabela 6 apresenta o pós-teste de Wilcoxon com correção de Holm. Neste

contexto, p de Holm corresponde ao *p-value* ajustado para múltiplas comparações e foi adotado como critério principal para determinar a significância estatística entre pares de *ensembles*. Além disso, a NLL (*Negative Log-Likelihood*), métrica que quantifica o erro probabilístico do modelo e foi utilizada para calcular $\Delta\mathrm{NLL}$, que indica a direção e a magnitude da diferença observada entre cada par (valores negativos favorecem o *ensemble* B).

Tabela 6. Pós-teste pareado (Wilcoxon em NLL) com correção de Holm, entre os ensembles apresentados

Comparação *ensemble A x B*	ΔNLL	p ajustado por Holm
stacking_simples vs stacking_aumento	-0.025	0.180
stacking_simples vs stacking_total	-0.018	1.57×10^{-8}
stacking_simples vs stacking_ajustado	-0.027	2.53×10^{-16}
stacking_aumento vs stacking_total	$+0.007$	8.04×10^{-7}
stacking_aumento vs stacking_ajustado	-0.001	8.26×10^{-18}
stacking_total vs stacking_ajustado	-0.008	0.007

$\Delta\mathrm{NLL} = \mathrm{NLL}_B - \mathrm{NLL}_A$; valores negativos favorecem B (menor NLL).

Podemos constatar, após esse teste, que o *stacking_ajustado* superou os demais, tendo em vista que sua *NLL* é maior que a de seus adversários e apresenta p ajustado menor que 0,05 em todas as comparações. Outra análise importante é o fato de não haver melhora significante entre usar apenas predições com aumento de dados ou sem aumento de dados; porém, ao usar as duas juntas, por produzir mais dados treináveis, apresenta melhores resultados.

5. Conclusão

Neste trabalho, foi desenvolvido e avaliado um método de classificação de imagens dos diferentes tipos de tecidos do TME de CC, usando *stacking ensembles*. Inicialmente, reproduziu-se um cenário de referência baseado nas condições reportadas por [Lou et al. 2025], estabelecendo-se um *baseline* comparável para as análises subsequentes. Em seguida, avaliou-se um conjunto de arquiteturas pré-treinadas que apresentaram ganhos consistentes, em especial *rsn101*, *dsn121* e *efn_b2*, com evidências de diferença estatisticamente significativa pelo teste de Wilcoxon pareado sobre o *F1-score*.

Ao incorporar aumento de dados *on-the-fly* via Albumentations, verificaram-se melhoras generalizadas no desempenho das arquiteturas em relação à sua versão sem o aumento, reforçando o papel do *data augmentation* em cenários balanceados, favorecendo a variabilidade das amostras. Essa técnica, fundamentada pelos testes estatísticos, serviu de validação para as escolhas dos membros para nosso *ensemble*, empilhando, assim, os dados que favorecem o meta-modelo a entender os padrões associativos das predições geradas pelas redes.

Desta forma, conclui-se que o uso de *stacking ensemble* favorece a generalização do modelo, permitindo combinar diferentes perspectivas de treinamento entre os modelos distintos. Essa estratégia possibilita a exploração complementar de diferentes pontos for-

tes das arquiteturas envolvidas, contribuindo, assim, para o desempenho global do modelo proposto.

Apesar dos resultados promissores, este estudo apresenta a limitação de ter usado apenas uma base de dados para CG, o que restringe a avaliação da capacidade de generalização do modelo proposto. Contudo, como trabalhos futuros, recomenda-se: (i) validar o comitê em *datasets* multi-institucionais de câncer gástrico; (ii) realizar um estudo de explicabilidade do meta-modelo proposto; e (iii) aplicar o mesmo desenho metodológico a outras patologias em patologia digital.

Referências

de Visser, K. E. and Joyce, J. A. (2023). The evolving tumor microenvironment: From cancer initiation to metastatic outgrowth. *Cancer Cell*, 41(3):374–403.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15. Springer Berlin Heidelberg, Berlin, Heidelberg.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, Austria. Disponível em: `https://arxiv.org/abs/2010.11929`.

Ferlay, J., Ervik, M., Lam, F., Laversanne, M., Colombet, M., Mery, L., Piñeros, M., Znaor, A., Soerjomataram, I., and Bray, F. (2024). Global cancer observatory: Cancer today. Site institucional da International Agency for Research on Cancer.

Guehria, S., Belleili, H., and Azizi, N. (2023). A survey on ensemble multi-label classifiers. In Abraham, A. et al., editors, *Proceedings of the 14th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2022)*, volume 648 of *Lecture Notes in Networks and Systems*, pages 100–109. Springer Nature Switzerland, Cham.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., and Adam, H. (2019). Searching for mobilenetv3. *arXiv preprint arXiv:1905.02244*.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708.

Iglovikov, V., Druzhinin, M., Buslaev, A., et al. (2025). Albumentations documentation. `https://albumentations.ai/docs/`. Acesso em: 12 ago. 2025.

Kablan, R., Miller, H. A., Suliman, S., and Frieboes, H. B. (2023). Evaluation of stacked ensemble model performance to predict clinical outcomes: A covid-19 study. *International Journal of Medical Informatics*, 175:105090.

Kather, J. N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.-A., and et al. (2019). Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Medicine*, 16(1):e1002730.

Lou, S., Ji, J., Li, H., and et al. (2025). A large histological images dataset of gastric cancer with tumour microenvironment annotation for ai. *Scientific Data*, 12:138.

Lou, S., Ji, J., Li, H., Zhang, X., Jiang, Y., Hua, M., Chen, K., Ge, K., Zhang, Q., Wang, L., Han, P., and Cao, L. (2024). Gastric cancer histopathology tissue image dataset (gchtid). Dataset.

Mandal, S., Baker, A.-M., Graham, T. A., and Bräutigam, K. (2025). The tumour histopathology "glossary" for ai developers. *PLOS Computational Biology*, 21(1):e1012708.

Mikhailov, I., Khvostikov, A., and Krylov, A. (2022). Methodical approaches to annotation and labeling of histological images in order to automatically detect the layers of the stomach wall and the depth of invasion of gastric cancer. *Archive of Pathology = Arkhiv patologii*, 84(6):67–73. In Russian.

Mohammed, M., Mwambi, H., Mboya, I. B., Elbashir, M. K., and Omolo, B. (2021). A stacking ensemble deep learning approach to cancer type classification based on tcga data. *Scientific Reports*, 11(1):15626.

Mumuni, A. and Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. *Array*, 16:100258.

Naimi, A. I. and Balzer, L. B. (2018). Stacked generalization: An introduction to super learning. *European Journal of Epidemiology*, 33(5):459–464.

Rong, R., Sheng, H., Jin, K. W., Wu, F., Luo, D., Wen, Z., Tang, C., Yang, D. M., Jia, L., Amgad, M., Cooper, L. A. D., Xie, Y., Zhan, X., Wang, S., and Xiao, G. (2022). A deep learning approach for histology-based nuclei segmentation and tumor microenvironment characterization. *bioRxiv*. Preprint.

Su, X., Mao, Q., Wu, Z., et al. (2025). Large language models driven neural architecture search for universal and lightweight disease diagnosis on histopathology slide images. *npj Digital Medicine*, 8:682.

Taifa, I. A., Setu, D. M., Islam, T., Dey, S. K., and Rahman, T. (2024). A hybrid approach with customized machine learning classifiers and multiple feature extractors for enhancing diabetic retinopathy detection. *Healthcare Analytics*, 5:100346.

Tan, M. and Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*. Disponível em: `https://arxiv.org/abs/1905.11946`.

Wang, Z., Peng, H., Wan, J., and et al. (2024). Identification of histopathological classification and establishment of prognostic indicators of gastric adenocarcinoma based on deep learning algorithm. *Medical Molecular Morphology*, 57:286–298.