

FUNDAÇÃO ANTÔNIO PRUDENTE
A.C.CAMARGO CANCER CENTER

ADRIANA PASSOS BUENO

AVALIAÇÃO DA REGRESSÃO TUMORAL POR
APRENDIZADO PROFUNDO EM
ADENOCARCINOMAS GÁSTRICOS: IMPLICAÇÕES
NA REPRODUTIBILIDADE E PROGNÓSTICO
PÓS-NEOAJUVÂNCIA

Tese de Doutorado, apresentada à Fundação Antônio Prudente como requisito para a obtenção do título de Doutor em Ciências. Área de concentração: Oncologia

Orientador: Dr. Israel Tojal da Silva

Coorientador: Dr. Felipe José Fernández Coimbra

São Paulo
2025

Bueno, Adriana .

Avaliação da regressão tumoral por aprendizado profundo em adenocarcinomas gástricos: implicações na reprodutibilidade e prognóstico pós-neoadjuvância. / Adriana Bueno. São Paulo, 2025.

91f.

Tese de Doutorado - Fundação Antônio Prudente. Curso de Pós-Graduação em Ciências - Área de concentração: Oncologia.

Orientador: Israel Tojal da Silva.

1. Adenocarcinoma Trato Gastrointestinal, 2. Escore de Regressão Tumoral, 3. Inteligência Artificial

CDU 616

Nome: Adriana Passos Bueno

Título: avaliação da regressão tumoral por aprendizado profundo em adenocarcinomas gástricos: implicações na reprodutibilidade e prognóstico pós-neoadjuvância

Aprovado em: 18/12/2025

Banca Examinadora

Orientador: Israel Tojal da Silva

Instituição: Fundação Antônio Prudente

Membro da banca: Dr. Clovis Antonio Lopes Pinto

Instituição: Fundação Antônio Prudente

Membro da banca: Dr. Paulo Kassab

Instituição: Santa Casa de Misericórdia de São Paulo, Departamento de Cirurgia.

Membro da banca: Dr. Zhao Liang

Instituição: Universidade de São Paulo, Faculdade de Filosofia Ciências e Letras de Ribeirão Preto

Membro da banca: Dr. Joaquim Cezar Felipe

Instituição: Universidade de São Paulo, Faculdade de Filosofia Ciências e Letras de Ribeirão Preto, Departamento de Computação e Matemática

*A ciência não pode prever o que vai acontecer.
Só pode prever a probabilidade de algo acontecer.*
César Lattes

*Este trabalho é dedicado às crianças adultas que,
quando pequenas, sonharam em se tornar cientistas.*

AGRADECIMENTOS

Agradeço aos médicos patologistas do Departamento de Anatomia Patológica do AC Camargo Cancer Center, Dr. Gabriel Oliveira dos Santos e Dr. Warley de Abreu Nunes, assim como à Dra. Mariangela Ottoboni Brunaldi, da Faculdade de Medicina de Ribeirão Preto, pela valiosa contribuição a este estudo, trazendo o olhar especializado fundamental tanto para as análises conduzidas quanto para o processo de aprendizado ao longo do trabalho.

Agradeço também a Luan Vinicius de Carvalho Martins, Renan Valieris Bueno de Almeida e Alexandre Defelicibus pelo suporte constante, pelas orientações e pelos conhecimentos compartilhados, que foram essenciais para o desenvolvimento desta pesquisa.

Agradeço também ao meu coorientador, Dr. Felipe José Fernández Coimbra, cuja ampla experiência em oncologia, com ênfase no aparelho digestivo alto, contribuiu de maneira decisiva para o aprofundamento da minha visão interdisciplinar. Suas orientações foram fundamentais ao longo de todo o desenvolvimento deste projeto.

Por fim, expresso meu profundo agradecimento ao meu orientador, Dr. Israel Tojal da Silva, pela orientação dedicada, pelos ensinamentos e pela contribuição decisiva que possibilitaram não apenas a realização deste projeto, mas também um importante avanço na minha formação acadêmica e profissional.

RESUMO

Bueno AP. **Avaliação da regressão tumoral por aprendizado profundo em adenocarcinomas gástricos: implicações na reprodutibilidade e prognóstico pós-neoadjuvância.** [Tese de Doutorado]. São Paulo; Fundação Antônio Prudente; 2025.

A avaliação da resposta patológica à terapia neoadjuvante em neoplasias do estômago é crucial para o manejo clínico, sendo geralmente realizada através do Escore de Regressão Tumoral (**ERT**), um método que, no entanto, apresenta desafios significativos de subjetividade e baixa reprodutibilidade interobservador. Com o objetivo de mitigar essas limitações e aprimorar a predição prognóstica, este estudo propôs o desenvolvimento de um modelo de Patologia Computacional (**M1**) baseado em Redes Neurais em Grafos (**GNNs**) associadas a Modelos de Fundação (**MFs**) para análise de imagens histológicas digitais (WSIs). A metodologia utilizou uma coorte de adenocarcinomas gástricos e colorretais de duas instituições, treinando o modelo para identificar 15 categorias histológicas e testando combinações de arquiteturas **GNN** (como **GIN** e **ARMA**) com extratores de *features* (Virchow 2 e UNI 2). Os resultados indicaram que o modelo proposto (**M1**) é capaz de identificar as 15 classes histológicas designadas, com alta precisão (F1-Score de até 0,94), sendo a arquitetura **GIN** em conjunto com o Virchow 2 a combinação com o melhor desempenho na coorte de validação interna. O modelo também reproduziu o **ERTG** com forte concordância em relação à avaliação de um patologista especialista ($p=0,88$). Além disso, a análise de sobrevida pós-cirúrgica revelou que a combinação de variáveis clínicas tradicionais com a quantificação automatizada das categorias histológicas pelo **M1** (**'cli + model'**) superou o poder preditivo do **ERT** isolado e do estadiamento TNM isolado. Conclui-se que a aplicação de **GNNs** sensíveis ao contexto topológico oferece uma ferramenta robusta para padronizar a avaliação patológica e refinar significativamente a estratificação prognóstica em oncologia e que, quando associado a outras variáveis do laudo anátomo-patológico é capaz de melhorar predição prognóstica de sobrevida pós cirurgia (**SPC**). Esses achados reforçam a relevância da análise do ecossistema histológico completo para uma estratificação prognóstica mais refinada, embora aspectos como o viés de anotação e a necessidade de maior diversidade de dados indiquem caminhos importantes para trabalhos futuros.

Palavras-chave: Adenocarcinoma Trato Gastrointestinal. Escore de regressão tumoral. Inteligência Artificial.

ABSTRACT

Bueno AP. **Deep learning assessment of tumor regression in gastric adenocarcinomas: implications for reproducibility and prognosis after neoadjuvant therapy.** [Thesis]. São Paulo; Fundação Antonio Prudente; 2025.

The assessment of pathological response to neoadjuvant therapy in stomach neoplasms is crucial for clinical management, typically performed using the Tumor Regression Score (**TRS**), a method that, however, presents significant challenges regarding subjectivity and low interobserver reproducibility. Aiming to mitigate these limitations and enhance prognostic prediction, this study proposed the development of a Computational Pathology model (**M1**) based on Graph Neural Networks (**GNNs**) associated with Foundation Models (**FM**s) for the analysis of digital histological images (Whole Slide Images - **WSIs**). The methodology utilized a cohort of gastric and colorectal adenocarcinomas from two institutions, training the model to identify 15 histological categories and testing combinations of **GNN** architectures (such as **GIN** and **ARMA**) with feature extractors (Virchow 2 and UNI 2). The results indicated that the proposed model (**M1**) is capable of identifying the 15 designated histological classes with high precision (F1-Score up to 0,94), with the **GIN** architecture combined with Virchow 2 being the best-performing combination in the internal validation cohort. The model also reproduced the **ERTG** with strong agreement compared to a specialist pathologist's evaluation ($\rho=0,88$). Furthermore, post-surgical survival analysis revealed that the combination of traditional clinical variables with the automated quantification of histological categories by **M1** ('**cli** + **model**') surpassed the predictive power of the TRS and TNM staging alone. It is concluded that the application of topologically context-aware **GNNs** offers a robust tool to standardize pathological evaluation and significantly refine prognostic stratification in oncology, and when associated with other variables from the pathology report, it is capable of improving the prognostic prediction of post-surgical survival (**SPC**). These findings reinforce the relevance of analyzing the complete histological ecosystem for more refined prognostic stratification, although aspects such as annotation bias and the need for greater data diversity point to important directions for future work.

Keywords: Gastrointestinal Tract Adenocarcinoma. Tumor Regression Score. Artificial Intelligence.

LISTA DE ILUSTRAÇÕES

Figura 1 – Componentes TGI	14
Figura 2 – Taxa de incidência e mortalidade dos cânceres mais incidentes em 2020	15
Figura 3 – Jornada do paciente	16
Figura 4 – Fluxo do material da ressecção cirúrgica	16
Figura 5 – Amostragem macroscópica	17
Figura 6 – Identificação de LT	18
Figura 7 – Escore modificado de <i>Ryan</i>	19
Figura 8 – Estratégias de avaliação do ERT	19
Figura 9 – Estratégias para realização do ERT	19
Figura 10 – Padrões de resposta do LT	20
Figura 11 – Diagrama de Venn IA	28
Figura 12 – Contextualização de <i>patch</i>	30
Figura 13 – Arquitetura em grafos de uma GNN	32
Figura 14 – Evolução visão computacional	34
Figura 15 – Distribuição da coorte em subgrupos	37
Figura 16 – Sistema de apoio de anotação	41
Figura 17 – Biblioteca de <i>tiles</i>	41
Figura 18 – Predição das classes em LT	44
Figura 19 – Testes realizados por subgrupo	52
Figura 20 – Mapa de calor <i>tiles</i> em treino e validação de M1_Dev	54
Figura 21 – Matriz de confusão (M1_Dev - Val)	56
Figura 22 – Matriz de confusão da coorte externa de teste	58
Figura 23 – Metaplasia intestinal em mucosa gástrica	59
Figura 24 – Performance modelos RF para predizer SPC	62
Figura 25 – Importância variáveis pelo método SHAP	64
Figura 26 – Caso de FMRP com área densa e “colagenizada”	68
Figura 27 – Classes mescladas em LT	69
Figura 28 – Diferentes padrões de resposta	69

LISTA DE TABELAS

Tabela 1 – Estadiamento patológico para tumores gástricos - AJCC 8 ^a edição . . .	25
Tabela 2 – Detalhamento dos testes realizados por subgrupo de M1_Dev	44
Tabela 3 – Descrição dos conjuntos de testes para avaliação entre os 5 obeserva- dores com respectivo objetivo	48
Tabela 4 – Descrição das combinações de variáveis para estudo de SPC	50
Tabela 5 – Distribuição da coorte por subgrupos	53
Tabela 6 – Distribuição do número de anotações manuais em cada subgrupo de M1	53
Tabela 7 – Comparação entre desempenho de MF para cada arquitetura de GNN	55
Tabela 8 – Performance do grupo M1_Dev - TestI	56
Tabela 9 – Performance do grupo M1_Dev - TestE. GNN X Virchow 2 para extração de features	57
Tabela 10 – Performance do grupo M1_Dev - TestE - GNN x UNI2	57
Tabela 11 – Concordância das avaliações entre os três patologistas	60
Tabela 12 – Concordância das avaliações entre os diferentes métodos de M1 com os três patologistas	61

LISTA DE SIGLAS E ABREVIATURAS

ACCCC	AC Camargo Cancer Center
CNN	Redes Neurais Convolucionais
ERTG	Escore de Regressão Tumoral Gástrico
EST	Topografias transição esôfago-gástricas e estômago
FM	Foundation Models
GNN	Redes Neurais de Grafos
IA	Inteligência Artificial
LT	Leito Tumoral
PC	Peça Cirúrgica
PC	Patologia Computacional
PD	Patologia Digital
RETO	Topografias transição colorretal e reto
RF	Randon Forest
SPC	Sobrevida Pós-Tratamento Cirúrgico
SSL	Técnicas de Aprendizado Auto-Supervisionado
TGI	Trato Gastrointestinal
WSI	Whole Slide Images

SUMÁRIO

1	INTRODUÇÃO	14
1.1	Neoplasias do Trato Gastrointestinal	14
1.2	Escore de Regressão Tumoral	15
1.2.1	Visão geral	15
1.2.2	Avaliação macroscópica	16
1.2.3	Avaliação microscópica	17
1.2.4	Concordância entre observadores	20
1.2.5	Métricas para avaliação de concordância entre observadores	21
1.2.6	Predição Prognóstica	24
1.2.7	Importância do Estadiamento patológico	24
1.3	Patologia e Visão Computacional	25
1.3.1	Patologia Digital	25
1.3.2	Patologia Computacional	26
1.4	Racional do Estudo e Definição dos Objetivos	34
2	MATERIAIS E MÉTODOS	35
2.1	Definição da coorte do estudo	35
2.1.1	Distribuição da coorte para os diferentes testes	36
2.2	Construção do modelo de GNN M1	37
2.2.1	Definição das categorias histológicas	37
2.2.2	Distribuição de dados para desenvolvimento de M1	40
2.2.3	Treinamento, validação e teste interno de M1	42
2.2.4	Teste Externo de M1	43
2.2.5	Compreensão dos padrões de comportamento de M1	44
2.3	Performance de M1 para reprodução de ERTG	45
2.3.1	Definição do Leito Tumoral (LT)	45
2.3.2	Protocolo para realização do escore de regressão tumoral	46
2.3.3	Método de estimativa dos componentes do LT pelo modelo M1	46
2.3.4	Avaliação da concordância entre observadores	47
2.4	Performance de M1 para predição de sobrevida pós-tratamento cirúrgico	48
2.4.1	Avaliação do papel das variáveis	51
3	RESULTADOS	52
3.1	Definição da coorte do estudo	52
3.1.1	Distribuição da coorte para diferentes testes	52
3.2	M1 identifica categorias histológicas de pacientes com câncer gástrico	53

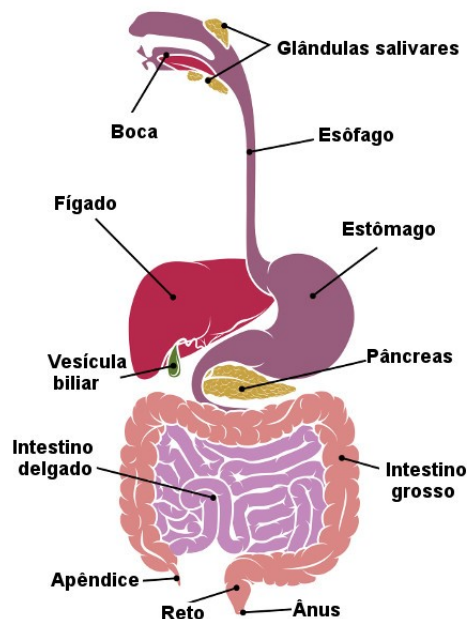
3.2.1	Biblioteca de <i>tiles</i> e distribuição de subgrupos para treino, validação e testes de M1	53
3.2.2	Treinamento, Validação e Teste Interno de M1	55
3.2.3	Teste Externo de M1	57
3.3	M1 é capaz de reproduzir ERTG	60
3.3.1	Concordância entre observadores	60
3.4	M1 com dados anátomo - clínicos melhora o poder preditivo prognóstico . .	62
3.4.1	Avaliação da importância das variáveis	63
4	DISCUSSÃO	66
5	CONCLUSÃO	74
	REFERÊNCIAS	75
	APÊNDICE A – ABSTRACT EXPOSTO NA USCAP	82
	APÊNDICE B – WSI2ML	86

1 INTRODUÇÃO

1.1 NEOPLASIAS DO TRATO GASTROINTESTINAL

O trato gastrointestinal (**TGI**) é composto por órgãos do tubo digestivo, incluindo a orofaringe, o esôfago, o estômago, o intestino delgado, o cólon e o reto (Figura 4). Existe uma grande variedade de neoplasias malignas que acometem as diferentes topografias do trato gastrointestinal, sendo que algumas delas apresentam alta taxa de incidência, morbidade e mortalidade. Segundo os dados do observatório global do câncer GLOBOCAN2020, ao considerar a taxa de incidência global, o câncer colorretal se apresentou como o quarto colocado alcançando uma taxa de incidência de 19,5/100 000 habitantes (Figura 2), ficando atrás do câncer de mama, próstata e pulmão. De maneira similar, o câncer de estômago foi o sexto colocado, com uma taxa de 11,1/100.000 habitantes (The Global Cancer Observatory 2020) (Figura 2). Tanto os tumores colorretais quanto os de estômago, além de serem altamente incidentes, possuem alta taxa de mortalidade, acometendo 6,2/100 000 habitantes e 6,1/100 000 habitantes, respectivamente (The Global Cancer Observatory 2020). Já as neoplasias do esôfago, embora menos comuns, são extremamente letais e responsáveis pela sexta causa de morte por câncer, com uma sobrevida em 5 anos de 15%-20% (Arnal et al. 2015).

Figura 1 – Componentes do trato gastrointestinal



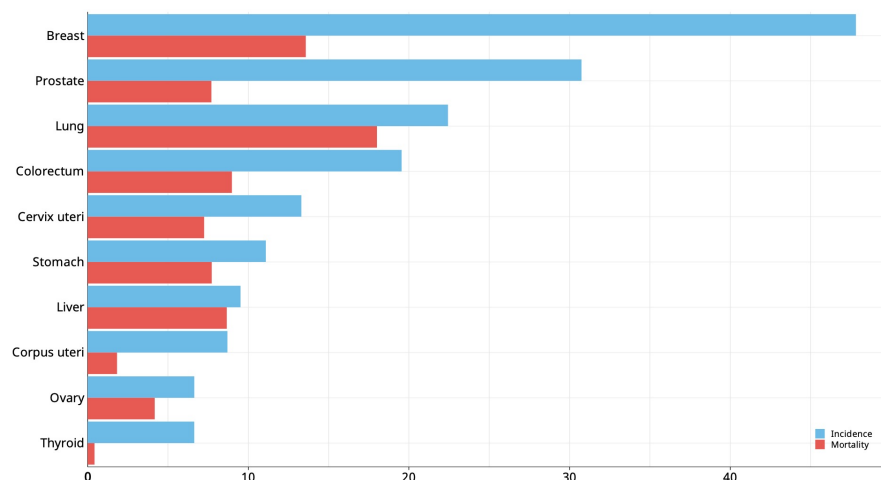
Fonte: Passos (2021).

Legenda: Componentes do trato gastrointestinal e os órgãos associados, como as glândulas salivares, o pâncreas, as vias biliares e o fígado.

O manejo dessas neoplasias varia conforme a topografia, o estadiamento e o tipo

histológico. O uso do tratamento sistêmico pré-cirúrgico, ou neoadjuvante, vem sendo cada vez mais utilizado nos cânceres esofágicos, do estômago e colorretais (Pai e Pai 2018). As vantagens do tratamento neoadjuvante incluem o aumento da chance de cirurgia curativa após a redução do tumor (*dowstaging*), a eliminação de micrometástases e a melhora dos sintomas relacionados aos tumores correspondentes. Adicionalmente, pacientes que recebem tratamento perioperatório têm mais chance de progressão livre de doença (Cunningham et al. 2016).

Figura 2 – Taxa de incidência e mortalidade dos cânceres mais incidentes em 2020



Fonte: The Global Cancer Observatory (2020).

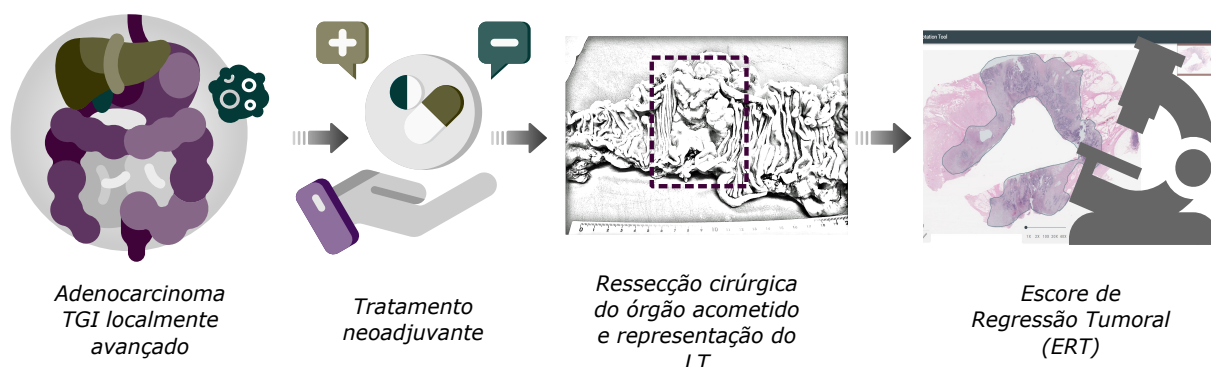
Legenda: Taxa de incidência e mortalidade dos cânceres mais incidentes em 2020.

1.2 ESCORE DE REGRESSÃO TUMORAL

1.2.1 Visão geral

Após a terapia neoadjuvante, a maior parte dos pacientes segue para o tratamento cirúrgico. Em seguida, busca-se avaliar o efeito do tratamento no tumor por meio de uma análise da peça cirúrgica pelo patologista, considerando, para tanto, um escore que quantifica a regressão do tumor. Essa avaliação fornece informações prognósticas e pode orientar a quimioterapia adjuvante (tratamento sistêmico após a cirurgia) (Pai e Pai 2018). A Figura 3 ilustra a jornada do paciente com adenocarcinoma de **TGI** localmente avançado, submetido à terapia neoadjuvante, com ressecção cirúrgica posterior. De importância, com o aumento do uso de terapia neoadjuvante em tumores do trato gastrointestinal, a avaliação histopatológica vem crescendo e constitui um dos principais métodos para a avaliação da regressão tumoral (Pai e Pai 2018). Por outro lado, há algumas dificuldades no uso dos protocolos disponíveis. A Figura ?? ilustra o fluxo do material da ressecção cirúrgica até a confecção do laudo com Escore de Regressão Tumoral (**ERT**).

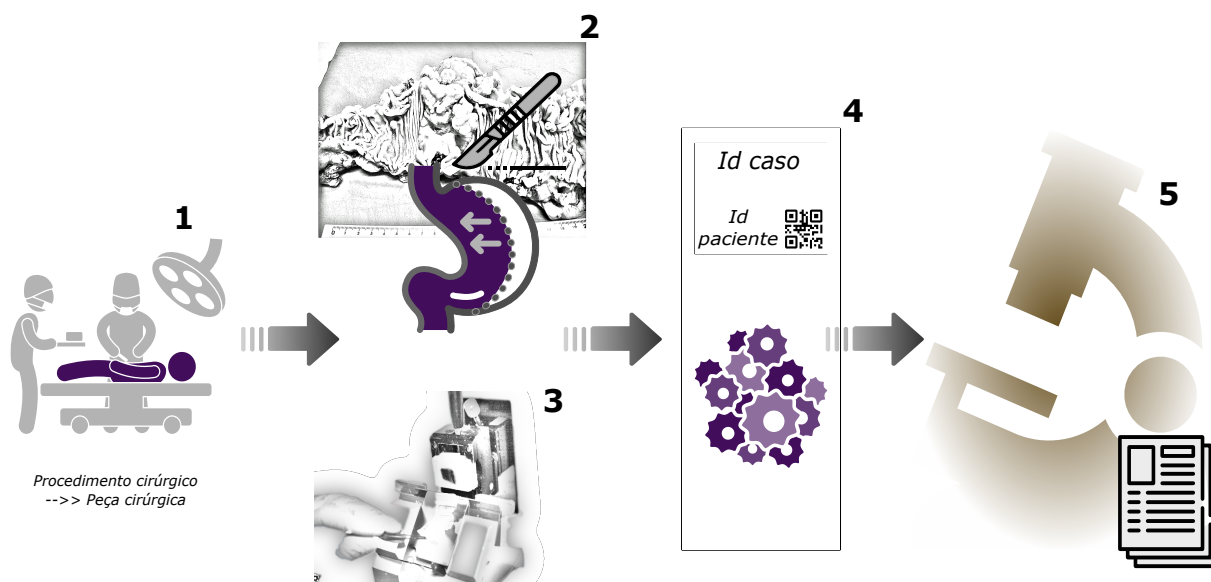
Figura 3 – Jornada do paciente



Fonte: Autoria própria.

Legenda: Jornada do paciente com adenocarcinoma de **TGI** localmente avançado.

Figura 4 – Fluxo do material da ressecção cirúrgica



Fonte: Autoria própria.

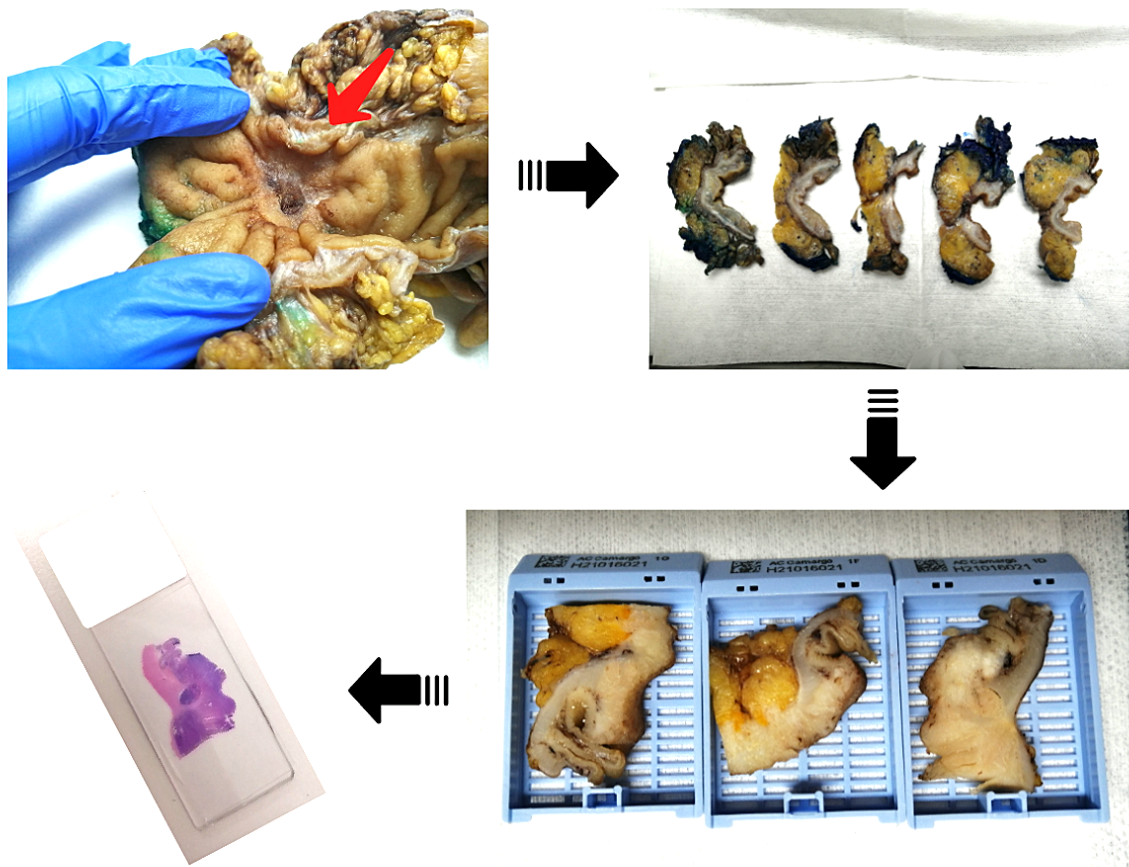
Legenda: Fluxo do material da ressecção cirúrgica até a confecção do laudo com **ERT**. 1- Ressecção cirúrgica do órgão acometido; 2- Macroscopia com identificação e representação do **LT**; 3- Processamento histológico para confecção de lâminas; 4- Lâminas representativas do órgão acometido, incluindo **LT**; 5- Execução de laudo anátomo-patológico com **ERT** a partir da avaliação microscópica das lâminas de microscopia.

1.2.2 Avaliação macroscópica

Atualmente, vários esquemas são utilizados para realizar a avaliação pós-terapia neoadjuvante em tumores do trato gastrointestinal (Pai e Pai 2018). Todos os métodos utilizados para a realização do **ERT** baseiam-se na avaliação dos componentes do Leito Tumoral (**LT**). Essa avaliação tem início no processo de macroscopia, no qual o **LT** do órgão ressecado será identificado, mensurado e representado conforme os protocolos de

macroscopia para pacientes pós-terapia neoadjuvante (Amin et al. 2017; Nagtegaal et al. 2002; Attaway e Fortuna 2024). Apesar do método, protocolos de macroscopia enfatizam que o **LT** deve ser amostrado extensamente e incluído em sua totalidade para a realização adequada do **ERT**. Consequentemente, a quantidade de lâminas necessárias para representar um **LT** está diretamente relacionada ao tamanho macroscópico do **LT**. A Figura 5 exemplifica um processo de macroscopia de peças de **TGI** pós-tratamento neoadjuvante.

Figura 5 – Amostragem macroscópica



Fonte: Autoria própria.

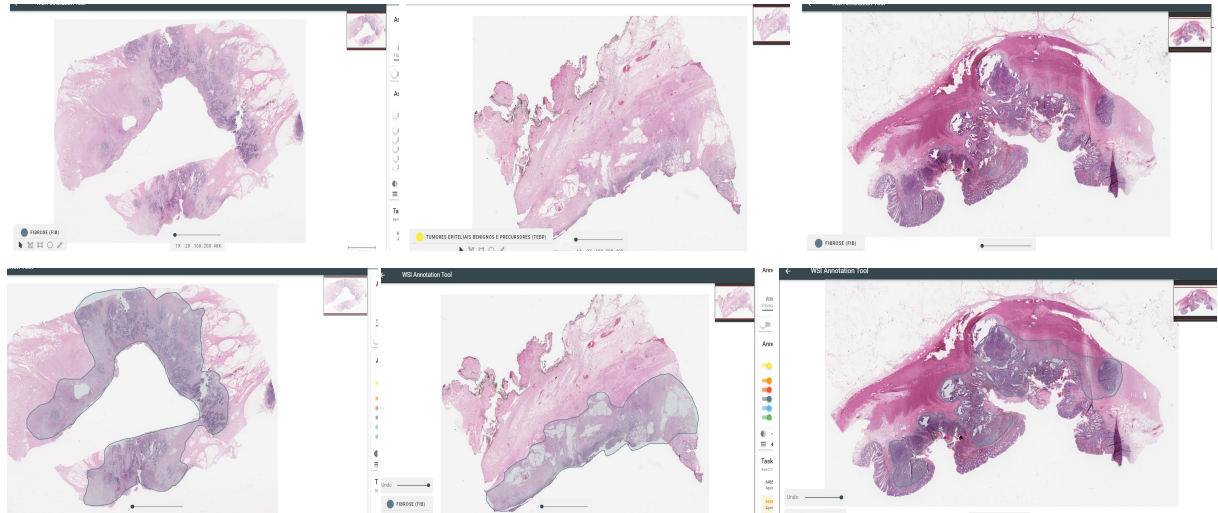
Legenda: Processo de amostragem macroscópica para ressecções cirúrgicas de órgãos do **TGI** após tratamento neoadjuvante.

1.2.3 Avaliação microscópica

Assim como na macroscopia, é necessária a identificação do **LT** nas lâminas de microscopia para a realização do **ERT**. Dentro da mesma lâmina, há um subconjunto que representa a área de **LT** (Figura 6), composta por tumor residual e por áreas que exibem processos reparativos e regenerativos. Esses processos refletem alterações histológicas que substituem a neoplasia invasiva destruída em resposta à ação citotóxica dos tratamentos

neoadjuvantes. Para a avaliação microscópica a partir de lâminas, existem duas estratégias principais estabelecidas para a análise histopatológica e para a realização do **ERT**.

Figura 6 – Identificação de **LT**



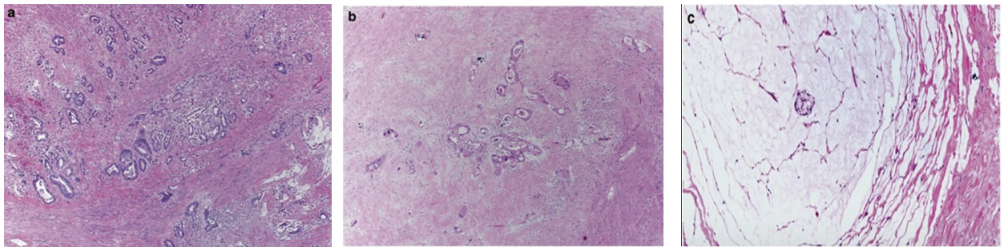
Fonte: Autoria própria.

Legenda: Identificação de **LT** em lâminas de microscopia de neoplasias do **TGI**. As fotos na linha superior mostram as lâminas sem nenhuma marcação. Na linha inferior há a identificação do **LT** ressaltada em cinza.

Na primeira estratégia, propôs-se um esquema que quantifica a proporção de fibrose existente, entendida como área de resposta, e de células tumorais viáveis. Dentro desse contexto, encontram-se os escores propostos por *Mandard*(Mandard et al. 1994), *Dworak*(Dworak et al. 1997), *Ryan*(Ryan et al. 2005) e pelo **CAP** (*College of American Pathologists*) (College of American Pathologists 2023a; College of American Pathologists 2023b; College of American Pathologists 2023c). Todos esses esquemas utilizam termos descritivos e subjetivos para relatar a relação entre fibrose e células tumorais viáveis, por exemplo: 'fibrose com apenas células tumorais esparsas' ou 'fibrose supera a quantidade de tumor residual' (Figura 7). Portanto, a ausência de terminologia específica e de uma medida quantitativa precisa torna a implementação desses esquemas desafiadora, resultando em baixa reprodutibilidade interobservador, além de dificultar a comparação entre os estudos. Outro problema dessa abordagem de avaliação diz respeito à dificuldade de diferenciar a fibrose induzida pelo tratamento da desmoplasia tumoral, um efeito estromal causado pelo tumor (Pai e Pai 2018).

Uma segunda estratégia consiste na quantificação do tumor residual em relação ao leito tumoral original, que corresponde à área que o tumor ocupava antes do tratamento neoadjuvante. Nesse contexto, se encaixam os esquemas de *Becker*(Becker et al. 2003), *Rodel*(Rödel et al. 2005) e o esquema modificado para câncer retal(Bateman et al. 2009). Em tais esquemas, é essencial determinar a extensão do leito tumoral, o que pode não ser viável na prática quando não há alterações macroscópicas na peça cirúrgica (Figura 9).

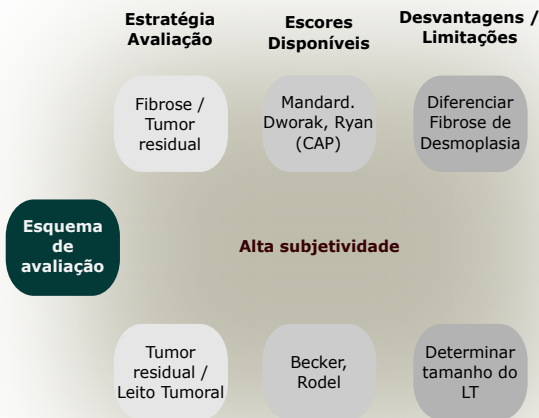
Figura 7 – Escore modificado de *Ryan*



Fonte: (Pai e Pai 2018).

Legenda: Exemplos do escore modificado de *Ryan* proposto pelo **CAP**. a) neoplasia residual extensa sem evidência de regressão tumoral - escore 3; b) neoplasia residual com regressão tumoral evidente, com predomínio de fibrose - escore 2; e c) pequeno grupo de células isoladas residuais - escore 1.

Figura 9 – Estratégias para realização do **ERT**.



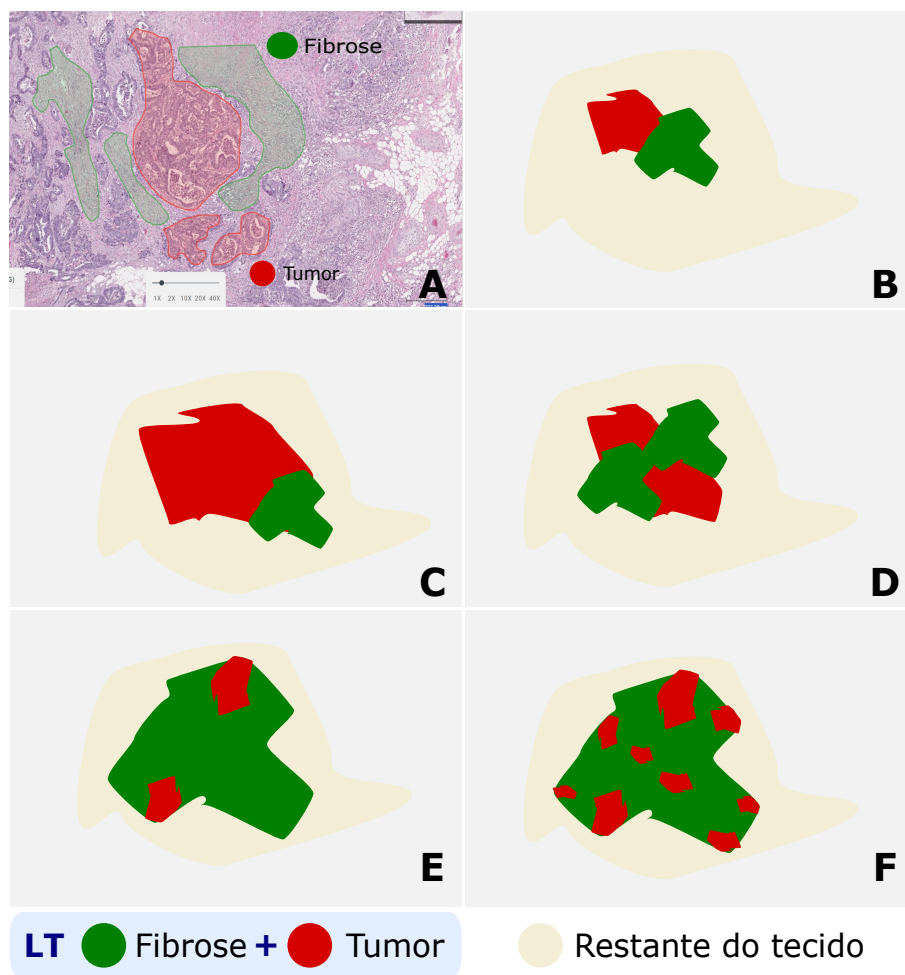
Fonte: Autoria própria.

Legenda: diferentes estratégias para avaliação do **ERT**. Na linha superior, esquemas que avaliam fibrose *versus* tumor residual no **LT**; na linha de baixo, estratégias que avaliam tumor residual frente ao **LT** macroscópico.

1.2.4 Concordância entre observadores

As duas estratégias apresentadas demonstram suas respectivas vantagens e desvantagens. Entretanto, independentemente do esquema de avaliação do **LT**, o padrão de resposta da neoplasia invasiva ao tratamento pode, por si só, representar um desafio significativo. É mais fácil estimar, a olho nu, os componentes do **LT** em uma situação de pouca resposta ao tratamento, na qual há um predomínio significativo do tumor invasivo, resultando em um bloco único, do que em um padrão de resposta em que o tumor residual se apresenta fragmentado, disperso e entremeado por áreas de resposta ao tratamento e fibrose (Figura 10).

Figura 10 – Padrões de resposta do **LT**



Fonte: Autoria própria.

Legenda: esquema representando possíveis padrões de resposta em um **LT**. A - Corte histológico ilustrando componentes de Tumor e Fibrose em um **LT**; B - Esquema representando **LT** com aproximadamente 50% de cada componente; C a F - Esquemas ilustrando disposições de **LT** gradativamente mais desafiadoras para a estimativa do **ERT**.

Embora o escore de regressão seja um método que auxilia no manejo de pacientes no pós-operatório, a concordância entre os avaliadores de um mesmo conjunto de amostras permanece desafiadora. Por exemplo, foi demonstrado que não houve consenso entre 17 patologistas quanto à aplicação dos sistemas de graduação de *Mandard* e *Dworak*, bem como do esquema modificado para câncer retal (Chetty et al. 2012). Nesse estudo, observou-se muito pouca concordância entre os três sistemas de classificação, com valores de κ de 0,28, 0,35 e 0,38 nos escores de *Mandard*, *Dworak* e do câncer retal modificado, respectivamente. Entre os escores citados acima, o de *Ryan*, originalmente criado para a avaliação de tumores retais, apresenta-se como o de maior reprodutibilidade. Dessa forma, o **CAP** sugere o uso do Escore de Regressão Tumoral de *Ryan* Modificado (**ERTRM**) para esôfago e estômago (Mandard et al. 1994; Dworak et al. 1997; Ryan et al. 2005).

1.2.5 Métricas para avaliação de concordância entre observadores

Diante dos desafios mencionados, um aspecto fundamental refere-se à concordância entre observadores na avaliação e na graduação do efeito do tratamento nos tumores do trato gastrointestinal. Existem muitas métricas disponíveis para avaliar a consistência das medidas entre diferentes observadores, algumas aplicáveis a variáveis contínuas e outras a variáveis discretas. Em seguida, há uma breve explicação das variáveis disponíveis:

Coeficiente *Kappa* (κ)

Também conhecida como *Kappa* de *Cohen*, essa estatística é a principal utilizada para avaliar a concordância interobservador em dados nominais ou categóricos entre dois observadores. O *Kappa* (κ) mede a concordância observada entre os observadores, subtraindo a concordância esperada pelo mero acaso.

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Em que:

- P_o = Proporção Observada de Concordância (a concordância real).
- P_e = Proporção Esperada de Concordância por Acaso.

Seus valores podem variar de 0 até 1, com a seguinte interpretação:

- $\kappa = 1$: Concordância perfeita.
- Classificação (Escala de Landis & Koch):
 - 0.00: Pobre

- 0.00 – 0.20: Leve
 - 0.21 – 0.40: Razoável
 - 0.41 – 0.60: Moderada
 - 0.61 – 0.80: Substancial (Boa)
 - 0.81 – 1.00: Quase Perfeita
- $\kappa = 0$: A concordância é apenas o que seria esperado pelo acaso.

Escore de Kendall

O Escore de *Kendall* ou Concordância de Rank, como o Tau (τ), é uma estatística não paramétrica usada principalmente para dados ordinais (rankings). O Kendall's Tau (τ) mede a força da associação (correlação) entre os rankings de dois observadores, indicando a probabilidade de que ambos classifiquem um par de sujeitos na mesma ordem. Varia de -1 a $+1$, onde $+1$ indica concordância perfeita no ranking e 0 indica que os rankings são independentes.

Correlação

A correlação pode ser utilizada em variáveis contínuas para medir a força e a direção da relação linear entre duas variáveis. O papel principal da correlação é determinar:

1. Associação Linear: Se uma variável tende a mudar em proporção constante com a outra;
2. Direção: Se a relação é positiva (ambas aumentam) ou negativa (uma aumenta e a outra diminui);
3. Força: Quão próximos os dados estão de formar uma linha reta (quão consistente é o relacionamento).

O principal coeficiente para esta análise é o Coeficiente de Correlação de Pearson (r), que mede o grau de associação linear entre duas variáveis. Se apresenta como um intervalo que varia de -1 a $+1$ e sua interpretação segue os parâmetros detalhados a seguir:

- $r = +1$: Correlação linear positiva perfeita.
- $r = -1$: Correlação linear negativa perfeita.
- $r = 0$: Ausência de correlação linear (pode haver outra forma de relação).

Além do Coeficiente de Correlação de Pearson (r), as métricas MSE, R^2 , MAE e RMSE são métricas de erro ou desempenho usadas para quantificar a precisão de uma medição em relação a outra, ou a precisão de um modelo de previsão.

Erro Absoluto Médio (MAE) é o erro médio em unidades originais, tratando todos os erros (diferenças) com peso igual, sendo menos sensível a outliers.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Nesta equação, o símbolo $\sum$ (sigma) representa o somatório, indicando que a operação deve ser realizada somando-se os valores de cada iteração, desde a primeira ($i = 1$) até a última observação (n). A variável n denota o número total de observações (tamanho da amostra), y_i refere-se ao valor real observado e $\hat{y}_i$ indica o valor previsto pelo modelo para a i -ésima observação.

Erro Quadrático Médio (MSE) é a média das diferenças quadráticas, penalizando erros grandes mais severamente do que o MAE (devido ao quadrado).

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Analogamente ao MAE, utiliza-se o somatório ($\sum$) das diferenças quadráticas entre os valores reais (y_i) e previstos ($\hat{y}_i$), dividindo-se pelo número total de amostras (n).

Raiz do Erro Quadrático Médio (RMSE) é a raiz quadrada do MSE, trazendo a métrica de erro de volta à unidade original dos dados, tornando-a mais fácil de interpretar do que o MSE. Ainda assim, mantém a penalidade mais alta para erros grandes.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Coeficiente de Determinação (R^2) mede a qualidade do ajuste, indicando a proporção da variância na variável dependente que é explicada pelas variáveis independentes. Em modelos de regressão linear simples (com apenas uma variável preditora), o R^2 é o quadrado do coeficiente de correlação de Pearson (r^2). Um R^2 alto sugere que a linha de regressão (e, portanto, a correlação) explica bem os dados.

$$R^2 = 1 - \frac{SQ_{res}}{SQ_{tot}}$$

Nesta fórmula, SQ_{res} representa a soma dos quadrados dos resíduos (a variação não explicada pelo modelo) e SQ_{tot} é a soma total dos quadrados (a variação total dos dados observados). Kendall

1.2.6 Predição Prognóstica

O principal objetivo do **ERT** é quantificar e qualificar o grau de resposta de uma neoplasia ao tratamento neoadjuvante e, assim, fornecer informações prognósticas. Para o câncer de mama, está bem estabelecido que o **ERT** é um melhor preditor de sobrevida global do que outras medidas, entre elas o estadiamento patológico pós-tratamento (Symmans et al. 2007; Akbar et al. 2019). No entanto, para tumores do trato gastrointestinal, não há uma relação clara entre os escores disponíveis e a evolução do paciente. Nesse cenário, existem estudos demonstrando que a resposta patológica completa, ou regressão tumoral total, está associada a uma melhor evolução em pacientes com tumores de esôfago, estômago e reto (Maas et al. 2010; Campos-Lobato et al. 2011; Stark et al. 2019). Porém, a resposta patológica completa é observada em apenas 15-20% dos pacientes (Agarwal et al. 2013), de modo que, para a maioria dos pacientes, não há uma estratificação prognóstica clara de acordo com os diferentes graus de resposta.

1.2.7 Importância do Estadiamento patológico

A avaliação do **ERT** é acompanhada pelo estadiamento patológico (vide Tabela 1 para detalhes da classificação patológica de neoplasias gástricas) de acordo com as diretrizes da oitava edição do AJCC (American Joint Committee on Cancer 2017). Ao comparar os sistemas de avaliação de *Ryan*, *Dworak* e *Mandard* com o estadiamento patológico pós-tratamento (ypTNM), foi demonstrado que nenhum desses esquemas se mostrou superior na predição prognóstica em comparação ao estadiamento patológico (Kim et al. 2016). Sendo assim, a despeito da realização do escore pós-tratamento neoadjuvante e da tentativa de padronizá-lo, o estadiamento ainda é o melhor preditor prognóstico para pacientes com adenocarcinoma gástrico pós-terapia neoadjuvante. Essa divergência se dá, em parte, porque os sistemas atuais de análise histológica pós-terapia neoadjuvante não consideram a avaliação dos linfonodos, que é um dos principais indicadores prognósticos (Kim et al. 2010; Huebner et al. 2012; Kim et al. 2016).

Tabela 1 – Estadiamento patológico para tumores gástricos - AJCC 8ª edição

Código	Descrição
Tumor primário (pT)	
TX	O tumor primário não pode ser avaliado
T0	Sem evidência de tumor primário
Tis	Carcinoma in situ: tumor intraepitelial sem invasão da lâmina própria, displasia de alto grau
T1	Tumor invade a lâmina própria, muscularis mucosae ou submucosa
T1a	Tumor invade a lâmina própria ou muscularis mucosae
T1b	Tumor invade a submucosa
T2	Tumor invade a muscularis propria
T3	Tumor penetra o tecido conjuntivo subseroso sem invasão do peritônio visceral ou estruturas adjacentes
T4	Tumor invade a serosa (peritônio visceral) ou estruturas adjacentes
T4a	Tumor invade a serosa (peritônio visceral)
T4b	Tumor invade estruturas/órgãos adjacentes
Linfonodos regionais (pN)	
NX	Os linfonodos regionais não podem ser avaliados
N0	Ausência de metástase em linfonodos regionais
N1	Metástase em um ou dois linfonodos regionais
N2	Metástase em três a seis linfonodos regionais
N3	Metástase em sete ou mais linfonodos regionais
N3a	Metástase em sete a 15 linfonodos regionais
N3b	Metástase em 16 ou mais linfonodos regionais

Fonte: (American Joint Committee on Cancer 2017).

1.3 PATOLOGIA E VISÃO COMPUTACIONAL

1.3.1 Patologia Digital

A prática da anatomia patológica tem experimentado uma transformação significativa, catalisada pelo advento da Patologia Digital (**PD**). Este desenvolvimento consolidou-se com a introdução de scanners capazes de converter lâminas de vidro em imagens digitalizadas de alta resolução chamadas **WSI** (do inglês *Whole Slide Images (WSIs)*) (Rashidi et al. 2025). Este avanço na digitalização das lâminas, observado desde o início dos anos

2000, foi crucial, pois eliminou uma barreira significativa, permitindo assim a emergência da Patologia Computacional (**PC**), um campo que utiliza a inteligência artificial (**IA**) para resolver problemas em patologia (Pantanowitz et al. 2025; Rashidi et al. 2025; Hanna et al. 2025). Impulsionada pela disponibilidade de grandes conjuntos de dados e pelo rápido progresso nas pesquisas em **IA**, atualmente, a **PC** que representa uma mudança de paradigma na oncologia de precisão e no diagnóstico laboratorial (Chen et al. 2024a; Lu et al. 2024; Pantanowitz et al. 2025).

Atualmente, os sistemas de **IA** são empregados em uma vasta gama de funções, desde tarefas fundamentais de visão computacional — como a detecção de mitoses, segmentação de núcleos celulares e quantificação de linfócitos infiltrantes de tumor (**TILs**) (Saltz et al. 2018; Corredor et al. 2019)) — até aplicações complexas de suporte à decisão clínica, incluindo a graduação automática de tumores (por exemplo, o escore de Gleason para câncer de próstata (Nagpal et al. 2019)) e a predição de desfechos prognósticos baseados em biomarcadores digitais (Bera et al. 2019; Valieris et al. 2024). Além da análise morfológica, a **IA** generativa tem sido utilizada para a criação de dados sintéticos e modelos de treinamento, enquanto sistemas multimodais buscam integrar dados histopatológicos com informações genômicas e clínicas, visando um diagnóstico holístico e personalizado (Bera et al. 2019; Hanna et al. 2025; Rashidi et al. 2025; Pantanowitz et al. 2025).

1.3.2 Patologia Computacional

A evolução de modelos de visão computacional (subcampo da **IA** voltada para o manejo de dados visuais, como imagens e vídeos (Choi et al. 2020)), dialoga diretamente com a evolução da **PC**. A fim de compreender os desdobramentos da segunda é necessário entender a evolução da primeira. O campo da visão computacional, cujos primórdios datam desde 1920 com as tratativas inaugurais de processamento de imagem com o primeiro sistema de transmissão de imagens por cabo de Bartlane (Gonzalez e Woods 2018), apresentou um desenvolvimento expressivo nas últimas décadas devido a disponibilidade de grandes conjuntos de dados ("*big data*", diminuição do custo de armazenamento e melhoria da capacidade computacional (Gonzalez e Woods 2018)).

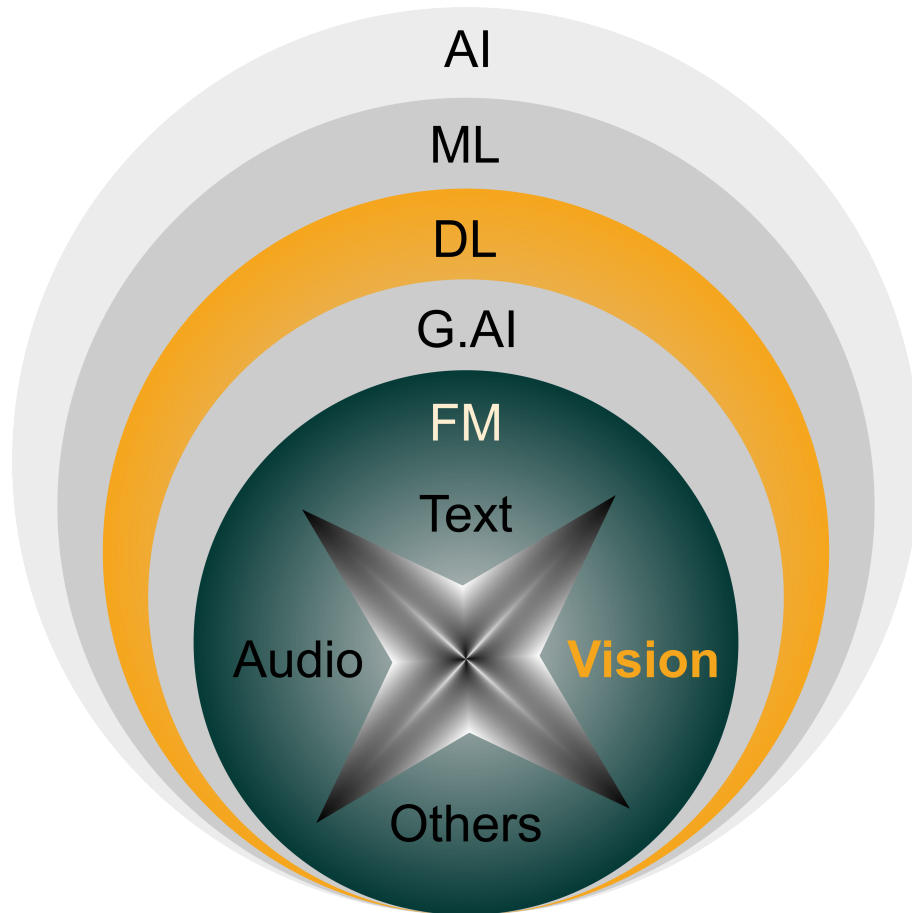
Dentro da **VC** existem diferentes subcampos: *Machine Learning* (**ML**) é a área que foca no aprendizado da **IA** com o desenvolvimento de algoritmos que identificam padrões (Choi et al. 2020). A abordagem mais inicial de **ML** se deu com técnicas *hand-crafted* (**HC**), na qual a extração de características (*features*) das imagens é realizada de forma manual (Nanni et al. 2017). Desde seu início e até hoje, existe na **PC** o uso de técnicas **HC** (Aswathy e Jagannath 2021), com extração de *features* texturais, morfométricas e de cor, para tarefas como segmentação de tecido e análise celular (Haralick et al. 1973; Tosta et al. 2021; Diamond et al. 2004). Em contraste, nas técnicas de *Deep Learning* (**DL**), uma técnica de aprendizado profundo que se inspirou nas redes neurais biológicas

e apresenta uma estrutura em Redes Neurais (**RN**) (Choi et al. 2020), existem camadas profundas que realizam essa extração de *features* (Nanni et al. 2017). Dentro de **DP**, se destacam as Redes Neurais Convolucionais (**CNN**), que trouxeram grandes avanços no campo da **VC**.

CNN é um tipo de **RN** que explora a estrutura dos dados para aprender automaticamente filtros que extraem características proeminentes dos dados a fim de aprender filtros de convolução que revelam padrões (como linhas e bordas) que podem ser utilizadas como *features* para executar tarefas como classificação (LeCun et al. 1995). Hoje, sabe-se que as camadas iniciais aprendem padrões mais simples, enquanto camadas subsequentes albergam aspectos mais elaborados (como formas e texturas) (Zeiler e Fergus 2014; Nguyen et al. 2016). Os modelos de **DL**, particularmente as **CNNs** ganharam grande força com o desafio *ImageNet Large Scale Visual Recognition Challenge (ILSVRC)* de 2010 em 2017 (Russakovsky et al. 2015), e modelos de **CNN** pré-treinados com pesos do grande dataset *ImageNet* (Deng et al. 2009) são utilizados ainda hoje para transferência de aprendizagem e *fine-tuning* - processo que implica em pegar uma **RN** já treinada e adaptá-la para uma nova tarefa - pautando-se na premissa de que os padrões simples e as *features* aprendidas nas camadas mais superficiais poderiam ser generalizados para outras tarefas. No campo da **PC**, as **CNN** desempenharam um papel relevante no desenvolvimento de modelos de **DL** voltados para tarefas específicas, como predição de biomarcadores, identificação de estruturas (como os **TILs** (Saltz et al. 2018; Corredor et al. 2019)) e predição de biomarcadores (Bera et al. 2019; Valieris et al. 2024). Mais recentemente, novos modelos de **ML** foram introduzidos na forma de Transformadores Visuais (Visual Transformers - **ViT**) (Liu et al. 2022; Aitazaz et al. 2023), que detém o estado da arte no campo até 2025.

Os Modelos de Fundação (**MF**, também conhecidos como *Foundation Models* (**FM**), são grandes modelos de Redes Neurais treinados com uma abordagem não supervisionada ou auto-supervisionada. Atualmente, representam a abordagem mais recente para a construção de modelos de **MF**, usados como alternativa ao *fine-tuning* com redes treinadas na ImageNet - dataset em larga escala (14 milhões de imagens) para pesquisa de aplicações visuais (Deng et al. 2009). Para a **PC**, os **MFs** em anatomia patológica trouxeram uma contribuição significativa. Esses **MF** são criados com o objetivo de extrair características significativas e discriminativas o que, por sua vez, possibilita mitigar a limitação imposta pela escassez de dados rotulados, ao gerar representações ricas e robustas dos tecidos, que possam ser transferidas e adaptadas para uma ampla gama de tarefas clínicas, como a classificação de subtipos de câncer, com uma necessidade mínima de anotações adicionais (aprendizado *few-shot* ou *zero-shot*). Um exemplo notável é o modelo CHIEF (Wang et al. 2024), que foi submetido a pré-treinamento com 60.530 **WSIs** de 14 coortes para prever desfechos de sobrevivência em 15 coortes de validação. Essa abordagem contrasta significativamente com abordagens prévias, que treinavam um modelo ponta a ponta para uma única tarefa.

Figura 11 – Diagrama de Venn **IA**



Fonte: Autoria Própria.

Legenda: Diagrama de Venn ilustrando subáreas de **IA**. As cores enfatizam métodos explorados no trabalho: **DL** e **MF** unimodais, treinados exclusivamente em imagens.

Além do CHIEF, dois outros **MF** são o UNI 2 (Chen et al. 2024b) e o Virchow 2 (Zimmermann et al. 2024a). O UNI2, também conhecido como ViT-Large, é um codificador visual auto-supervisionado que foi previamente treinado em um extenso conjunto de dados denominado Mass-100K. Este dataset é composto por aproximadamente 16 milhões de patches de tecido, que foram extraídos de mais de 21 442 imagens de amostras de tecido digitalizadas (**WSIs**). Por outro lado, o Virchow 2, juntamente com suas variantes, como o Virchow 2 que possui 632 milhões de parâmetros, visa maximizar tanto a escala quanto a diversidade dos dados. Este modelo foi treinado utilizando aproximadamente 3 milhões de **WSIs** provenientes de diversas instituições, cada uma com diferentes técnicas de coloração (Zimmermann et al. 2024a). O Virchow 2 implementa modificações algorítmicas específicas, como o *Extended-Context Translation* (**ECT**), com o objetivo de reduzir a distorção morfológica das células durante o processo de treinamento. Essas características permitem ao Virchow 2 alcançar um desempenho de ponta em 12 tarefas distintas de classificação de tiles.

Existem muitas técnicas de **ML** para lidar com **WSI**, que, devido ao seu tamanho massivo não podem ser processadas diretamente. É necessário extrair pequenos recortes

chamados tiles ou patches, gerando um "pacote de tiles" (do inglês *bag-of-tiles*) que representa a lâmina. O processo envolve a escolha de nível de zoom (MPP), tamanho do tile, sobreposição (*overlap*) e filtragem de fundo/artefatos. Dentre as técnicas podemos citar:

1. **CNNs** e Transformers para segmentação: classificam cada tile individualmente para formar Regiões de Interesse (**ROI**) maiores.
2. Multiple Instance Learning (**MIL**): Recebe o *bag-of-tiles* como entrada e aprende uma função para mapeá-lo para um rótulo único da lâmina, usando métodos baseados em instâncias (agregação de votos/probabilidades) ou baseados em embeddings (aprendendo uma representação para o *bag*).

Além do amplo uso das **CNN** e do surgimento dos **MFs**, outro método de **DL** vem chamando a atenção no campo da **PC**: as Redes Neurais em Grafos (do inglês *Graph Neural Networks*, **GNN**). Grafos são representações de objetos altamente customizáveis, podendo conter informação adicional: nós são passíveis de conter *features* que descrevem seu conteúdo, como as *features* morfológicas para tiles de imagens histológicas; arestas podem incluir informação que descreve a relação entre elementos, como peso. Um grafo G é representado por um conjunto de nós que representa os elementos do sistema e arestas, representando as interações entre esses elementos. As **GNN** tentam trazer a representação obtida em Redes Neurais para o domínio de dados em grafos. Em sua aplicação na **PC**, diferentemente das técnicas de **ML** de **CNN** e Transformers e **MIL**, as **GNNs** requerem a construção de grafos a partir das lâminas, geralmente conectando tiles vizinhos ou baseados em distância, permitindo modelar a estrutura e o microambiente do tecido.

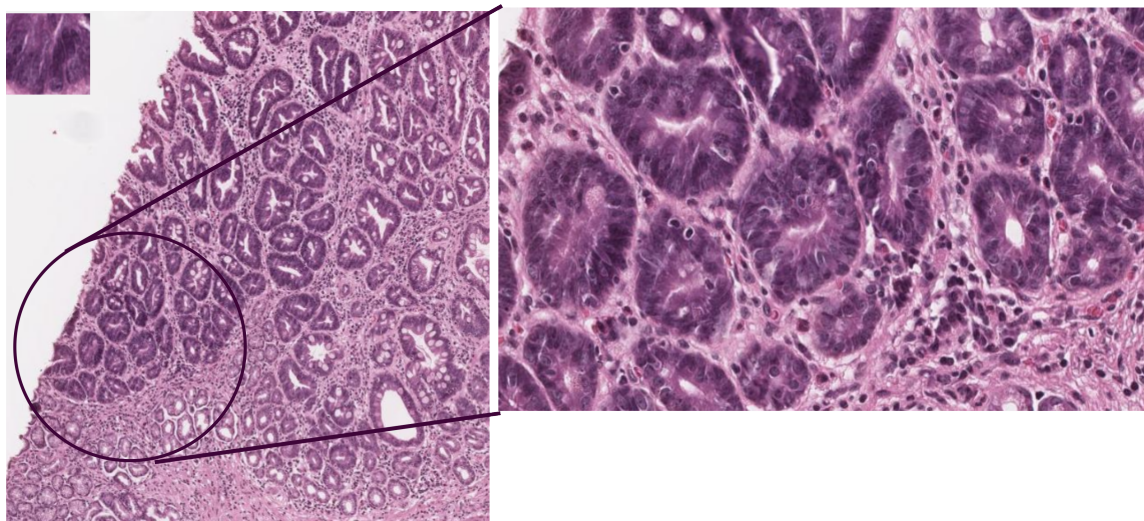
O uso das **GNNs** na anatomia patológica ainda não está tão consolidado como o das **CNNs**, contudo já existem estudos validando sua aplicação (Ahmedt-Aristizabal et al. 2022). Baranwal et al. (Baranwal et al. 2021) apresentam um trabalho com o uso de (*Graph Convolution Network*, **GCN**) (Kipf e Welling 2017) para agregar as características a nível de nó em um arranjo em grafos para graduação de doenças pancreáticas. Wang et al (Wang et al. 2020) utiliza camadas de **GCN** para classificação de *tissue micro-arrays* de próstata (TMA). Zhou et al (Zhou et al. 2019) propõe um novo método de convolução baseado em GraphSAGE que demonstrou melhor performance preditiva para graduação de câncer colorretal em comparação com outras abordagens que utilizaram **CNNs**.

Embora recente, o interesse das **GNNs** para a anatomia patológica pauta-se em potenciais vantagens das **GNNs**, como modelagem direta de interações ponto a ponto, discernimento eficaz das estruturas topológicas de tecido e celular, aquisição de representações sensíveis ao relacionamento (*relationship-aware*), modelagem do **WSI** inteira de uma só vez, captura da estrutura global do **WSI**, mitigação do viés de patching, adequação para lidar com dados irregulares e não sequenciais, modelagem hierárquica e interpretabilidade em nível de entidade (*entity-wise*) (Brussee et al. 2024). Além de

indicar validação, alguns trabalhos no domínio da anatomia patológica demonstram uma melhor performance para um objetivo específico de **GNNs** quando comparados com outros modelos como *base-line* (Zhou et al. 2019; Wu et al. 2023; Alzoubi et al. 2024).

RNs como **CNNs** e **ViTs** lidam com dados regulares, com formato em grid, como matrizes que representam dados tabulares, imagens e áudio. Ou seja, esses dados estão no domínio Euclidiano e, por conseguinte, apresentam composicionalidade o que, por sua vez, permite que **RNs** aprendam funções a partir desses dados (LeCun et al. 2015). Dados em grafos não apresentam essa propriedade devido à sua natureza irregular. Em geral, dados em grafos do mundo real, como redes sociais, representam sistemas com relações complexas (Amaral 2022). A análise histológica de tecidos realizada pelo médico patologista pode ser interpretada como um sistema complexo, com dados de natureza irregular que depende do contexto espacial e das interações entre os diferentes tecidos. Em outras palavras, a tomada de decisão diagnóstica não se baseia na análise isolada dos componentes, mas sim na avaliação desses componentes dentro do ecossistema de todo o tecido.

Figura 12 – Contextualização de *patch*



Fonte: Autoria própria.

Legenda: Exemplo de *patch* fora de contexto com significado histológico ambíguo. A interpretação correta do *patch* no canto superior esquerdo só é possível no contexto de menor aumento, o que evidencia que se trata de área de metaplasia intestinal na mucosa gástrica.

Células avaliadas de forma isolada, desconectadas do contexto, podem ter interpretações histológicas ambíguas. Exemplificando, podemos citar um *patch* - pequena subimagem ou região retangular extraída de um **WSIs** - apresentando células com características epiteliais atípicas. Sem o contexto apropriado, tais células podem representar tanto atipia reativa, quanto displasia, ou ainda células de uma neoplasia invasiva. A verdadeira natureza do processo patológico subjacente só pode ser identificada através da avaliação dessas células dentro de seu contexto (Figura 12). Essa concepção, que valoriza os entornos do objeto de análise, dialoga com a consideração do contexto de vizinhança

de **GNNs** - que modela o **WSI** como um grafo, no qual os *patches* ou células são considerados nós conectados por relações de vizinhança ou similaridade - e como o esquema de **MP**, que promove a comunicação da informação entre esses nós. Assim, as **GNNs** têm a capacidade de modelar o raciocínio relacional, capturando as dependências espaciais intrincadas de uma imagem histológica devido a sua capacidade de aprendizado topológico (Figura 13).

O maior desenvolvimento das **GNNs** se deu na evolução dos esquemas de Passagem de Mensagens (do inglês *Message Passing*, **MP**), que teve seu início com a abordagem espectral pioneira de (Bruna et al. 2014), na qual modelou-se a convolução como processamento de sinais via transformação de *Fourier*, embora com alto custo computacional. Para mitigar isso, a ChebNet (2016) (Defferrard et al. 2016) introduziu filtros polinomiais localizados via expansão de *Chebyshev*, seguida pela GCN (2017) (Kipf e Welling 2016), que simplificou o processo para uma aproximação de primeira ordem eficiente. Essa trajetória culminou na formalização do esquema de **MP**, onde os nós trocam e agregam informações, permitindo o surgimento de arquiteturas como GraphSAGE (Hamilton et al. 2017) (amostragem para grandes grafos), GAT (Veličković et al. 2017) (mecanismos de atenção) e GIN (Xu et al. 2018) (focada na expressividade e injetividade), avançando recentemente para filtros racionais mais flexíveis e robustos, como a **GNN ARMA** (Bianchi et al. 2019), que supera limitações dos filtros polinomiais tradicionais ao capturar melhor a estrutura global e respostas de frequência complexas.

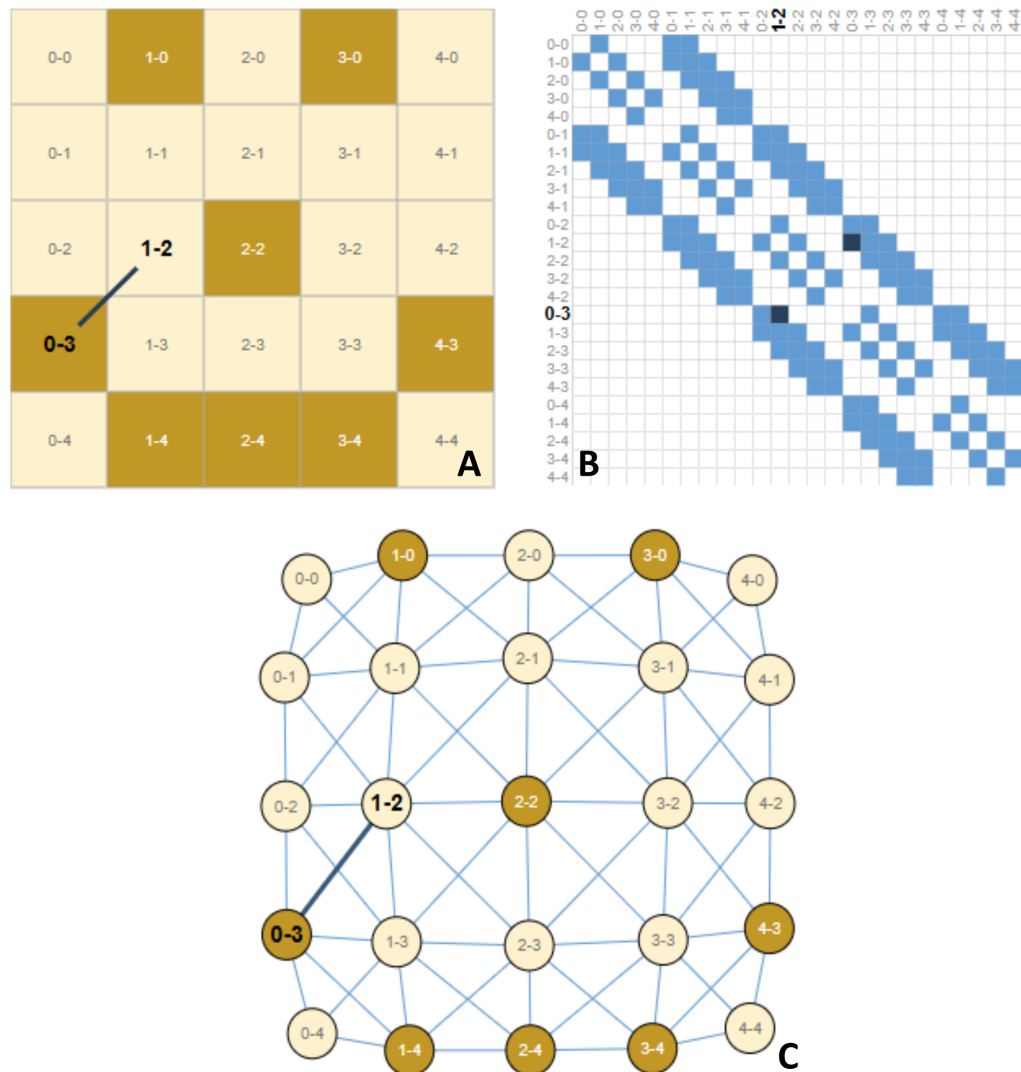
A abordagem de **MP** modela a interação entre nós como um processo de troca e consolidação de informações dividido em três etapas sequenciais (Gilmer et al. 2017): primeiro, ocorre o Cálculo da Mensagem, onde os nós vizinhos processam suas características e as enviam para o nó central ou atual; em seguida, realiza-se a Agregação da Mensagem, na qual o nó central coleta todas as mensagens recebidas e as combina em uma representação de tamanho fixo utilizando funções invariantes à permutação e ao número de vizinhos, como soma, média ou máximo; por fim, executa-se a Atualização do Nó, etapa em que o nó central processa a representação agregada da vizinhança para atualizar seu próprio estado ou valor, gerando o *embedding* que será utilizado nas camadas subsequentes da rede.

1.3.2.1 Revisão histórica da visão computacional na Patologias Computacional

A seguir reorganizaremos os eventos de relevância para o desenvolvimento da **VC** e **PC**, com maior detalhamento das diferentes arquiteturas de **GNN**, que são foco desse trabalho.

- 1973 - (Haralick et al. 1973) (*Hand-crafted features* - **HCF**): utilização de características manuais (*"hand-crafted"*), como características de textura, para a análise de imagens histopatológicas antes do advento de **DL**.

Figura 13 – Arquitetura em grafos de uma **GNN**



Fonte: Sanchez-Lengeling et al. (2021).

Legenda: esquema que representa a arquitetura estruturada em grafos de uma **GNN**. A indicada a matriz com pixels com uma relação entre dois pixels correspondentes à área ressaltada na matriz adjacente em B. C demonstra essa relação em forma de grafo.

- Anos 1990: Proposta inicial das **CNNs** (LeCun et al. 1995): primeira proposta das **CNN**, que funcionam aproveitando a estrutura de composicionalidade dos dados de entrada, embora sua popularidade dominante só tenha ocorrido na década de 2010.
- 2009 - *Dataset ImageNet* (Deng et al. 2009): esforço pioneiro na compilação de um dataset em larga escala (14 milhões de imagens) para pesquisa de aplicações visuais, permitindo o desenvolvimento de novas abordagens de avaliação e visão computacional.
- 2010 – 2017: **ILSVRC** (Russakovsky et al. 2015): desafio que promoveu o desenvolvimento de novas abordagens para visão computacional, levando à popularidade dos modelos de **DL**, particularmente as **CNNs**, que dominaram os rankings.

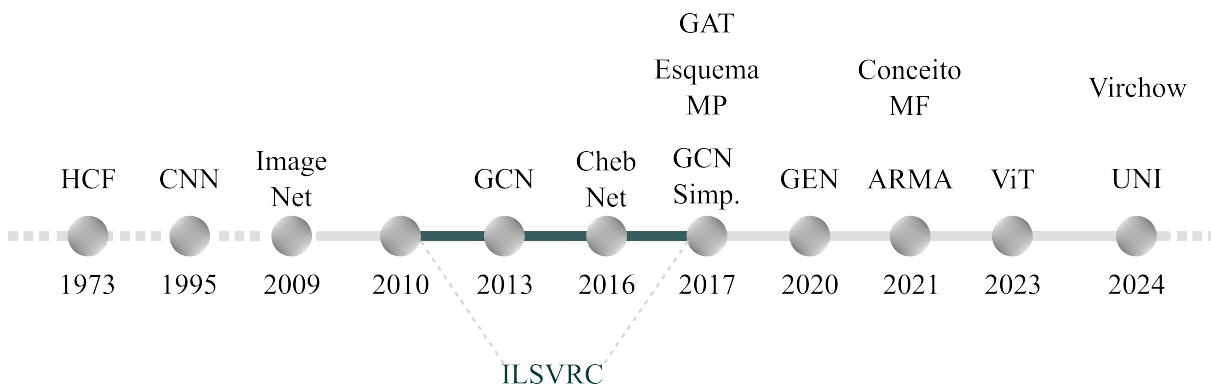
- 2013 - Trabalho pioneiro de (Bruna et al. 2014): modelou a *Graph Convolution Network* (GCN) como tarefas de processamento de sinal entre uma matriz de atributos de nós e pesos (filtro), marcando o início das melhorias no campo das Redes Neurais Convolucionais em Grafos.
- 2016 - ChebNet (Defferrard et al. 2016): introduziu o uso de polinômios de Chebyshev para aproximar filtros espectrais. Isso reduziu os requisitos computacionais ao evitar o cálculo da transformação de Fourier e permitiu filtros localizados a K -saltos de distância (K -hops).
- 2017 - Esquema **MP** (Gilmer et al. 2017): formalização das simplificações introduzidas ao longo dos anos nas **GNNs**, modelando a interação como troca de informações entre nós.
- 2017 - GCN *Graph Convolution Network* (Kipf e Welling 2016) – Simplificou propostas anteriores limitando a convolução a uma aproximação de primeira ordem e compartilhando pesos entre o nó atual e os vizinhos.
- 2017 - GAT - *Graph Attention Network* (Veličković et al. 2017): propõe uso de mecanismos de atenção para ponderar as contribuições de diferentes nós vizinhos no grafo durante a agregação de informações.
- 2020 - GEN *DeeperGCN (GEN layer)* (Li et al. 2020): Introduz novas funções de agregação (*SoftMax* e *PowerMean*) que combinam linearmente funções clássicas, permitindo o treinamento de redes mais profundas.
- 2021 - ARMA *Graph Neural Network* (Bianchi et al. 2019) – Propõe uma camada convolucional baseada em filtros ARMA (*Auto-Regressive Moving Average*) racionais. Diferente dos filtros polinomiais (como GCN e ChebNet), a camada ARMA oferece uma resposta de frequência mais flexível, maior robustez a ruído e captura melhor a estrutura global do grafo sem a necessidade de redes profundas, utilizando uma implementação recursiva e distribuída.
- 2021 - Conceito de **MF**: abordagem de construção de grandes modelos de Redes Neurais treinados de forma não supervisionada ou auto-supervisionada com o objetivo explícito de extrair características significativas.
- 2023 - ViT *Visual Transformers* (Liu et al. 2022; Aitazaz et al. 2023): modelos de aprendizado profundo inspirados no processamento de linguagem natural que utilizam mecanismos de atenção para analisar imagens, alcançando o estado da arte em diversos benchmarks.
- 2024 - UNI *Foundation Model* (Chen et al. 2024a): modelo de Fundação para histopatologia desenvolvido pelo *Mahmood Lab*, treinado com mais de 100.000 lâ-

minas histológicas, servindo como base para extração de características em novas aplicações.

- 2024: *Virchow 2* e *Virchow 2G* (Paige & Microsoft) (Zimmermann et al. 2024a) – modelos de fundação em escala massiva (até 1,9 bilhões de parâmetros) treinados com 3,1 milhões de imagens de lâmina inteira (WSI) de diversas origens globais e múltiplas ampliações (5x, 10x, 20x, 40x). Demonstrou que a diversidade de dados e adaptações específicas de domínio no treinamento auto-supervisionado superaram modelos anteriores no estado da arte de tarefas de patologia.

A Figura 14 ilustra a evolução dos eventos de relevância da VC para PC.

Figura 14 – Evolução visão computacional



Fonte: Autoria própria.

Legenda: Linha do tempo de evolução da visão computacional ilustrando a evolução dos eventos de relevância da **VC** para **PC**.

1.4 RACIONAL DO ESTUDO E DEFINIÇÃO DOS OBJETIVOS

Por esta motivação, o presente estudo propõe explorar a melhor combinação das diferentes arquiteturas de **GNNs** com seus diferentes esquemas de **MP**, associadas a dois **MFs** para extração de features para reproduzir o **ERTG**. Hipotetizou-se que a análise mais contextualizada e topológica das **GNNs** associadas ao **MF** (cujo pré-treinamento em vastos e diversos conjuntos de dados histológicos mitiga a escassez de dados rotulados) seria capaz de mitigar as limitações inerentes dos **ERTG**. Esse estudo também explorou se a reprodução do **ERTG** mediante uso de **M1** associada a outras variáveis anátomo-clínicas do laudo poderia resultar em um ganho preditivo prognóstico significativo para sobrevida pós tratamento cirúrgico (**SPC**), em comparação com as variáveis anátomo-clínicas ou do **ERTG** isolados.

2 MATERIAIS E MÉTODOS

2.1 DEFINIÇÃO DA COORTE DO ESTUDO

A coorte deste estudo foi delineada para atingir dois objetivos principais: o treinamento de um modelo de **GNN** (daqui em diante referido como **M1**) e a avaliação de seu desempenho na **ERTG**. 452 casos de ressecções cirúrgicas pós-tratamento neoadjuvante de pacientes com diagnóstico de adenocarcinomas de **EST** e **RETO** identificados no banco de dados da anatomia patológica do **ACCCC** foram solicitados no arquivo da instituição. Dentre esses, 246 foram encaminhados ao patologista para análise, conforme os critérios de inclusão estabelecidos para o estudo. Após a revisão dos casos, 115 foram descartados por não atenderem aos critérios de inclusão, dos quais 77 devido à documentação incompleta. Os 255 casos remanescentes foram então digitalizados para prosseguir com o estudo.

Para o treinamento, validação e teste do **M1**, foi selecionado um conjunto diversificado de pacientes diagnosticados com carcinomas do trato gastrointestinal, especificamente adenocarcinomas esofago-gástricos, gástricos e colorretais. Esta seleção fundamenta-se nas semelhanças observadas na avaliação de adenocarcinomas entre esses órgãos do trato digestivo, bem como na similaridade dos protocolos de **ERT** aplicados a esses órgãos, o que permite o intercâmbio de escores entre eles. Este argumento é corroborado pelo fato de que o protocolo recomendado pelo Colégio Americano de Patologistas (**CAP**) (College of American Pathologists 2023c) para o estômago é uma adaptação do protocolo de *Ryan*, originalmente desenvolvido para tumores retais.

Além de os protocolos de **ERT** serem semelhantes e intercambiáveis, a estrutura histológica dos órgãos envolvidos é comparável, tratando-se de órgãos ocos e/ou tubulares do trato gastrointestinal, com uma estrutura histológica similar, composta de um revestimento mucoso na parte luminal, seguido por camadas de tecido muscular, tecido conjuntivo e tecido adiposo, que podem estar recobertos ou não por serosa. Ademais, para a construção do **M1**, buscou-se explorar todo o ecossistema de lâminas histológicas a fim de validar a hipótese de que, para maximizar o potencial de uma **GNN**, com sua arquitetura baseada em nós e arestas e sua capacidade de capturar padrões espaciais e contextuais, seria ideal incluir no treinamento a maior diversidade possível de componentes histológicos presentes em lâminas de pacientes com diagnósticos de adenocarcinoma pós-tratamento neoadjuvante. Para alcançar esse objetivo, fazia-se necessária uma ampla variedade de componentes para o desenvolvimento do modelo.

Considerando que muitos dos componentes histológicos presentes em tecidos que respondem ao tratamento apresentam elevada variabilidade, sem distribuição consistente (algumas características histológicas surgem apenas esporadicamente), optou-se por desenvolver o **M1** com dados de adenocarcinomas de esôfago baixo, estômago, transição

colorretal e reto, visando a uma amostra abrangente da diversidade de componentes que constituem um leito tumoral pós-tratamento neoadjuvante. Para o segundo objetivo, que visa destacar a relevância clínica do **M1** no contexto da **ERTG**, o grupo de estudo foi composto exclusivamente por pacientes com adenocarcinoma do estômago e da junção esôfago-gástrica.

Os critérios de inclusão para ambos os objetivos do estudo concentraram-se em pacientes com diagnóstico confirmado de adenocarcinoma, que foram estadiados de acordo com o sistema pTNM, até o estágio ypT3 após o tratamento. Ambas as instituições envolvidas no projeto aderiram rigorosamente aos protocolos estabelecidos pela **AJCC** e pelo **CAP**, utilizando o escore modificado de *Ryan* para a análise macroscópica e a avaliação histológica. Casos em que as lâminas estavam danificadas ou apresentavam coloração inadequada, o que impossibilitava a reavaliação diagnóstica e a avaliação do **ERTG**, foram excluídos do estudo. Um patologista revisou todos os casos para assegurar o cumprimento dos critérios predefinidos. As lâminas de vidro adequadas foram posteriormente digitalizadas com o digitalizador de alta resolução Aperio da *Leica Biosystems*, disponibilizando-as em uma plataforma personalizada desenvolvida especificamente para o projeto.

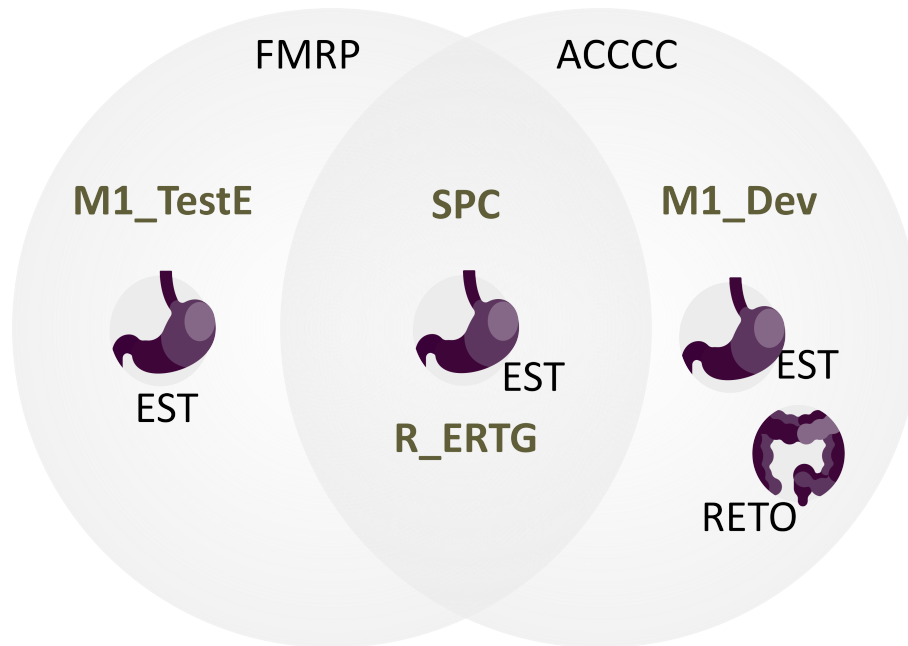
2.1.1 Distribuição da coorte para os diferentes testes

Após a aplicação dos critérios de inclusão e exclusão, a coorte foi distribuída para testar as diferentes hipóteses formuladas ao longo do estudo. A coorte foi dividida nos seguintes grupos a fim de atender às premissas:

1. **Modelo M1**: subgrupo selecionado para treinamento, validação e teste de um modelo de **GNN**, com o objetivo de identificar categorias histológicas relacionadas à resposta ao tratamento presentes em lâminas histológicas de pacientes com diagnóstico de adenocarcinoma pós-terapia neoadjuvante (Figura 15).
 - a) Desenvolvimento do modelo de **GNN** M1 (**M1_Dev**): amostra de casos exclusivamente do **ACCCC**, com casos de adenocarcinomas de transição esôfago-gástricas e estômago (agrupados sob a sigla **EST**) e de transição colorretal e reto (agrupados sob a sigla **RETO**).
 - b) Teste externo de M1 (**M1_TestE**): amostra de casos exclusivamente de RP, com exemplares de adenocarcinomas de transição esôfago-gástricas (**TEG**) e de estômago (**EST**). Nesse teste externo, já havia interesse em avaliar o comportamento do modelo na reprodução do **ERTG**.
2. **Reprodução do ERTG (R_ERTG)**: subgrupo selecionado para avaliar se o **ERTG** estimado pelo modelo **M1** apresenta concordância com as avaliações de um conjunto de patologistas. Nesse subgrupo, pacientes incluídos no treinamento e na validação de **M1** foram excluídos.

3. **Predição de sobrevida pós-tratamento cirúrgico (SPC)**: subgrupo selecionado para avaliar qual é a melhor combinação de variáveis para prever a sobrevida pós-ressecção cirúrgica. O conjunto de variáveis inclui dados contidos nos laudos anatomo-patológicos, além da avaliação do **ERT**, das estimativas do **ERT** do modelo **M1** e do **ERT** realizado por patologistas.

Figura 15 – Distribuição da coorte em subgrupos



Fonte: Autoria própria.

Legenda: Diagrama de Venn ilustrando a distribuição da coorte em subgrupos.

2.2 CONSTRUÇÃO DO MODELO DE GNN M1

2.2.1 Definição das categorias histológicas

Quinze categorias foram estabelecidas para amostrar o ecossistema de lâminas de vidro: Mucosa Gástrica (**MG**), Mucosa Colônica (**MCo**), Mucosa Escamosa (**ME**), Estroma (**EST**), Estrutura Linfóide (**LINF**), Fibra Muscular (**MUSC**), Debris (**DEB**), Mucina Acelular (**MUC**), Tumor Invasivo (**TUMIN**), Fibrose (**FIB**), Tecido Adiposo (**ADIP**), Necrose (**NEC**) e Calcificação (**CAL**). Os critérios patológicos detalhados para cada categoria podem ser encontrados a seguir.

1. Mucosa Gástrica (**MG**):
 - a) Mucosa gástrica sem particularidades histológicas, compreendendo:
 - i. Glândulas mucinosas
 - ii. Glândulas oxínticas

- iii. Foveolas
- iv. Lâmina própria

2. Mucosa Colônica (**MC**):

- a) Mucosa colônica sem particularidades histológicas, compreendendo:
 - i. Epitélio de superfície
 - ii. Criptas
 - iii. Lâmina própria
- b) Áreas de metaplasia de padrão intestinal completa ou incompleta, podendo conter displasia de alto grau sem componente invasivo, em transição esôfago-gástrica e estômago.

3. Mucosa escamosa sem particularidade histológicas (**ME**):

- a) Esôfago
- b) Ânus

4. Estroma (**EST**):

- a) Tecido conjuntivo sem sinais de processo reparativo em resposta ao tratamento prévio.
 - i. Marcações predominantemente em lâminas das margens de peças cirúrgicas.
- b) Tecido neural

5. Fibra muscular (**MUSC**):

- a) Camada muscular própria
- b) Camada muscular da mucosa
- c) Parede calibrosa de vasos

6. Estrutura linfoide (**EL**):

- a) Estrutura linfoide primária
- b) Estrutura linfoide secundária
- c) Estrutura linfóide terciária

7. Debrí (**DEB**):

- a) Agregados hemorrágicos.
- b) Remanescentes celulares.

- c) Exsudato supurativo.
- d) Placas fibrinoleucocitárias.

8. Mucina Acelular (**MUC**)

9. Tumor invasivo (**TUMIN**):

- a) Esôfago
- b) Estômago - subtipos
 - i. Adenocarcinoma tubular
 - ii. Adenocarcinoma de células parietais
 - iii. Adenocarcinoma papilífero SOE
 - iv. Carcinoma micropapilífero SOE
 - v. Carcinoma mucinoso
 - vi. Carcinoma com células em anel de sinete
 - vii. Carcinoma pouco coeso
 - viii. Carcinoma medular com estroma linfoide
 - ix. Adenocarcinoma hepatoide
 - x. Carcinoma de células de Paneth
- c) Cólon - subtipos
 - i. Adenocarcinoma serrilhado
 - ii. Adenocarcinoma adenoma-like
 - iii. Adenocarcinoma Micropapilar
 - iv. Adenocarcinoma Mucinoso
 - v. Adenocarcinoma medular
 - vi. Carcinoma pouco coeso
 - vii. Carcinoma com células em anel de sinete

10. Fibrose (**FIB**):

- a) Definição: extensa deposição de colágeno.
- b) Fibrose associada à regressão tumoral deve ser definida quando intimamente associada a um ou mais componentes histológicos:
 - i. Lagos de mucina (com ou sem células tumorais).
 - ii. Necrose
 - iii. Macrófagos espumosos.
 - iv. Fendas de colesterol.
 - v. Células gigantes multinucleadas do tipo corpo estranho.

vi. Focos isolados/ilhas de tumor residual embebidos em áreas de fibrose.

11. Processo Inflamatório Reparativo/Inespecífico (**PIRI**):

- a) Histiócitos espumosos.
- b) Tecido de granulação.
- c) Granuloma.
- d) Células gigantes multinucleadas.
- e) Macrófagos espumosos.
- f) Fendas de colesterol.
- g) Células gigantes multinucleadas do tipo corpo estranho.
- h) Tecido edemaciado, com infiltrado inflamatório crônico, que não está agregado o suficiente para ser considerado EL.

12. Tecido adiposo (**ADIP**)

13. Artefato (ART):

- a) Materiais exógenos, artefatos pré-analíticos ou problemas no corte que impossibilitem a identificação do tecido subjacente.
- b) Áreas de tatuagem
- c) Tinta

14. Necrose (**NEC**)

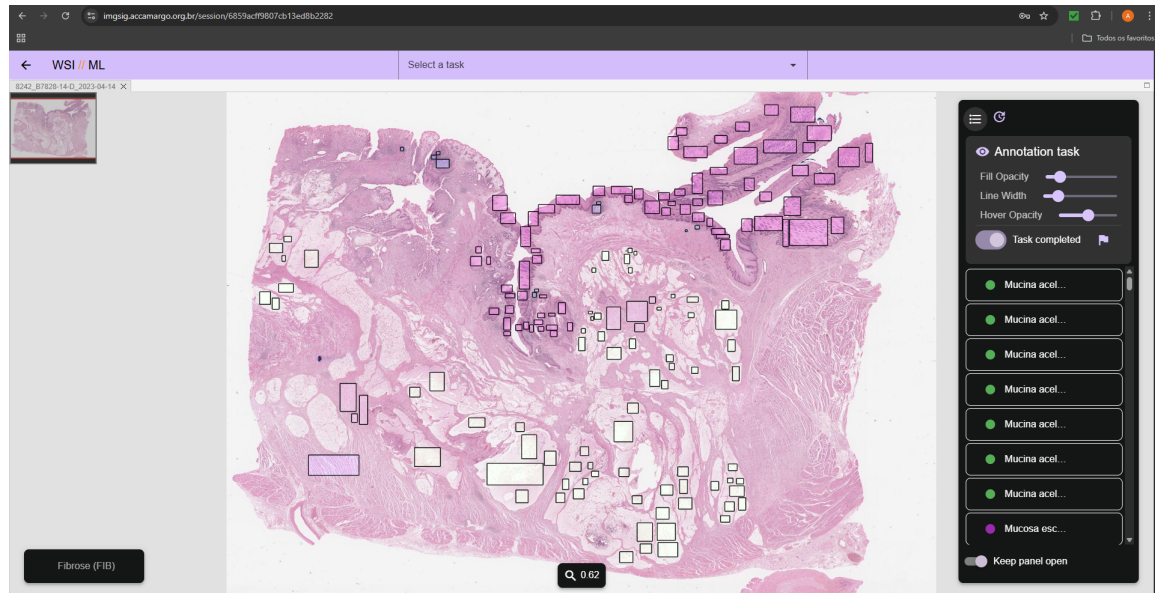
15. Calcificação distrófica (**CAL**)

2.2.2 Distribuição de dados para desenvolvimento de M1

Com o objetivo de viabilizar o treinamento (**Train**), a validação (**Val**) e o teste interno (**TestI**) do modelo em desenvolvimento, procedeu-se à anotação manual de uma amostra de casos digitalizados. Essas anotações foram conduzidas em uma plataforma customizada, desenvolvida internamente pelo grupo do laboratório de bioinformática ao longo da execução do projeto (Figura 16 (Martins et al. 2023)). Subsequentemente, cada anotação foi segmentada em *tiles* (fragmentos de imagem com dimensões de 224×224 pixels e resolução de $0,5 \mu\text{m}/\text{pixel}$), resultando na construção de uma biblioteca de *tiles* (Figura 17).

O conjunto de dados para treinamento, validação e teste (denominado subgrupo **M1_Dev**) foi constituído exclusivamente por casos provenientes do **ACCCC**, abrangendo as topografias **EST** e **RETO**. Esse grupo foi dividido em aproximadamente 50% para o grupo de treino (**M1_Dev - Train**), 30% para o grupo de validação (**M1_Dev - Val**) e 20% para o grupo de teste interno (**M1_Dev**).

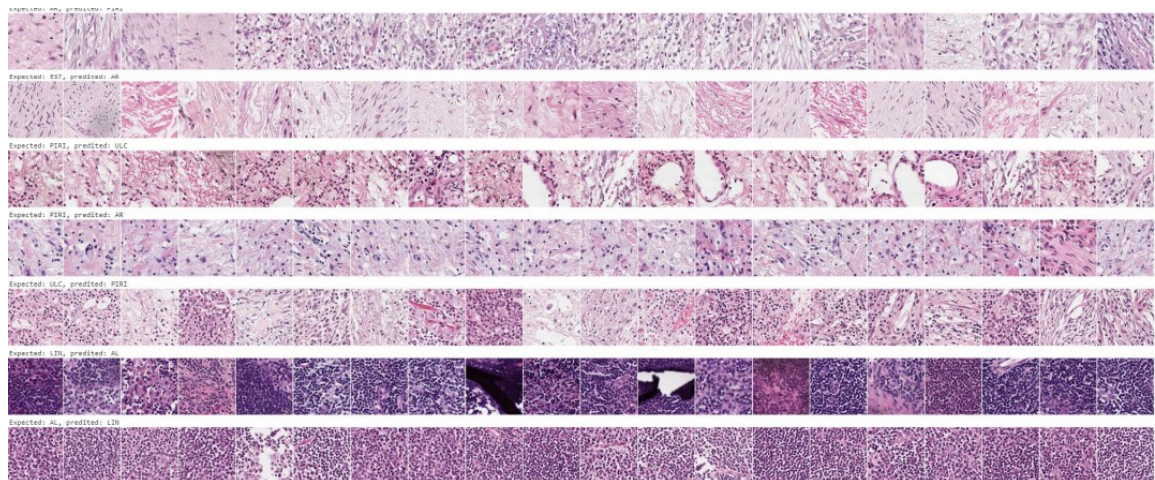
Figura 16 – Sistema de apoio de anotação



Fonte: Autoria própria.

Legenda: sistema de apoio à anotação disponibiliza uma interface com funcionalidades para que o especialista anote manualmente as regiões em lâminas do **LT**.

Figura 17 – Biblioteca de *tiles*



Fonte: Autoria própria.

Legenda: Exemplos de tiles gerados a partir das anotações manuais, constituindo uma biblioteca de *tiles*.

A anotação desses casos foi realizada por dois patologistas internos, A e W. Considerando a dificuldade inerente à distinção de certas características morfológicas do leito tumoral (**LT**), optou-se pela inclusão de um terceiro patologista externo (Patologista F), proveniente de uma instituição externa (o Hospital Israelita Albert Einstein — **M1**), para a anotação do subgrupo de teste interno de **M1** na tentativa de mitigar o viés de anotação e avaliar a generalização do modelo. Por conseguinte, este subgrupo de teste foi anotado pelos patologistas A e F. Tal estratégia permitiu avaliar se a predição de categorias pelo Modelo **M1** estaria restrita à concordância com as anotações do grupo do **ACCCC** (o que refletiria, consequentemente, a fidelidade à interpretação institucional), ou se o modelo seria capaz de se equiparar à interpretação de um patologista externo que recebeu apenas um treinamento baseado nas definições de categorias, para a realização das anotações.

2.2.3 Treinamento, validação e teste interno de M1

A metodologia empregada baseou-se na construção de um modelo de rede neural em grafos utilizando *Python* 3.12, *PyTorch* 2.7.1, *PyTorch Geometric* 2.7.0 e *CUDA* 12.9 para aceleração computacional. Os grafos foram gerados a partir de tiles de 256 *pixels*, sem sobreposição e com resolução de 0,5 MPP, permitindo representar cada região como um subgrafo estruturado. A arquitetura da **GNN** foi composta por três camadas de convolução em grafos com ativação *leaky ReLU*, seguidas por uma camada linear final, adotando 256 hidden channels para garantir capacidade expressiva adequada. O treinamento foi conduzido por até 100 épocas, com taxa de aprendizado de 0,0005, acumulação de gradientes a cada 10 iterações e *batch size* de 32 subgrafos para treinamento e 2 grafos completos para validação, assegurando estabilidade e eficiência no processo de otimização.

A extração de características (*features*) foi realizada com dois modelos de fundação para anatomia patológica (modelos que já tiveram *fine-tuning* voltado para lâminas de patologia), diferentes: Virchow 2 (Zimmermann et al. 2024b) e UNI2 (Chen et al. 2024c). Optou-se por utilizar **GNN** no desenvolvimento do modelo, denominado **M1**, partindo da premissa de que um método de aprendizado que opera diretamente sobre estruturas de grafos poderia alcançar melhores resultados. Isso porque as **GNNs** são capazes de aprender e inferir a partir das relações e conexões entre as entidades representadas no grafo (Corso et al. 2024). As **GNNs** (Defferrard et al. 2016; Kipf e Welling 2016; Hamilton et al. 2017; Veličković et al. 2017; Xu et al. 2018; Bianchi et al. 2019; Li et al. 2020) testadas foram treinadas e validadas a partir das áreas anotadas para identificar as 15 categorias definidas no projeto.

Após a determinação da combinação ideal entre o método de extração de características e a categorização, foi elaborada uma matriz de confusão (**MC**) com base no grupo de validação (**M1_Dev - Val**). Essa matriz exhibe a comparação entre as predições do modelo e as anotações manuais prévias, com o intuito de avaliar os erros e o desempenho de cada classe. Enquanto as métricas de avaliação do modelo permitem quantificar seu

desempenho de maneira global, a **MC** fornece uma análise mais detalhada do comportamento do modelo, possibilitando o rastreamento mais preciso de erros específicos.

Para o teste interno (**M1_Dev - TestI**), também se procedeu à comparação entre os diferentes modelos de **GNN**. Contudo, frente à performance superior com o **Virchow 2** para extração de features no grupo de validação, e devido ao fato de que essa coorte também era inteiramente proveniente da mesma instituição (**ACCCC**), de forma que as lâminas apresentavam um padrão similar, optou-se por não testar o uso do **UNI** para extrair as *features* nesse grupo de teste. Para essa bateria de teste, o intuito era avaliar o desempenho geral do modelo, a fim de verificar se **M1** era capaz de identificar as categorias que compõem as lâminas de pacientes com diagnóstico de adenocarcinoma do **TGI** pós-tratamento neoadjuvante. Diante dessa premissa, não foi gerada uma matriz de confusão para avaliar o comportamento do modelo.

2.2.4 Teste Externo de M1

O teste externo (**TestE**) foi realizado com a premissa de avaliar 2 instâncias. A primeira seria a performance de **M1**, já focada em neoplasias gástricas, a fim de dar início à avaliação do comportamento do nosso modelo, voltado para o objetivo principal do estudo, a **RERTG**. A segunda instância seria avaliar o desempenho do modelo na identificação das categorias em pacientes com câncer gástrico, em uma coorte externa, e, assim, realizar uma validação externa. Nesse contexto, foram realizados testes no subgrupo designado como **M1_Dev - TestE**, novamente com diferentes arquiteturas de **GNN** utilizando **UNI 2** (Chen et al. 2024c) e **Virchow 2** (Zimmermann et al. 2024b).

Esse dataset era composto exclusivamente por lâminas provenientes de outra instituição externa, possuindo, assim, características específicas relacionadas ao pré-processamento dessas lâminas. Essas características, como coloração, espessura de corte e tamanho do fragmento da lâmina, por exemplo, podem tornar as imagens geradas a partir da digitalização das lâminas diferentes das do **ACCCC**. Consequentemente, a despeito da aplicação de técnicas de pré-processamento de imagem, como a normalização de cores, poderia haver influência dessas dissimilaridades no desempenho do modelo **M1**. Dito isso, optou-se por incluir o método **UNI 2** na extração de características nesse grupo de teste externo (**M1_Dev - TestE**) apesar do desempenho superior do **Virchow 2** no grupo de validação (**M1_Dev - Val**). A Tabela 2 apresenta um resumo das combinações dos testes efetuados por subgrupo da coorte **M1_Dev**, juntamente com a especificação do objetivo de cada combinação.

Tabela 2 – Detalhamento dos testes realizados por subgrupo de M1_Dev. Ext. Feat. (Método utilizado para extração de *features*); MC (Se foi gerado uma Matriz de Confusão (MC) para avaliar comportamento de M1 no subgrupo designado); Objetivo (Objetivos dos testes no subgrupo).

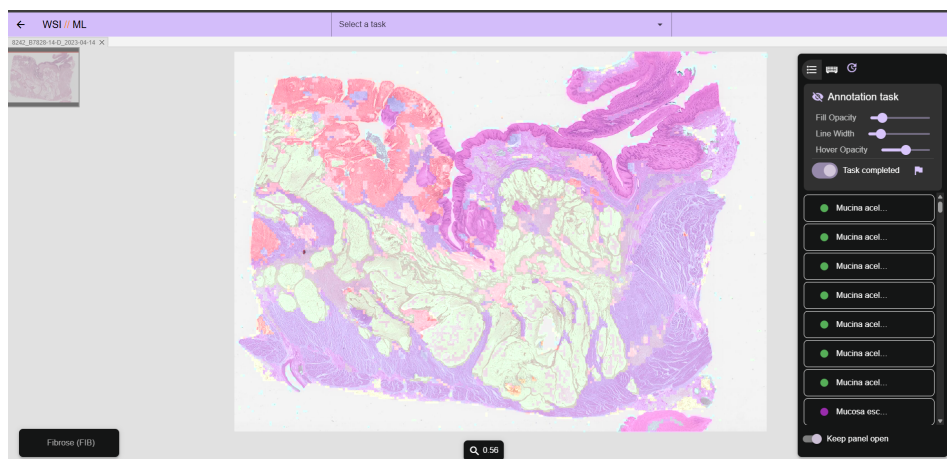
Subgrupo	Ext. Feat.	MC	Objetivo
M1_Dev - Val	UNI e Virchow 2	Sim	Validação e Aval_Comp - TGI
M1_Dev - TestI	Virchow 2	Não	TestI
M1_Dev - TestE	UNI e Virchow 2	Sim	TestE e Aval_Comp - EST

Fonte: Autoria própria.

2.2.5 Compreensão dos padrões de comportamento de M1

Para obter uma avaliação mais precisa do comportamento do modelo M1 observado a partir das matrizes de confusão geradas nos subgrupos **M1_Dev - Val** e **M1_Dev - TestE**, bem como para uma compreensão mais aprofundada das implicações desse desempenho para a estimativa geral do **ERTG**, foi realizada uma análise visual das predições das classes pelo modelo **M1**. Este processo incluiu a extensão das predições sobre imagens digitalizadas, utilizando uma plataforma desenvolvida pelo grupo. Esta análise foi conduzida com o objetivo de identificar e detalhar potenciais padrões de erro, permitindo uma compreensão mais acurada do funcionamento do modelo **M1**. Essa abordagem facilitou a identificação de tendências de viés do modelo **M1** e a correlação dos erros detalhados nas **MCs** com os processos patológicos subjacentes observados em lâminas de pacientes submetidos a tratamento neoadjuvante. A Figura 18 apresenta um exemplo de distribuição das classes preditas pelo modelo **M1** em uma lâmina de tecido de **EST**.

Figura 18 – Predição das classes em **LT**



Fonte: Autoria própria.

Legenda: predição das classes em lâmina de **LT** previamente anotada. As classes estimadas são destacadas pelas cores associadas, conforme o painel no canto direito.

2.3 PERFORMANCE DE M1 PARA REPRODUÇÃO DE ERTG

Para avaliar a estimativa do **ERTG** com o modelo **M1**, procedeu-se à análise da concordância entre a estimativa gerada pelo modelo **M1** e o **ERTG** dos laudos originais, na qual **M1** assumiu uma função semelhante à de um patologista. Considerando a já conhecida alta variabilidade interobservador na mensuração dos componentes do **LT** e, consequentemente, na execução do **ERTG**, tornou-se necessário estabelecer uma "verdade fundamental" (Ground Truth - **GT**) mais robusta. Subsequentemente, três patologistas especializados em **TGI**, designados como '**G**', '**W**' e '**M**', revisaram o **ERTG** em casos de **EST** incluídos no subgrupo **R_RET**, sem acesso aos resultados originais. Esse grupo de três patologistas especializados em **TGI** foi designado como **Patol_TGI**. A concordância entre esses três patologistas foi então analisada, assim como a concordância da estimativa de **ERTG** gerada pelo modelo **M1**, que foi considerado como um quarto patologista. Para garantir um **GT** mais robusto, uma segunda análise de concordância foi realizada apenas entre a estimativa do modelo **M1** e os casos em que houve consenso entre os três patologistas.

2.3.1 Definição do Leito Tumoral (LT)

O Leito Tumoral (**LT**) é definido como a área do tecido na lâmina que engloba o tumor invasivo residual após o tratamento prévio, associada a componentes teciduais que indicam resposta terapêutica, notadamente a fibrose (**FIB**) e a necrose (**NEC**). Para viabilizar a reprodução do **ERTG**, a determinação dos componentes do **LT** considerou parâmetros estabelecidos em protocolos existentes de **ERTG**, as características mapeadas do Modelo **M1** nos conjuntos de validação e de teste, e uma análise da literatura sobre constituintes teciduais relevantes. Embora a mucina acelular (**MUC**) e necrose (**NEC**) sejam reconhecidas como áreas de resposta tecidual, elas não são usualmente incluídas nos protocolos de **ERTG**. Adicionalmente, a **NEC** frequentemente apresenta semelhanças morfológicas com a **FIB**, pois ambas podem representar estágios sequenciais do mesmo processo reparador. Tal similaridade morfológica pode contribuir para uma dificuldade intrínseca de distinção entre essas categorias, de modo que o desempenho de **M1** nessas características tende a refletir a dificuldade dos próprios patologistas. Em virtude dessas considerações, optou-se por estimar os componentes do **LT** como Tumor Invasivo (**TUMIN**), **FIB** e **NEC**. Dada a ausência de consenso definitivo entre os diversos protocolos de Escore de Resposta Tumoral (**TRS**) no trato gastrointestinal, a eficácia do **M1** na reprodução do **GCTRS** foi avaliada ao tratar o **GCTRS** como observador (patologista) independente.

2.3.2 Protocolo para realização do escore de regressão tumoral

Frente à grande diversidade de **ERTs** existentes e à ausência de padronização, optou-se por utilizar o protocolo da instituição **ACCCC** para a reprodução do **ERTG** pelos três patologistas gastrointestinais. Esta decisão foi baseada no fato de que dois (W e G) dos três patologistas designados para esta avaliação independente são afiliados ao **ACCCC**, o que garante uma familiaridade maior com a metodologia em questão. Além disso, esta escolha facilita a comparação entre as novas medições e os resultados do modelo e do laudo original. No protocolo estabelecido pelo **ACCCC**, o **LT** é definido pelos componentes **LT**, **NEC** e **TUMIN**. As proporções estimadas de **TUMIN**, **FIB** e **NEC** para cada lâmina são determinadas pelos patologistas e, em seguida, a média dessas proporções é calculada dividindo-se a soma das estimativas individuais de cada componente por lâmina pelo número total de lâminas analisadas no caso em questão. Sendo assim, o resultado final da estimativa do **ERTG** é uma variável contínua, cuja soma dos três componentes necessariamente é 100%. Segue um exemplo ilustrativo de como o **ERTG** é reportado em laudo anatomo-patológico:

- Tumor Invasivo: 60%
- Fibrose: 30%
- Necrose: 10%

Diante da seleção do protocolo **ACCCC** para a avaliação do **ERTG**, os valores associados ao **ERTG** foram tratados como variáveis contínuas, permitindo, assim, a análise de correlação entre as diferentes medidas, bem como a utilização do escore de *Kendall*. Para a aplicação do coeficiente *Kappa* (κ), as variáveis contínuas foram organizadas em categorias segundo o protocolo de *Becker* (Becker et al. 2003), inicialmente desenvolvido para o estômago, com categorias definidas por percentis. As categorias do protocolo de *Becker* são:

- Categoria 1: 0% de tumor viável;
- Categoria 2: <10% de tumor viável;
- Categoria 3: 10-50% de tumor viável.
- Categoria 4: >50% de tumor viável.

2.3.3 Método de estimativa dos componentes do LT pelo modelo M1

Na realização do **ERTG**, apenas o **LT** é avaliado. Em outras palavras, o **LT** corresponde ao total (100%) da área considerada, e o escore resultante reflete a proporção de tumor invasivo remanescente (**TUMIN**) e da área de resposta (**AR**) dentro do **LT**. Os

componentes da **AR** podem ser subdivididos em **FIB** e **NEC**, a depender do protocolo ou premissa adotados.

No presente estudo, optou-se por subdividir a **AR** em **NEC** e **FIB** durante a avaliação conduzida pelos patologistas. Essas duas classes foram posteriormente agrupadas para compor a **AR** do **LT**. Para reproduzir o **ERTG** a partir das classes identificadas pelo modelo, foram empregadas duas metodologias distintas para a quantificação das proporções dos componentes do **LT**:

1. Método Global (**MGlob**): Esta abordagem consistiu na quantificação dos *tiles* presentes nas lâminas dos pacientes e no cálculo subsequente das proporções de **FIB**, **NEC** e **TUMIN** no conjunto de dados.
2. Método dos Patologistas (**MPatol**): Desenvolvido para emular o processo de cálculo tipicamente adotado pelos patologistas na rotina clínica. Neste método, as proporções estimadas de **TUMIN**, **FIB** e **NEC** para cada lâmina foram calculadas e, em seguida, calculou-se a média dessas proporções, dividida pelo número total de lâminas.

2.3.4 Avaliação da concordância entre observadores

A consistência das estimativas de **ERTG** foi realizada estabelecendo 5 observadores diferentes:

1. '**pat**': Avaliação do **ERTG** do laudo original, realizada em rotina por médico patologista do **ACCCC**;
2. Patologista **W**: um dos três patologistas com especialização em **TGI**, que realizou a avaliação cegada do **ERTG**;
3. Patologista **G**: um dos três patologistas com especialização em **TGI**, que realizou a avaliação cegada do **ERTG**;
4. Patologista **M**: um dos três patologistas com especialização em **TGI**, que realizou a avaliação cegada do **ERTG**;
5. Modelo **M1**: estimativa do **ERTG** pelo modelo, interpretada como se fosse um patologista. Na análise de concordância foram incluídas as duas metodologias para quantificação dos componentes do **LT**, **MGlob** e **MPatol**. A identificação dos dois métodos nos testes se deu da seguinte forma:
 - a) **M1_Glob**: estimativa do **ERTG** pelo modelo **M1** utilizando o método **MGlob**.
 - b) **M1_Patol**: estimativa do **ERTG** pelo modelo **M1** utilizando o método **MPatol**.

Com base na análise comparativa da consistência das medições realizadas por esses cinco observadores, foram determinados 3 conjuntos para testes específicos com diferentes objetivos de avaliação. O primeiro conjunto, denominado **Conc_3Patol**, refere-se à concordância na execução do **ERTG** entre os observadores **M**, **G** e **W** do grupo **Patol_TGI**. Esta avaliação visa identificar a presença de concordância entre os três observadores **M**, **G** e **W** na aferição do **ERTG**. O segundo grupo, designado como **Conc_ERTG**, está relacionado ao grau de concordância do modelo **M1**, incluindo suas variantes **M1_Glob** e **M1_Patol**, em relação aos diferentes **ERTG** disponíveis ('pat'; aferições de **M**, **G** e **W**). Essa análise permite avaliar a performance de **M1** em comparação individualizada com cada patologista, tanto em relação ao laudo original quanto aos membros do grupo **Patol_TGI**. O terceiro conjunto, denominado **Conc_GT**, aborda a comparação exclusiva com os **ERTG** em que **M**, **G** e **W** demonstraram concordância mútua, definida como **GT**. Esta estratégia, embora compare **M1** individualmente com cada patologista do grupo **Patol_TGI**, oferece uma comparação mais agregada, considerando apenas casos em que os três patologistas concordaram entre si. A Tabela 3 ilustra as diferentes combinações de testes.

Tabela 3 – Descrição dos conjuntos de testes para avaliação entre os 5 observadores com respectivo objetivo. Conjunto (Descrição do conjunto para teste específico); Variáveis (Variáveis que compõem o conjunto teste designado); Objetivo (Objetivo de avaliação da concordância no conjunto de teste designado).

Conjunto	Variáveis	Objetivo - Concordância
Conc_3Patol	M, G, W	Membros de Patol_TGI
Conc_ERTG	LO, M, G, W, M1, (M1_Glob, M1_Patol)	M1 e diferentes ERTG
Conc_GT	GT, M1 (M1_Glob, M1_Patol)	M1 e ERTG designados como GT

Fonte: Autoria própria.

2.4 PERFORMANCE DE M1 PARA PREDIÇÃO DE SOBREVIVÊNCIA PÓS-TRATAMENTO CIRÚRGICO

Atualmente, o estadiamento patológico é considerado o principal indicador prognóstico para pacientes diagnosticados com adenocarcinoma gástrico localmente avançado que passaram por ressecção cirúrgica após tratamento neoadjuvante. Além do sistema de estadiamento **TNM**, um laudo anátomo-patológico (**AP**) referente à avaliação de uma peça cirúrgica (**PC**) ressecada inclui diversas outras informações. O conjunto de dados contidos em um laudo de **AP** excluindo as informações referentes a ao **ERTG** foi denominado '**cli**', conforme detalhado a seguir:

1. Medidas macroscópicas do **LT**: durante o processo de amostragem macroscópica da peça cirúrgica, faz parte dos protocolos prever 2 medidas referentes à área do **LT**;

2. Invasões:
 - a) Invasão linfática: presente ou ausente;
 - b) Invasão vascular: presente ou ausente;
 - c) Invasão perineural: ausente ou presente;
3. Tipo histológico do tumor;
4. Grau de diferenciação do tumor: bem diferenciado; moderadamente diferenciado e pouco diferenciado;
5. Classificação de Lauren (Ma et al. 2016; Lauren 1965): difuso ou intestinal;
6. Presença ou ausência de célula em anel de sinete;
7. Linfonodos totais ressecados;
8. Linfonodos comprometidos com metástase;
9. Estadiamento patológico de acordo com a oitava edição do AJCC (Amin et al. 2017):
 - a) Estadiamento patológico T (ypT) - indica a profundidade de infiltração da neoplasia nas camadas da parede do estômago (para detalhes da classificação, vide Tabela 1);
 - b) Estadiamento N (ypN): classificação do comprometimento linfonodal por metástase de adenocarcinoma gástrico (para detalhes da classificação, vide Tabela 1)

Neste estudo hipotetizou-se que o uso de de um conjunto dessas variáveis ('cli') associadas ao **ERTG** apresentaria um maior potencial de predição prognóstica do que o sistema **TNM** ou ao **ERTG** isoladamente. Para testar esta hipótese, foram realizados ensaios no subgrupo de pacientes **SPC**, com o intuito de determinar a combinação mais eficaz de variáveis para a predição da sobrevida em dias subsequentes à ressecção cirúrgica em pacientes de **EST**. Esta sobrevida é definida como o intervalo de tempo, em dias, a partir da data da ressecção cirúrgica, que é o tratamento definitivo após a neoadjuvância, denominado Sobrevida pós-cirurgia (**SPC**). As variáveis consideradas nos respectivos ensaios incluíram o '**pat**', '**cli**', e várias combinações das categorias identificadas por **M1**.

Além da elevada variabilidade interobservador na avaliação do **LT** e da ausência de padronização dos protocolos para essa análise, este estudo propõe-se a investigar se a integração de todos os componentes da lâmina, indo além do **LT**, poderia exercer um papel significativo no poder preditivo prognóstico de **M1**. Em outras palavras, avalia-se se os tecidos além das categorias que constituem o **LT** - tradicionalmente a única região da lâmina analisada no **ERTG** - poderiam contribuir para o poder de predição prognóstica

de **M1**. Esta hipótese partiu da premissa de que os patologistas não se limitam a uma avaliação focada exclusivamente no **LT** durante o processo decisório, mas consideram o contexto integral da lâmina.

Adicionalmente, foi proposto que a utilização de **GNNs**, em vez de modelos mais convencionais como **CNNs**, poderia facilitar a compreensão contextual e topológica, devido à sua arquitetura baseada em grafos. Assim, este subconjunto de testes teve como objetivo analisar diferentes combinações das categorias previstas de **M1**, sem necessariamente seguir protocolos preexistentes, mas sim explorando diversas combinações para testar nossa hipótese sobre o uso da **GNN** na compreensão topológica e contextual, mimetizando o *modus operandi* do médico patologista para obter um valor prognóstico aprimorado.

Para a análise **SPC**, o seguimento dos pacientes foi categorizado em diferentes grupos:

- Morte pela doença (adenocarcinoma gástrico);
- Morte por outra causa (diferente de morte pelo câncer de estômago);
- Vivo com doença;
- Vivo sem doença;

Os pacientes que morreram por outra causa foram excluídos. Para determinar a combinação com maior poder prognóstico, as variáveis foram agrupadas em conjuntos para testes comparativos. Cada combinação (descrita na Tabela 4) foi submetida a diferentes algoritmos de *"Random Forest"* (Tay et al. 2025; Therneau 2024; Kuhn e Vaughan 2025), com o objetivo de esclarecer qual arranjo é mais adequado para a predição de **SPC**. Para avaliar o poder de generalização do modelo, a coorte **SPC** foi dividida em um grupo interno, apenas com casos do **ACCCC**, e em uma coorte externa, com casos da **FMRP** para teste externo.

Tabela 4 – Descrição das combinações de variáveis para estudo de **SPC**.

Combinação	Descrição
cli	Variáveis do laudo anátomo-patológico diferentes do ERTG
pat	ERTG do Laudo Original
model	15 categorias identificadas e quantificadas por M1 ;
cli + pat	'cli' + 'pat'
cli+model	'cli' + quantificação das 15 categorias por M1
mscore	'cli' + apenas as categorias contidas no LT identificadas e quantificadas utilizando M1 ;

Fonte: Autoria própria.

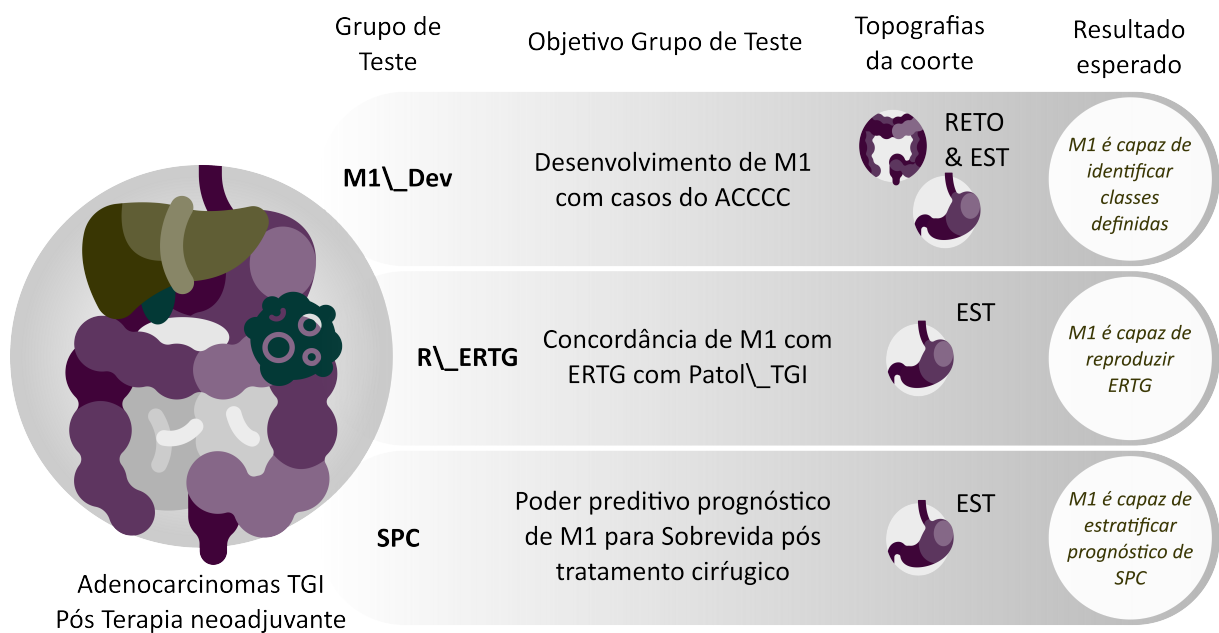
2.4.1 Avaliação do papel das variáveis

Utilizamos a técnica **SHAP** (Lundberg e Lee 2017) para interpretar de forma consistente a importância e o impacto de cada variável nas previsões do modelo. As variáveis provenientes do laudo original foram designadas como '**C__**' e as variáveis do modelo como '**__M**' para permitir a distinção entre sua origem. Foram obtidos o *SHAP value* para explicar cada previsão individual e o $mean(|SHAP\ value|)$ para avaliar a importância geral de uma característica para o modelo como um todo.

3 RESULTADOS

Com o intuito de aprimorar a compreensão dos diversos testes conduzidos para cada subgrupo deste estudo, elaborou-se um infográfico (Figura 19) que salienta os testes executados para cada subgrupo, juntamente com seus respectivos objetivos e resultados esperados.

Figura 19 – Testes realizados por subgrupo



Fonte: Autoria própria.

Legenda: infográfico ilustrando os testes realizados para cada subgrupo da coorte.

3.1 DEFINIÇÃO DA COORTE DO ESTUDO

3.1.1 Distribuição da coorte para diferentes testes

Os casos foram distribuídos nos subgrupos **M1_Dev**; **M1_TestE**; **R_ERTG** e **SPC**, de acordo com as premissas descritas na seção de materiais e métodos. A distribuição dos casos em cada subgrupo com a origem do material e o órgão representado está descrita detalhadamente na tabela 5.

Tabela 5 – Distribuição da coorte em subgrupos para diferentes testes. Grupo (subgrupo de teste realizado); Órgãos (Topografia incluída no subgrupo); Origem (Instituição de origem do material); N° pacientes (N° de pacientes do subgrupo); N° de lâminas (Número de lâminas dos pacientes no subgrupo).

Grupo	Órgãos	Origem	N° pacientes	N° lâminas
M1_Dev	EST e RETO	ACCCC	131	662
M1_TestE	EST	FMRP	27	91
R_ERTG	EST	ACCCC & FMRP	38	207
SPC	EST	ACCCC & FMRP	118	205

Fonte: Autoria própria.

3.2 M1 IDENTIFICA CATEGORIAS HISTOLÓGICAS DE PACIENTES COM CÂNCER GÁSTRICO

3.2.1 Biblioteca de *tiles* e distribuição de subgrupos para treino, validação e testes de M1

Dos 158 pacientes cujas lâminas do **LT** da ressecção cirúrgica já foram digitalizadas, 108 foram inicialmente designados para o desenvolvimento do modelo de . Em sua totalidade, esses pacientes apresentaram um conjunto de 753 lâminas, nas quais foram feitas 46 934 anotações manuais. Essas anotações resultaram em 15 533 392 *tiles* para o treinamento e a validação do modelo. A distribuição da quantidade de anotações e *tiles* gerados por subgrupo está detalhada na tabela 6. A distribuição da quantidade de *tiles* de cada categoria por paciente, para os grupos de treino e validação, é apresentada na Figura 20.

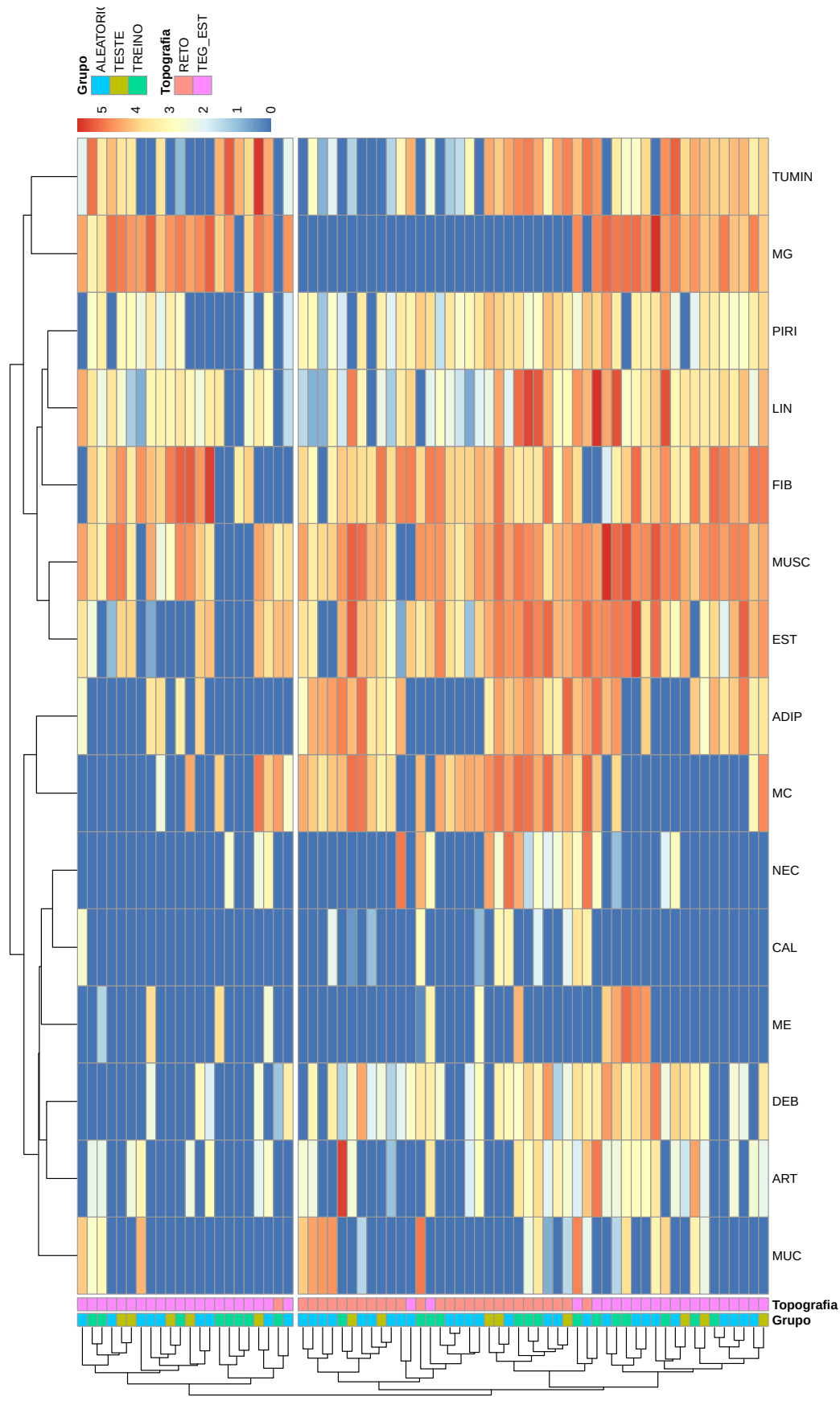
O subgrupo **M1_Dev - Total**, constituindo a totalidade de pacientes de **EST** e **RETO** dedicados para desenvolvimento de **M1**, constitui-se de 108 pacientes com suas respectivas 588 lâminas. Essa coorte foi posteriormente subdividida em grupos para treino (72 pacientes, 404 lâminas), validação (36 pacientes, 179 lâminas) e teste (23 pacientes, 79 lâminas).

Tabela 6 – Distribuição do número de anotações manuais em cada subgrupo de M1. Órgãos (Órgãos incluídos no subgrupo); N° pacientes (Número de pacientes no subgrupo); N° lâminas (Número de lâminas dos pacientes no subgrupo); N° Anot (Número de anotações manuais nas lâminas dos pacientes no subgrupo)

Teste	Órgãos	Origem	N° pacientes	N° lâminas	N° anotações	N° tiles
M1_Dev - Total	EST e RETO	ACCCC	131	662	46.934	13.463.885
M1_Dev - Train	EST e RETO	ACCCC	72	404	26.170	8.101.949
M1_Dev - Val	EST e RETO	ACCCC	36	179	15.399	3.804.491
M1_Dev - TestI	EST e RETO	ACCCC	23	79	5.365	1.557.445
M1_TestE	EST	FMRP	27	91	6.539	2.069.507

Fonte: Autoria própria.

Figura 20 – Mapa de calor *tiles* em treino e validação de M1_Dev



Fonte: Autoria própria.

Legenda: mapa de calor da quantidade de tiles de cada categoria por paciente nos grupos de treino e validação de M1_Dev (em paisagem).

3.2.2 Treinamento, Validação e Teste Interno de M1

Foi realizado um extenso estudo para identificar a melhor combinação de arquitetura de **GNN** utilizando diferentes **MFs** para extração de features. Os modelos foram treinados e validados com 7 diferentes arquiteturas de **GNN** utilizando 2 diferentes **MFs**, UNI2 (Chen et al. 2024c) (treinado em 300 milhões de parâmetros com *Embed Size* de 1024) e Virchow 2 (Zimmermann et al. 2024b) (treinado em 632 milhões de parâmetros com *Embed Size* de 3584). O melhor desempenho foi observado com a arquitetura **GIN** em conjunto com o Virchow 2 no grupo de validação (**M1_Dev - Val**). Os resultados detalhados de F1-Score do grupo **M1_Dev - Val** para cada combinação de **GNN** com UNI 2 e Virchow 2 são apresentados na Tabela 7.

Tabela 7 – Comparação entre desempenho de **MF** para cada arquitetura de **GNN**

GNN	UNI 2	Virchow2
gcn	0,8619669	0,8290138
cheb	0,9169112	0,9115092
gat	0,8571190	0,8707201
gen	0,9289766	0,9311008
sage	0,9095424	0,9354346
gin	0,9284265	0,9401164
arma	0,9275937	0,9132741

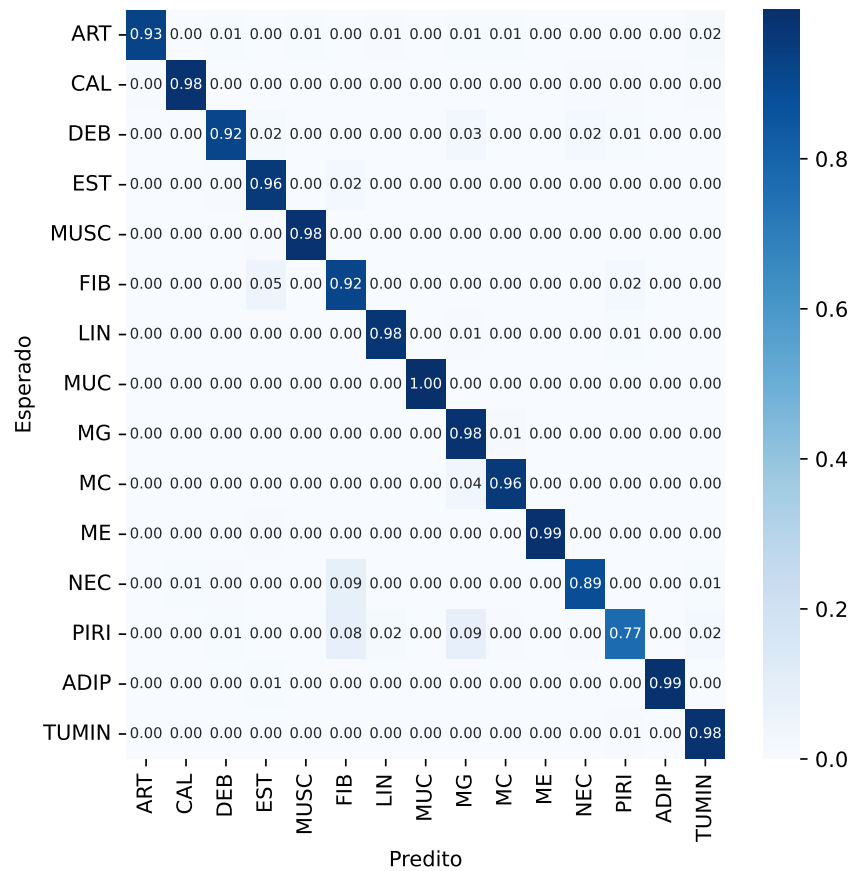
Fonte: Autoria própria.

As avaliações de performance realizadas com o grupo de validação revelaram que a arquitetura **GIN** juntamente com o modelo Virchow 2 apresentou o melhor desempenho para este subgrupo específico. Procedeu-se, então, com a construção da matriz de confusão (Figura 21) do subgrupo de validação com o intuito de avaliar o comportamento do modelo **M1** na identificação das categorias em lâminas de **TGI**. A análise da **MC** revela uma tendência do modelo **M1** em confundir as classes **PIRI**, **NEC** e **FIB** entre si. De fato, observa-se que o modelo **M1** classifica **NEC** e **PIRI** como **FIB**. Adicionalmente, a predição de **PIRI** como **NEC** também pode ser percebida.

Após treinamento e validação, foram realizados testes no grupo de teste interno (**M1_Dev - TestI**). Frente à melhor performance do Virchow 2 na extração de features no grupo de validação, optou-se por seguir apenas com os testes com o Virchow 2 nessa coorte. Os resultados detalhados dos testes do grupo **M1_Dev - TestI** para as diferentes combinações de **GNN** com Virchow 2 são apresentados na Tabela 8.

De forma geral, o modelo **M1** performou bem no grupo de teste interno, com F1-Score mais baixo de 0,79 e todos os resultados de AUROC acima de 0,92. Especificamente, apresentou um F1-Score de 0,87 com a **GNN GIN**. Com esses testes, conclui-se que **M1** é capaz de identificar as categorias presentes em um **LT** de lâminas de pacientes pós-tratamento neoadjuvante com adenocarcinoma do **TGI**.

Figura 21 – Matriz de confusão (M1_Dev - Val)



Fonte: Autoria própria.

Legenda: matriz de confusão normalizada do grupo de validação (M1_Dev - Val), evidenciando a tendência do modelo M1 de maior confusão entre as classes PIRI, NEC e FIB.

Tabela 8 – Performance do grupo M1_Dev - TestI. M1_Dev - TestI (grupo para teste Interno). AUROC (Area Under the Receiver Operating Characteristic); ACC (Acurácia).

GNN	AUROC	ACC	F1Score	Recall	Precision
cheb	0,9283353	0,9276202	0,8963674	0,9276202	0,8930355
arma	0,9296099	0,9154287	0,8854218	0,9154287	0,8991919
gin	0,9282554	0,9213023	0,8748363	0,9213023	0,8729081
gcn	0,9237598	0,8706564	0,8537281	0,8706564	0,8706958
gat	0,9219110	0,7963595	0,7934688	0,7963595	0,8140035
gen	0,9276776	0,8942051	0,8486735	0,8942051	0,8689204
sage	0,9276844	0,8991560	0,8514756	0,8991560	0,8877797

Fonte: Autoria própria.

3.2.3 Teste Externo de M1

Após o desenvolvimento de **M1** com a coorte do **ACCCC**, com casos de **EST** e **RETO** (conjunto de pacientes **M1_Dev**), passou-se ao teste de **M1** em uma coorte externa da **FMRP**, composta exclusivamente por **EST**. Foram realizados testes no sub-grupo designado como **M1_Dev - TestE**, novamente com diferentes arquiteturas de **GNN** utilizando UNI 2 e Virchow 2. Optou-se por incluir UNI 2 para extração de *features* nesse cenário, a despeito do melhor desempenho de Virchow 2 no grupo de validação (**M1_Dev - Val**), por se tratar de lâminas provenientes de instituição externa. A **GNN** com a arquitetura **GIN** apresentou F1-Score de 0,92 em conjunto com o Virchow 2. Os resultados detalhados dos testes do grupo **M1_Dev - TestE** para as diferentes combinações de **GNN** com Virchow 2 e UNI 2 são apresentados nas Tabelas 9 e 10.

Tabela 9 – Comparação de performance de 7 **GNNs** distintas com **Virchow 2** para extração de features. AUROC (Do inglês *Area Under the Receiver Operating Characteristic* - Área Sob a Curva Característica de Operação do Receptor); ACC (Acurácia); M-CK (Multiclass CohenKappa); M-MCC (Multiclass Matthews Corr Coef).

GNN	AUROC	ACC	F1-Score	Recall	Precision	M-CK	M-MCC
arma	0,998	0,921	0,933	0,921	0,953	0,958	0,958
gin	0,997	0,913	0,918	0,913	0,940	0,961	0,961
cheb	0,997	0,899	0,909	0,899	0,940	0,956	0,956
gen	0,995	0,869	0,889	0,869	0,939	0,951	0,952
gat	0,995	0,855	0,876	0,855	0,921	0,937	0,938
gcn	0,995	0,862	0,875	0,862	0,909	0,946	0,946
sage	0,996	0,851	0,862	0,851	0,932	0,954	0,955

Fonte: Autoria própria.

Tabela 10 – Comparação de performance de 7 **GNNs** distintas com **UNI** para extração de features. AUROC (Do inglês *Area Under the Receiver Operating Characteristic* - Área Sob a Curva Característica de Operação do Receptor); ACC (Acurácia); M-CK (Multiclass CohenKappa); M-MCC (Multiclass Matthews Corr Coef).

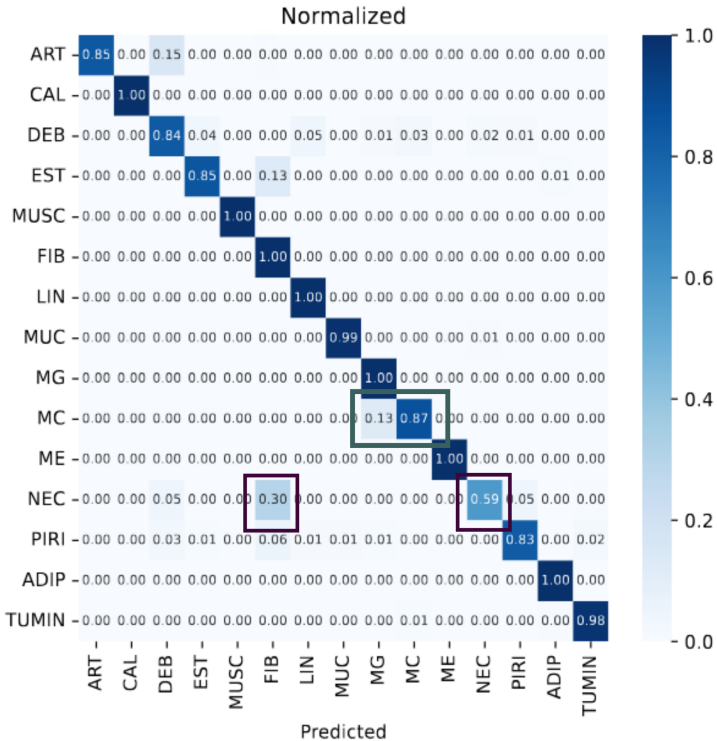
GNN	AUROC	ACC	F1-Score	Recall	Precision	M-CK	M-MCC
gen	0,993	0,862	0,863	0,862	0,922	0,951	0,951
gin	0,995	0,853	0,850	0,853	0,897	0,942	0,942
sage	0,992	0,858	0,846	0,858	0,872	0,944	0,944
arma	0,987	0,848	0,845	0,848	0,909	0,928	0,929
gat	0,991	0,829	0,833	0,829	0,890	0,932	0,933
cheb	0,994	0,830	0,832	0,830	0,874	0,918	0,919
gcn	0,990	0,807	0,803	0,807	0,851	0,918	0,919

Fonte: Autoria própria.

Gerou-se uma matriz de confusão desse grupo a fim de melhor compreender o padrão de comportamento de **M1** para a identificação das classes em coorte externa, conforme os resultados detalhados na Figura 22. A matriz de confusão (Figura 22) reforça o desempenho do modelo, de forma semelhante aos resultados obtidos no grupo de validação com coorte interna, com tendência a exibir maior confusão entre as classes **DEB**, **PIRI** e **NEC**, corroborando os resultados discutidos anteriormente. Contudo, no grupo com as lâminas da coorte **FMRP** houve maior confusão entre as classes Mucosa Gástrica (**MG**) e Mucosa Colônica (**MCo**). Ainda, ao observar a matriz de confusão (Figura 22) e compará-la à obtida com o desempenho no grupo de validação, nota-se uma piora significativa na distinção entre Necrose (**NEC**) e Fibrose (**FIB**). Esse cenário pode ser explicado por um único caso (identificado como CASO_NEC) com extensa necrose densa.

Embora o grupo de teste externo seja composto por 91 lâminas com 6 539 anotações, poucas lâminas apresentam áreas significativas de necrose para anotação e, consequentemente, teste. Esse caso em específico apresentava uma necrose já densa e "colagenizada", facilmente confundida com fibrose. As anotações desse caso configuram 100% do total de anotações de **NEC** do grupo de teste externo.

Figura 22 – Matriz de confusão da coorte externa de teste

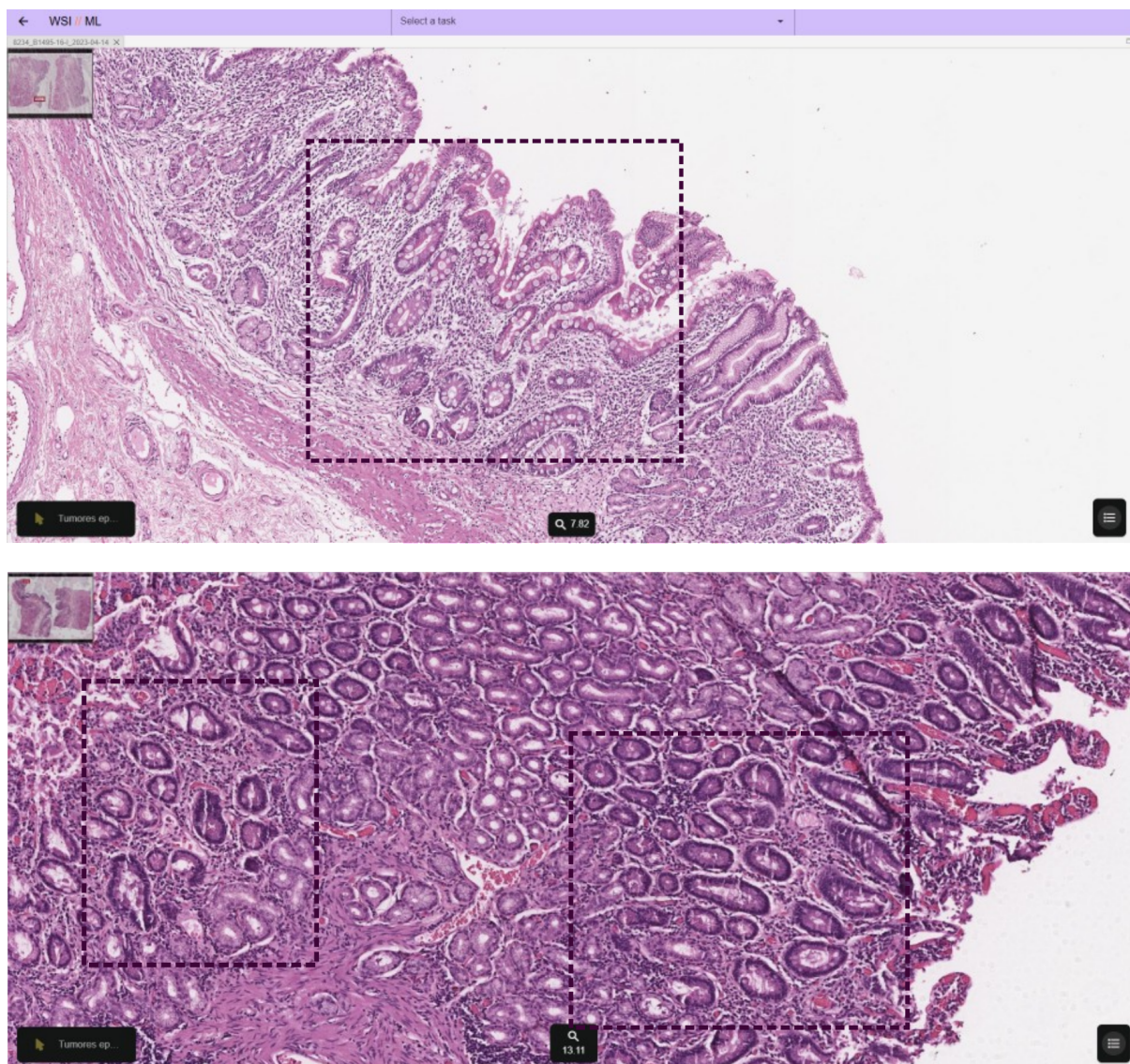


Fonte: Autoria própria.

Legenda: Matriz de confusão da coorte externa de teste, evidenciando a mesma tendência que a validação do modelo **M1** com a coorte interna: maior confusão entre as classes **PIRI**, **DEB**, **NEC** e **FIB**. Maior confusão entre **MCo** e **MG**. Além disso, houve maior dificuldade na distinção entre **NEC** e **FIB** (ressaltado em vermelho), devido ao caso CASO_NEC com 2 lâminas.

Para melhor compreender essa confusão entre **MCo** e **MG**, as lâminas que apresentavam a maioria dos tiles identificados de forma incorreta foram analisadas uma a uma, que evidenciaram uma característica particular das lâminas da instituição parceira, nas quais havia muitas áreas de metaplasia intestinal, sendo esta muitas vezes incompleta. Estipula-se que a arquitetura mais irregular e, principalmente, a maior variação em tamanhos das células caliciformes (ou células em *goblet*) da metaplasia incompleta (Shah et al. 2020) podem ter contribuído para a dificuldade do modelo em diferenciar **MG** de **MCo** nesses cenários específicos (Figura 23).

Figura 23 – Metaplasia intestinal em mucosa gástrica



Fonte: Autoria própria.

Legenda: exemplos de epitélio gástrico com áreas de metaplasia intestinal completa e incompleta nos casos de **FMRP**. As figuras ressaltam áreas de metaplasia intestinal.

3.3 M1 É CAPAZ DE REPRODUZIR ERTG

3.3.1 Concordância entre observadores

A avaliação da concordância foi realizada entre cinco observadores ('pat', **M**, **G**, **W** e **M1**, sendo este último subdividido em **M1_Global** e **M1_Patoll**, de acordo com a divisão dos conjuntos de testes **Conc_3Patol**, **Conc_ERTG**, **Conc_GT**). Em todos os conjuntos de testes, a avaliação da consistência das medidas dos observadores foi conduzida por meio de uma análise estatística pareada. Nesta análise, investigou-se a associação linear entre as variáveis contínuas e os níveis de concordância, utilizando-se o escore de *Kendall* para variáveis contínuas e o coeficiente *kappa*, que transforma os dados conforme o **ERTG** de *Becker:kappa* (*Becker*).

No conjunto de testes designados como **Conc_3Patol**, cujo objetivo foi avaliar a concordância entre os três patologistas integrantes do grupo **Patol_TGI**, observou-se uma associação linear forte, mas com níveis variados de concordância. Os coeficientes de correlação de Pearson (*r*), que quantificam a relação linear, apresentaram valores consistentemente elevados, variando de 0,70 (**M** vs **G**) a 0,78 (**W** vs **G**). O erro de medida foi igualmente baixo, com a Raiz do Erro Quadrático Médio (RMSE) exibindo valores entre 0,24 e 0,25. No entanto, a concordância ordinal, avaliada pelo coeficiente de *Kendall*, situou-se na faixa de moderada a boa, sendo mais robusta nas comparações envolvendo o observador **W** (**W** vs **G**: 0,69; **W** vs **M**: 0,64). Os valores de todos os testes, com precisão de duas casas decimais, estão descritos na tabela 11.

Tabela 11 – Concordância das avaliações entre os três patologistas. P1 (Patologista 1 na comparação); P2 (Patologista na comparação).

P1	P2	mse	r ²	mae	mse	<i>r</i>	<i>kendall</i>	κ (>0.5)	κ - <i>Becker</i>
M	G	0,06	0,12	0,19	0,25	0,70	0,56	0,43	0,35
W	M	0,06	0,51	0,19	0,24	0,75	0,65	0,60	0,58
W	G	0,06	0,47	0,16	0,25	0,78	0,69	0,59	0,45

Fonte: Autoria própria.

O coeficiente *Kappa* de *Becker*, ajustado pela probabilidade de concordância ao acaso e aplicado às categorias do protocolo de escore **ERTG** de *Becker*, demonstrou uma reprodutibilidade limitada, com valores variando entre as classificações Moderada a Razoável. A concordância categórica foi classificada como Moderada para o par de patologistas **W** e **M** (0,58) e como Razoável para o par **M** e **G** (0,35). Esta discrepância entre o alto coeficiente de correlação de *Pearson* e o menor valor do *Kappa* sugere que, embora os patologistas **M**, **G** e **W** apresentem uma tendência similar em suas classificações, alcançar uma concordância categórica exata é desafiador, com o patologista **W** mostrando-se o mais consistente em relação aos seus pares.

Na avaliação do grupo **Conc_ERTG**, cujo principal objetivo foi avaliar a concordância das métricas **M1** (**M1_Glob** e **M1_Patol**) em relação às diversas estimativas de **ERTG** fornecidas pelos patologistas designados como **M**, **G**, **W** e **'pat'**, constatou-se que o observador **W** demonstrou uma maior congruência com os modelos propostos. O observador **W** obteve os mais elevados coeficientes de correlação de *Pearson*, atingindo até 0,88 para a métrica **M1_Glob**, assim como os coeficientes de correlação de *Kendall*, alcançando até 0,72, quando comparado às abordagens estudadas. Este alinhamento foi corroborado por um alto coeficiente de determinação (R^2) observado para **W**, com valores chegando até 0,69, o que indica que os modelos propostos explicam uma proporção significativa da variância observada nas avaliações de **W**. Em contrapartida, os resultados de concordância observados para **M**, e de maneira mais pronunciada para **'pat'**, variaram entre níveis de Pobre a Razoável. O coeficiente *Kappa* (*Becker*) foi consistentemente baixo para todos os pares de comparações envolvendo os modelos, atingindo um valor mínimo de 0,08 na comparação entre **'pat'** e **M1_Glob**, o que classifica essa concordância como Pobre/Leve, sugerindo que o acordo se aproxima do aleatório. Os valores negativos do coeficiente de determinação R^2 observados para **M** (variando até -0,39) e para **'pat'** (até -0,24) indicam que as abordagens **M1_Glob** e **M1_Patol** não são adequadas para prever ou explicar a variância nas avaliações desses observadores, apresentando um ajuste inferior ao de um modelo nulo.

Tabela 12 – Concordância das avaliações entre os diferentes métodos de **M1** com os três patologistas.

MM1	P	mse	r ²	mae	rmse	r	kendall	kappa (>.5)	κ - Becker	κ (agree >.5)
M1_Glob	G	0,11	0,03	0,22	0,32	0,69	0,59	0,50	0,31	0,44
M1_Patol	G	0,11	-0,02	0,23	0,33	0,67	0,55	0,44	0,26	0,38
M1_Glob	W	0,04	0,69	0,12	0,19	0,88	0,72	0,78	0,36	0,35
M1_Patol	W	0,04	0,64	0,14	0,21	0,87	0,70	0,70	0,30	0,29
M1_Glob	M	0,08	-0,36	0,23	0,29	0,71	0,65	0,52	0,19	0,18
M1_Patol	M	0,08	-0,39	0,24	0,29	0,70	0,64	0,44	0,18	0,17
M1_Glob	'pat'	0,12	-0,25	0,27	0,35	0,65	0,56	0,33	0,08	-0,14
M1_Patol	'pat'	0,12	-0,21	0,27	0,35	0,67	0,57	0,35	0,10	-0,27

Fonte: Autoria própria.

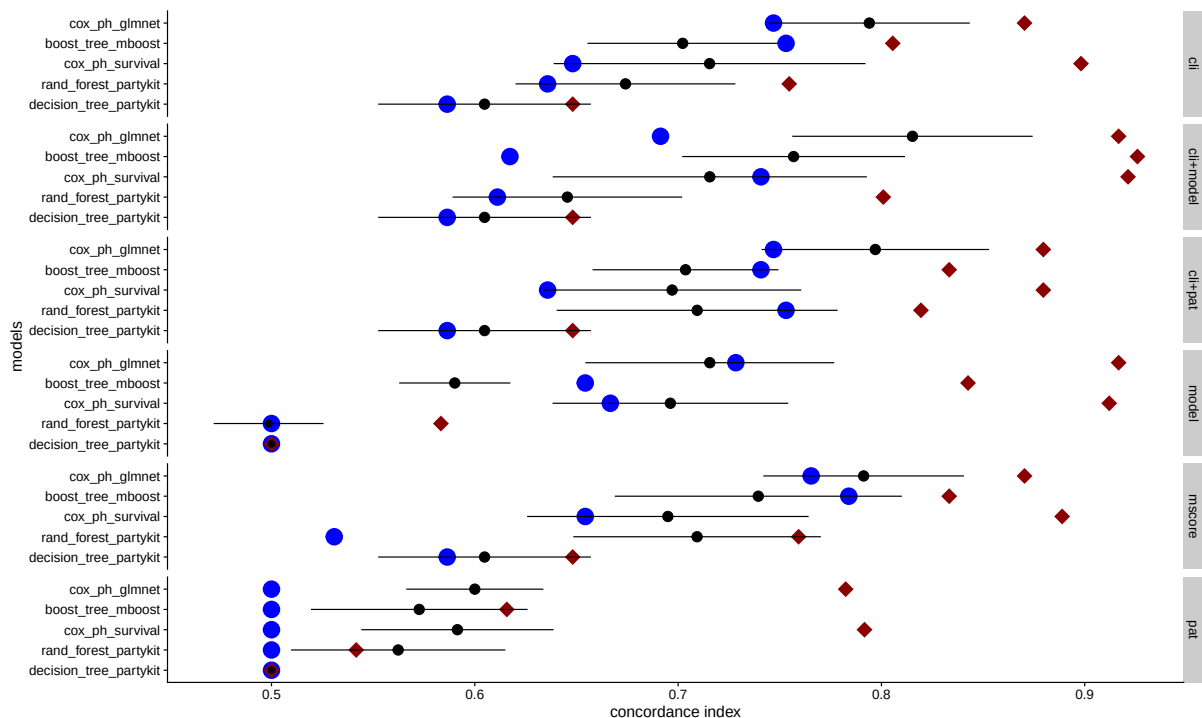
A terceira análise referente à concordância (conjunto **Conc_GT**), centrada na comparação entre o coeficiente *Kappa* ajustado pela chance conforme o protocolo de *Becker* (*kappa* (*Becker*)) e o coeficiente *Kappa* calculado para o subconjunto de alta concordância, determinados quando *kappa* (*Becker*) > 0,5 (*kappa* (agree >0,5)), demonstra padrões diversos de reprodutibilidade em função da qualidade da concordância no grupo **Patol_TGI**. Para o Patologista **G**, a métrica *kappa* (agree >0,5) (0,44 e 0,38) superou o *kappa* (*Becker*) (0,30 e 0,26) em ambas as abordagens (**M1_Glob** e **M1_Patol**).

Em contraste, a concordância para os Patologistas **W** e **M** permaneceu relativamente estável ou experimentou uma ligeira diminuição neste subconjunto (por exemplo, **M** vs **M1_Glob**: 0,19 para *Becker* vs 0,18 para *kappa* ($>0,5$)). O achado mais significativo ocorreu na comparação com '**pat**'. Embora o *kappa* (*Becker*) para '**pat**' já fosse extremamente baixo (0,08 e 0,10), a métrica *kappa* (agree $>0,5$) resultou em valores negativos (até -0,27 para **M1_Patoll**).

3.4 M1 COM DADOS ANÁTOMO - CLÍNICOS MELHORA O PODER PREDITIVO PROGNÓSTICO

A análise descritiva dos resultados de predição de sobrevida utilizando modelos de *Random Forest* (**RF**), conforme detalhado na figura 24, demonstra variações notáveis no poder preditivo (*Concordance Index* - **CI**) em função dos conjuntos de variáveis utilizados. Os modelos **RF** foram avaliados em seis combinações distintas de features: '**cli**', '**pat**', **model** (15 categorias quantificadas por **M1**), '**cli** + **pat**', '**cli** + **model**', e '**mscore**' ('**cli**' + apenas a categorias contidas no **LT** identificadas e quantificadas utilizando **M1**).

Figura 24 – Performance modelos **RF** para predizer **SPC**



Fonte: Autoria própria.

Legenda: Performance da combinação de classes estimadas por **M1** com dados ('**cli**') com modelos **RF** para predizer **SPC**. **CI** (Índice de Concordância); Diamante Vermelho: Representa a performance no teste interno (oriundos do **ACCCC**); Bola Azul: Representa a performance no teste externo (em dados de **FMRP**); Ponto Preto: Representa a performance da validação *k-fold*, uma técnica de validação cruzada que mede a estabilidade do modelo.

Os achados revelam que o maior poder de predição foi alcançado pelas combinações que integram o modelo **M1** com suas 15 categorias ('**model**') e dados '**cli**'. Especifica-

mente, a combinação '**cli** + **model**' apresentou o desempenho mais robusto para os modelos de **RF**, atingindo valores de **CI** próximos ou acima de 0,90 no teste interno (diamante vermelho). Embora o desempenho no teste externo (bola azul) seja ligeiramente inferior (aproximadamente 0,85), ele ainda representa o melhor resultado de generalização. Em contraste, as combinações '**mscore**' e '**pat**' resultaram nos valores de **CI** mais baixos para os modelos de **RF**. A combinação '**pat**' mostrou o desempenho mais fraco, com os resultados de *k-fold validation* (ponto preto) e teste externo (bola azul) se aproximando de 0,50 a 0,55 indicando poder preditivo marginalmente superior ao acaso. As combinações que incluem apenas '**cli**') ou apenas as categorias identificadas por **M1** ('**model**') situaram-se em um nível intermediário, com o desempenho no teste interno consistentemente superando os resultados de generalização (teste externo e validação *k-fold*).

A interpretação desses resultados sugere uma forte dependência do poder preditivo do **RF** em relação à abrangência e complexidade do conjunto de *features*. O desempenho superior da combinação **cli** + **model**' indica que a inclusão das categorias quantificadas por **M1**, em conjunto com os dados clínicos, fornece as informações mais valiosas para a predição de sobrevida. Por outro lado, o desempenho fraco das combinações ('**pat**' e **mscore** ('**cli**')) sugere que essas variáveis isoladamente são insuficientes para construir um modelo robusto de predição. Além disso, observa-se que, em quase todas as combinações, a performance no teste interno (diamante vermelho) é notavelmente superior à performance no teste externo (bola azul) e na validação *k-fold* (ponto preto). Este padrão pode indicar que **M1** esteja sujeito a certo grau de *overfitting* ou sobreajuste aos dados internos.

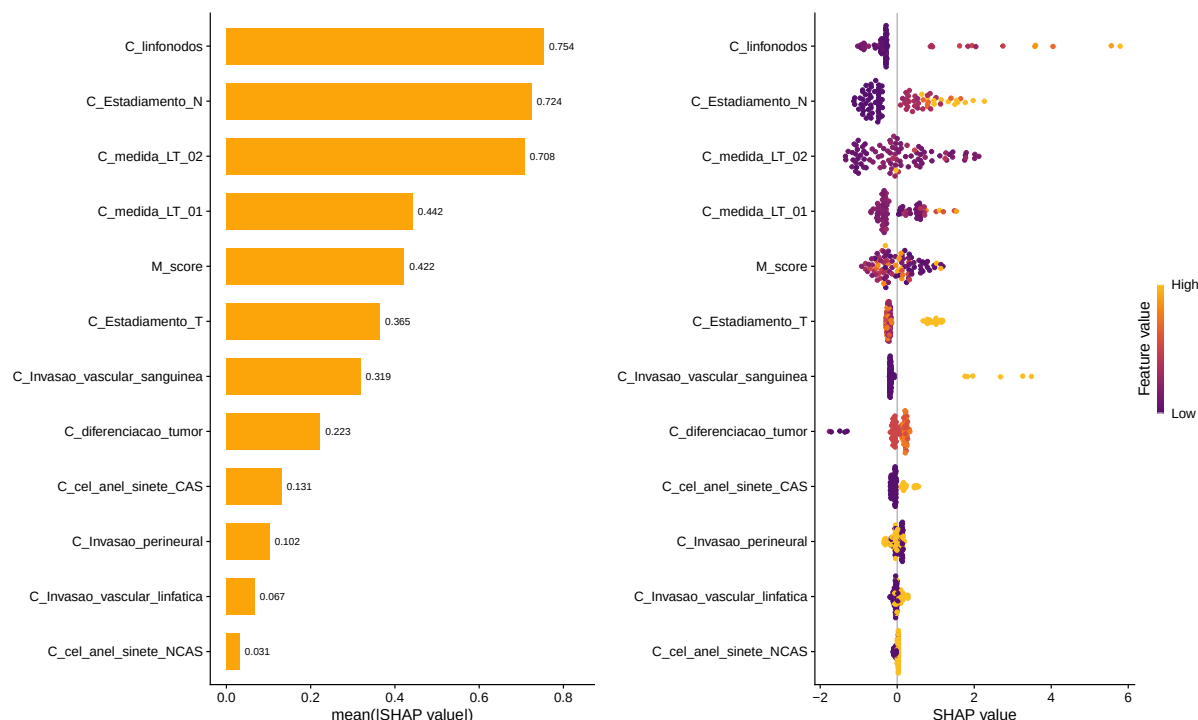
3.4.1 Avaliação da importância das variáveis

A metodologia **SHAP** (*SHapley Additive exPlanations*) foi aplicada para avaliar a contribuição e a direção do impacto de cada variável na predição de sobrevida utilizando o modelo **RF**. A análise descritiva dos valores médios de **SHAP** (Figura 25) revela uma hierarquia claramente definida de importância das características. As variáveis mais influentes, indicadas pelos valores médios de **SHAP** mais elevados, estão consistentemente associadas ao status dos linfonodos e às métricas quantificadas do leito tumoral (**LT**).

As variáveis predominantes são status linfonodal (variável '**cli**' C_linfonodos - valor médio de **SHAP**: 0,754) e variável '**cli**' estadiamento N (C_Estadiamento_N - valor médio de **SHAP**: 0,724), seguidas de perto pelas variáveis '**cli**' de medida macroscópica do **LT** (C_medida_LT_02 - 0,708 e C_medida_LT_01 -0,442). Em contraste, variáveis clinicopatológicas frequentemente citadas, como invasão vasuclar linfática e células (0,067) ausência de células em anel de sisnete (N_CAS - 0,031), demonstram o menor impacto na predição de sobrevida do modelo **RF**.

Ao analisar o gráfico de distribuição de **SHAP** (diagrama de pontos), observa-se um padrão no qual altos valores (indicados pelas cores amarela/laranja) nas variáveis de maior relevância (status linfonodal (C_linfonodos), estadiamento N (C_Estadiamento_N)

Figura 25 – Importância variáveis pelo método **SHAP**



Fonte: Autoria própria.

Legenda: Análise gráfica da importância das variáveis pelo método **SHAP**.

e medidas macroscópicas do **LT** (C_medida_LT_02) contribuem predominantemente para um valor de **SHAP** positivo, influenciando a predição em direção a um determinado resultado. Em contraste, valores baixos (indicados pelas cores roxo/borgonha) contribuem para um valor de **SHAP** negativo.

A análise interpretativa dos resultados obtidos através do método **SHAP** sugere que o poder preditivo do modelo **RF** é majoritariamente sustentado por variáveis extraídas do conjunto de dados '**cli**'. As características de maior relevância, C_linfonodos (0,75) e C_Estadiamento_N (0,72), são variáveis clássicas de estadiamento que pertencem ao conjunto '**cli**', corroborando que o *status* linfonodal permanece como o principal determinante prognóstico no modelo de análise de sobrevida. A importância elevada das medidas macroscópicas do **LT** (C_medida_LT_02 - 0,70 e C_medida_LT_01 - 0,44), em associação com o status linfonodal, indica que o modelo **RF** confere prioridade ao estadiamento nodal e às métricas dimensionais macroscópicas do **LT** em detrimento de outras características.

De importância, a quinta variável significativa é o **mscore**, que representa o '**cli**', exclusivamente associado a categorias presentes no **LT**, identificadas e quantificadas pelo método **M1**. Em outras palavras, apesar do status linfonodal e das medidas do **LT** serem as variáveis de maior poder prognóstico de forma independente, a interação entre **M1** e '**cli**' também se destaca como uma combinação de considerável importância na predição prognóstica.

Adicionalmente, o mapeamento **SHAP** (cores quentes, ou altos valores das *features*, tendendo a valores **SHAP** positivos) sugere que o modelo **RF** associa corretamente piores condições clínicas (por exemplo, maior número de linfonodos envolvidos ou maior medida quantificada do leito tumoral) a um desfecho prognóstico pior. Variáveis clinicopatológicas consideradas marcadores de agressividade tumoral, como C_invasao_vascular_linfatica (0,06) e C_cel_anel_sinete_NCAS (0,03), apresentaram impacto preditivo marginal. Isto indica que a informação prognóstica dessas *features* de menor relevância pode já estar totalmente encapsulada nas variáveis de estadiamento **cli** de alta hierarquia (C_linfonodos, C_Estadiamento_N) e nas medidas de tamanho de **LT**, tornando-as redundantes dentro do contexto deste modelo multivariado específico. Em resumo, a capacidade de predição de sobrevida do modelo *Random Forest* depende fortemente da informação clínica categórica e quantitativa (**'cli'**), com particular ênfase no status linfonodal e nas dimensões macroscópicas do **LT**.

4 DISCUSSÃO

O câncer gástrico, sendo o quinto câncer mais comum e a quarta principal causa de mortes por câncer globalmente, apresenta elevadas taxas de mortalidade e incidência crescente entre adultos mais jovens, com baixas taxas de sobrevida devido ao diagnóstico tardio em estágios avançados (Heer et al. 2020; Morgan et al. 2022; Su e Yu 2022). Embora a ressecção cirúrgica seja uma estratégia central, o tratamento neoadjuvante é recomendado para tumores localmente avançados, visando o downstaging e aumentando a probabilidade de uma ressecção curativa (Cunningham et al. 2006; Su e Yu 2022). No entanto, a avaliação pós-tratamento através do **ERTG** é problemática, devido à falta de padronização, alta variabilidade interobservador (Chetty et al. 2012; Pai e Pai 2018). Além disso, a despeito de seu potencial uso clínico, seu poder preditivo é inferior ao estadiamento patológico (Maas et al. 2010; Campos-Lobato et al. 2011; Stark et al. 2019).

O uso de IA tem se mostrado uma abordagem promissora na análise de **WSI**, oferecendo suporte ao diagnóstico e contribuindo para avaliações prognósticas mais precisas (Pantanowitz et al. 2025; Rashidi et al. 2025). Esse desenvolvimento foi acelerado com o surgimento dos **MF**, que trazem grandes promessas para a **PC** (Chen et al. 2024a). Dentro desse contexto e frente à premissa de que **GNNs** podem capturar melhor as dependências espaciais intrincadas de uma imagem histológica devido a sua capacidade de aprendizado topológico (Wu et al. 2023; Alzoubi et al. 2024; Brussee et al. 2024), optou-se pelo desenvolvimento de um modelo de **GNN** com comparação de diferentes **MF** para reprodução do **ERTG**. Hipotetizou-se que essa abordagem mitigaria as limitações inerentes dos **ERTG** e que esse modelo associado a variáveis anátomo-clínicas teria maior poder de predição prognóstica do que o estadiamento isolado.

A fim de testar essa premissa, desenvolveu-se **M1**, um modelo de **GNN** combinado com **MF** para extração de features capaz de indetificar as categorias histológica de pacientes com câncer de estômago pós terapia neoadjuvante. Os resultados obtidos a partir da validação indicaram que a arquitetura de redes neurais gráficas **GIN** (Xu et al. 2018), em combinação com o extrator de características Virchow 2 (Zimmermann et al. 2024a), exibiram o melhor desempenho na identificação das categorias histológicas mencionadas. Considerando que o desempenho das **GNNs** é altamente dependente da qualidade das *features*, o emprego de **MF** pré-treinados em patologia, como o Virchow 2 e UNI 2, provavelmente forneceu as representações de nó robustas necessárias para o sucesso do modelo. Postula-se que essa combinação foi fundamental para a robustez da performance do **M1**, evidenciada por um *F1-score* de 0,94 com a **GNN GIN** (Xu et al. 2018) no grupo de validação, 0,87 no teste interno e 0,92 externo, sempre apresentando uma AUROC superior a 0,92, o que ressalta a significância desses resultados. Esses resul-

tados corroboram que **M1** é competente em identificar as categorias histológicas de **WSI** de pacientes com câncer de estômago pós terapia neoadjuvante.

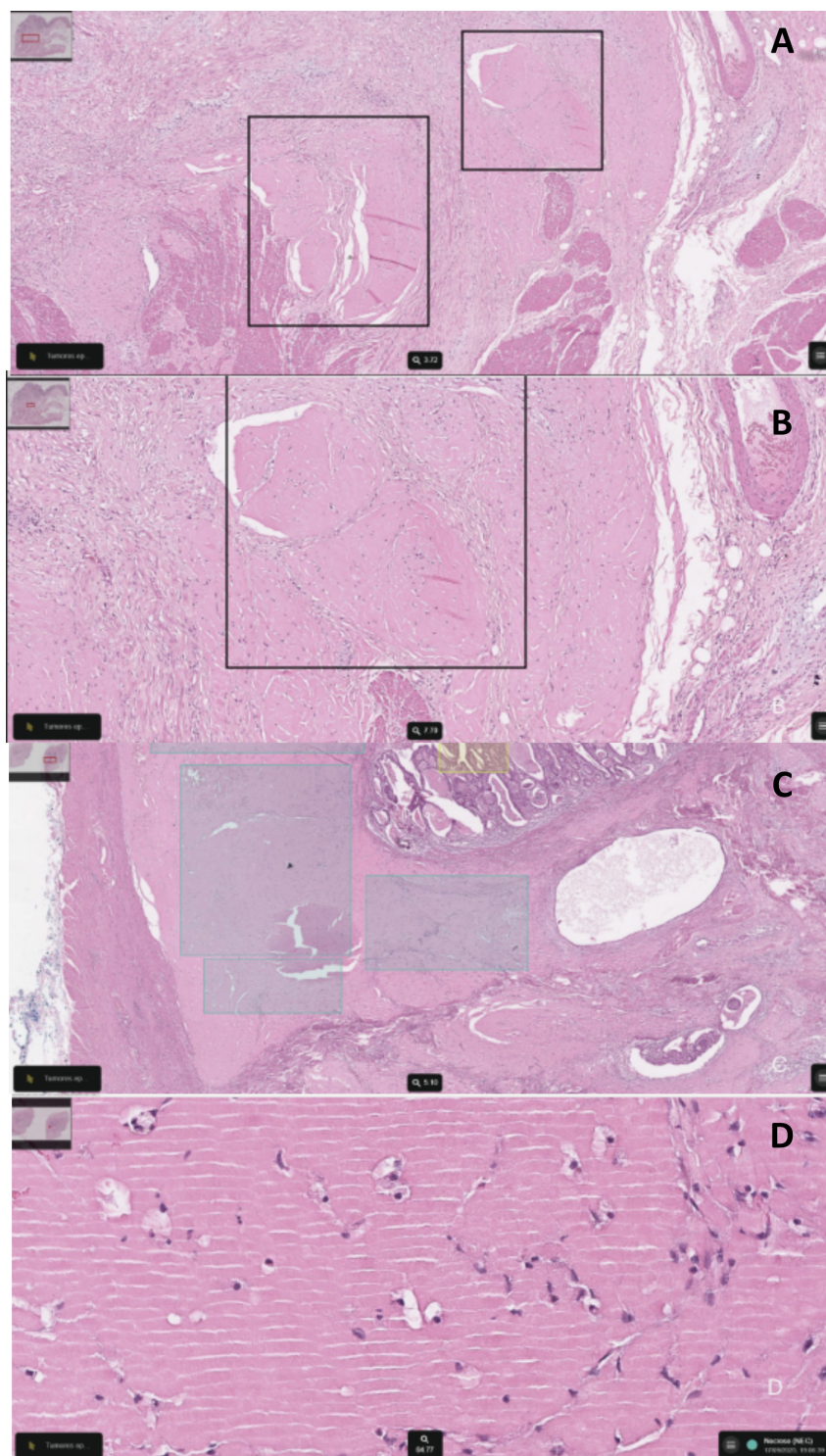
A despeito da boa performance de **M1** para identificar as categorias do **WSI**, uma análise aprofundada das predições geradas por **M1**, a partir das matrizes de confusão, revelou uma tendência significativa de confusão entre as classes **PIRI**, **NEC** e **FIB**. A fim de obter uma avaliação mais precisa do comportamento do modelo nesse cenário, bem como para uma compreensão mais aprofundada das implicações do desempenho de **M1** para a estimativa geral do **ERTG**, foi realizada uma análise visual das predições das classes pelo modelo **M1**. Este processo incluiu a extensão das predições sobre imagens digitalizadas, utilizando a plataforma desenvolvida pelo grupo (Martins et al. 2023). Esta análise foi conduzida com o objetivo de identificar e detalhar potenciais padrões de erro, permitindo uma compreensão mais acurada do funcionamento do modelo **M1**, facilitando a identificação de tendências de viés do modelo **M1** e a correlação dos erros detalhados nas **MCs** com os processos patológicos subjacentes observados em lâminas de pacientes submetidos a tratamento neoadjuvante.

A partir dessa análise, identificou-se que as regiões de maior confusão entre **PIRI**, **NEC** e **FIB** ocorreram em três cenários principais. No primeiro cenário, pelo menos duas das três classes confundíveis estavam presentes na área de **LT** de forma mesclada, sem limites claros de separação (conforme ilustrado na Figura 26).

No segundo cenário, as classes confundíveis estavam representadas por grandes anotações de áreas com um padrão histológico de difícil classificação dicotômica, ou seja, onde o aspecto histológico entre as classes confundíveis é limítrofe. A Figura 27 ilustra o caso da lâmina CASO_NEC de **FMRP**, que foi responsável pela maioria dos erros entre **NEC** e **FIB** observados na matriz de confusão do grupo **M1_Dev - TestE**. Esta área específica refere-se a uma área cicatricial pós-necrose, que apresentava uma área extensivamente colagenizada. Após revisão para consenso entre dois patologistas, concluiu-se que se trata de uma área cicatricial após extensa necrose, mas ainda não tecido fibrótico. Esta situação ilustra o cenário em que algumas características histológicas podem exibir uma histologia muito semelhante e até sobreposta. As anotações deste caso configuram 100% do total de situações em que houve confusão entre **NEC** e **FIB** no grupo de teste, justificando a aparente performance inferior do modelo para esta situação em casos externos. No terceiro cenário, observaram-se a ocorrência de distintos padrões de resposta. Em uma única lâmina, foi possível identificar regiões com resposta reduzida, onde a neoplasia difusa se manifesta em forma de lençol, além de áreas com resposta aumentada ao tratamento, onde o tumor está entremeado com diferentes classes histológicas (Figura 28).

Os achados discutidos estão diretamente relacionados às características intrínsecas do processo histológico reparativo subsequente ao tratamento sistêmico. A resposta fibrótica e inflamatória à necrose demonstra sobreposição morfológica às alterações fibróticas específicas associadas ao tratamento. Embora os processos responsivos ao tratamento ten-

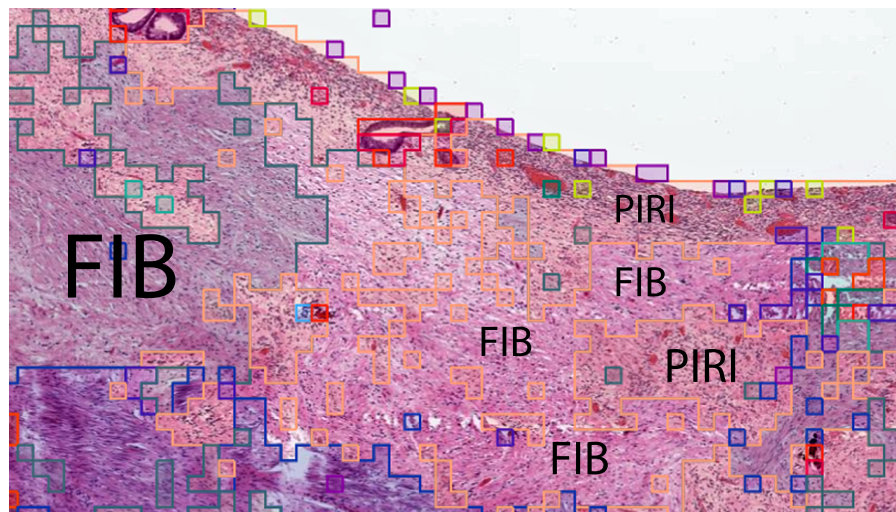
Figura 26 – Caso de **FMRP** com área densa e “colagenizada”



Fonte: Autoria própria.

Legenda: Lâminas do caso de **FMRP** com área densa e “colagenizada”, representando os erros de classificação de **NEC** como **FIB**. A: aumento de 20x com áreas ressaltadas com retângulos pretos sem preenchimento; B: área de cicatriz após necrose em aumento de 40x; C: segunda lâmina do mesmo caso, com áreas de cicatriz ressaltadas em retângulos verde-água. D: detalhe da área de cicatriz em aumento de 40x.

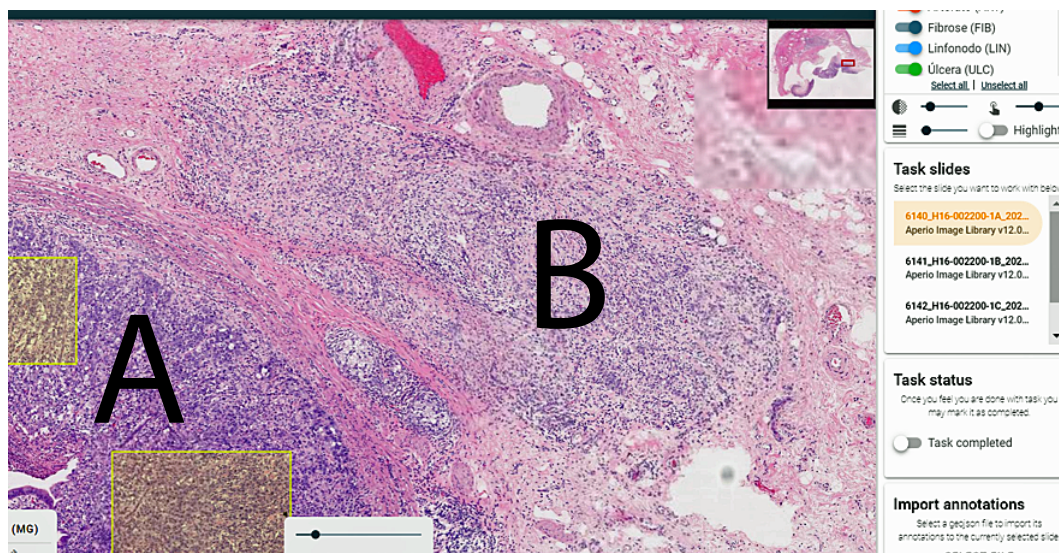
Figura 27 – Classes mescladas em **LT**



Fonte: Autoria própria.

Legenda: Imagem ilustrando situação em que duas das três classes confundidas se apresentam na área de **LT** de forma mesclada, sem limites claros entre si. **PIRI** (laranja claro) em meio a **FIB** (áreas sem cor em lâmina com tecido do **TGI**). Área mais extensa de **FIB** à esquerda em cinza.

Figura 28 – Diferentes padrões de resposta



Fonte: Autoria própria.

Legenda: Diferentes padrões de resposta para um mesmo tumor na mesma lâmina. A: área com baixa resposta, na qual a neoplasia difusa pode ser identificada na forma de um lençol. B: área de maior resposta ao tratamento, no qual o tumor (áreas mais roxas) está entremeadado com áreas de **PIRI** (áreas mais claras).

dam a evoluir para a fibrose tecidual, pode-se observar uma combinação de componentes e morfologias, com proporções variáveis de **FIB**, **PIRI** e **NEC**, dependendo do estágio do processo (Becker et al. 2003; Rödel et al. 2005; Pai e Pai 2018).

Adicionalmente, a presença da classe **NEC** não é necessariamente indicativa de resposta ao tratamento, visto que pode ser identificada em neoplasias que ainda não passaram por terapias prévias (Becker et al. 2003; Rödel et al. 2005; Pai e Pai 2018). Além disso, a resposta fibrótica e inflamatória à necrose também se sobrepõe, morfológicamente, às modificações fibróticas relacionadas ao tratamento. O mesmo se aplica à classe **PIRI**, já que áreas de regressão tumoral frequentemente exibem um infiltrado inflamatório agudo ou subagudo em resposta aos efeitos citotóxicos do tratamento, manifestando-se como tecido em que um processo fibroso-inflamatório substitui as células tumorais malignas destruídas (Becker et al. 2003; Pai e Pai 2018).

Em resumo, o processo de resposta ao tratamento resultará, eventualmente, em fibrose tecidual; entretanto, haverá uma combinação de componentes, com proporções variadas de **PIRI**, **NEC** e **FIB**, dependendo do estágio do processo. Ademais, a necrose **NEC** em si desencadeia uma reação inflamatória, o que torna ainda mais complexa a distinção entre essas classes. Portanto, entende-se que a dificuldade do modelo neste contexto reflete a complexidade intrínseca enfrentada por patologistas ao discernir esses processos, o que se deve, em grande parte, ao fato de serem, na realidade, estágios de um mesmo processo reparativo ou apresentarem-se de forma misturada, dificultando sua categorização dicotômica.

No conjunto de testes **R_ERTG**, cujo objetivo foi avaliar o desempenho do modelo **M1** na reprodução do **ERTG**, a concordância entre observadores — incluindo patologistas especialistas e o modelo **M1** — constituiu um pilar central na avaliação do desempenho de **M1**. Este processo visou estabelecer um *Ground Truth* (**GT**) mais robusto para comparar o desempenho da **IA**, considerando a conhecida alta variabilidade interobservador dos métodos de **ERTG**. Foram designados 3 grupos de testes, **Conc_3Patol** para avaliar a concordância das observações entre os 3 patologistas considerados especialistas em **TGI**, **Conc_ERTG** para avaliar concordância entre o modelo **M1** e patologistas (**'pat'**, **M**, **G**, **W**) e **Conc_GT**, na qual avaliou-se o desempenho de **M1** isolando apenas os casos onde os patologistas **M**, **G** e **W** demonstraram uma alta concordância mútua (χ (agree > .5)), definido como **GT**.

No grupo de testes **Conc_3Patol**, observou-se uma forte associação linear entre as avaliações dos patologistas. Contudo, a concordância ordinal, avaliada pelo *Kappa* de *kendall* (τ), e a concordância categórica, avaliada pelo *Kappa* (*Becker*), indicaram níveis de reprodutibilidade limitados, situados nas faixas de Moderada a Razoável. Especificamente, a concordância categórica variou de 0.35 (**M** vs **G**) a 0.45 (**W** vs **G**). Embora o patologista **W** tenha demonstrado a consistência mais robusta em suas classificações categóricas em relação aos seus pares, a discrepância entre o alto coeficiente de correlação de Pearson e o *kappa* mais baixo sugere que, embora os patologistas **M**, **G** e **W** apresentem

uma tendência semelhante em suas classificações, alcançar uma concordância categórica exata é desafiador. Isso ressalta a alta subjetividade inerente aos métodos de avaliação de **ERTG**, o que, por sua vez, dialoga diretamente com a dificuldade de identificação e quantificação de algumas categorias do **LT**.

Conc_ERTG revelou um padrão tendencioso, no qual o observador **W** apresentou a maior congruência com as estimativas do modelo **M1** em todas as abordagens analisadas, evidenciado pelos mais altos coeficientes de correlação de *Pearson* (atingindo r , de até 0,88 para **M1_Glob**, e coeficientes de *kendall* de até $\tau = 0,72$). Este alinhamento foi reforçado pelo elevado coeficiente de determinação (R^2) observado para **W**, que chegou a 0,69. Em contraste, os níveis de concordância observados para os patologistas **M** e '**pat**' variaram de Pobre a Razoável. O coeficiente *kappa* (*Becker*) foi consistentemente baixo, com valores mínimos de 0,085, classificando a concordância como Pobre e sugerindo uma proximidade ao acaso. Essa disparidade sugere fortemente que a performance de **M1** está intrinsecamente ligada ao patologista **W**, que foi o único patologista a participar das anotações manuais que treinaram o **M1**.

Os valores negativos do *kappa* encontrados nas comparações envolvendo '**pat**' indicam que as abordagens do **M1** tendem à discordância em relação ao '**pat**'. Um ponto crítico para os **ERTG** provenientes do laudo original é que muitos desses laudos são antigos, já que a coorte se estende desde o final de 2014. Consequentemente, os **ERTG** foram realizados por patologistas diferentes dos que participaram deste estudo, incorporando assim uma interpretação distinta das lâminas. Além disso, é importante notar que os protocolos de anatomia patológica não são estáticos e passam por atualizações regulares. O **CAP**, por exemplo, é atualizado anualmente (College of American Pathologists 2024). Portanto, uma parte significativa dos relatórios de '**pat**' pode não estar atualizada e ser sujeita a interpretações de patologistas que não fazem mais parte da instituição. Esta hipótese está alinhada com a baixa concordância entre '**pat**' e a patologista **M**, que não pertence à instituição **ACCCC**.

No grupo **Conc_GT** o patologista **G** apresentou uma melhora da concordância com **M1**, apresentando um *kappa* (*Becker*) de 0,31 e uma métrica χ (agree $>,5$) de 0,26 nas comparações com **M1_Glob** e **M1_Patoll**. A concordância para os patologistas **W** e **M** permaneceu relativamente estável ou apresentou uma ligeira diminuição quando comparada a **M1** neste subconjunto de casos mais claros. Na interpretação desses resultados, o aumento observado na concordância para o Patologista **G**, ao isolar os casos de alta reprodutibilidade inicial, indica que as abordagens de **M1** apresentam maior capacidade de concordância com as avaliações humanas quando há maior clareza e menos ambiguidade no próprio caso (conforme refletido pela concordância *kappa* $> 0,5$).

Em uma análise abrangente dos ensaios **R_ERTG**, conclui-se que o modelo **M1** demonstra uma capacidade robusta de replicar o desempenho observado durante seu treinamento na reprodução do **ERTG**, atuando como um espelho fiel do médico patologista **W** que contribuiu para seu treinamento. Essa capacidade é atribuída, provavelmente, à

utilização de arquiteturas avançadas de **GNNs**, como **ARMA** e **GIN**, que permitiram a **M1** explorar as complexas relações espaciais do **LT**. A discordância observada entre **M1** e os demais patologistas, **M** e **G**, pode ser interpretada como um reflexo da variabilidade inerente entre observadores na avaliação do **ERTG**, fenômeno corroborado pelos resultados do subconjunto de testes **Conc_3Patol**. Além disso, a acentuada discordância de **M1** em relação a '**pat**' não implica, necessariamente, em uma performance inadequada de **M1**, mas pode ser consequência da desatualização e de diferentes interpretações do **ERTG** em casos mais antigos. Conclui-se, portanto, que **M1** é competente na reprodução do **ERTG**, ainda que apresente a limitação de operar de maneira muito semelhante ao patologista **W**, incorporando seus vieses. Esse achado reforça a notável capacidade de uma **GNN** em aprender a partir do contexto: **M1** se tornou tão eficiente em reproduzir a interpretação de **W**, com quem foi treinado, que reflete essa atuação de forma consistente, apresentando discordância significativa em relação a outros patologistas, especialmente à patologista **M**, que não integra o **ACCCC**.

Os ensaios realizados no subconjunto **SPC** demonstraram que a combinação de **M1** com dados anátomo-clínicos (**cli**) melhora o poder preditivo prognóstico pós tratamento neoadjuvante. Ao avaliar seis diferentes combinações de características, a análise revelou variações significativas no poder preditivo. A combinação que se destacou para a predição de sobrevida foi consistentemente a '**cli** + **model**'. Esta combinação integra as variáveis do laudo anatomopatológico original ('**cli**') com as 15 categorias histológicas identificadas e quantificadas pelo **M1** ('**model**'). A combinação '**cli** + **model**' demonstrou o desempenho mais robusto para os modelos RF, atingindo valores de Índice de Concorância (**CI**) próximos ou superiores a 0,90 no teste interno. Embora o desempenho no teste externo tenha sido ligeiramente inferior, situando-se em torno de 0,85, ainda assim representou um nível robusto de predição. Em forte contraste, as combinações que dependiam unicamente do **ERTG** estimado pelo patologista ('**pat**') ou que incluíam apenas as categorias do **LT** quantificadas por **M1** demonstraram o pior desempenho, com valores de **CI** entre 0,50 e 0,55. Isso indica que a capacidade preditiva dessas variáveis isoladas é apenas marginalmente superior ao acaso.

A superioridade da combinação '**cli** + **model**' sugere que a inclusão das características quantificadas pelo **M1** ('**model**'), juntamente com os dados clínicos e patológicos básicos '**cli**', configura a combinação com maior poder de predição de sobrevida. Esta descoberta é particularmente relevante no contexto da premissa do estudo de que todos os componentes de uma lâmina, e não apenas aqueles contidos no **LT**, poderiam contribuir para uma interpretação aprimorada do **ERTG** e aumentar seu potencial prognóstico. Observa-se que '**cli** + **model**' superou '**mscore**', no qual apenas as variáveis contidas no **LT** foram incluídas na análise, em conjunto com as informações '**cli**'.

A análise de contribuição das variáveis através da metodologia **SHAP** revelou uma hierarquia de importância bem definida, mostrando que o poder preditivo do modelo **RF** é fortemente dependente de informações que vão além do **ERTG**, contidas em '**cli**'. As

variáveis com os valores médios de **SHAP** mais elevados e, portanto, as mais influentes foram o status linfonodal, o estadiamento N e as dimensões macroscópicas do **LT**. O *status* linfonodal, que traz a relação entre o número de linfonodos comprometidos e o número de linfonodos ressecados, permanece o principal determinante prognóstico na predição de sobrevida de forma isolada. Esse achado dialoga diretamente com dados da literatura, demonstrando que o estadiamento, hoje, é o melhor preditor prognóstico de pacientes pós terapia neoadjuvante. Sendo assim, ressalta-se a importância do status linfonodal e do tamanho do **LT** para a predição prognóstica (Maas et al. 2010; Campos-Lobato et al. 2011; Stark et al. 2019).

Em resumo, os achados deste estudo corroboram e sustentam as hipóteses iniciais propostas. O *status* linfonodal continua sendo o indicador prognóstico isolado mais relevante para a predição de **SPC**. Além disso, observou-se que a medida macroscópica do **LT** é uma variável de extrema importância. Finalmente, quando essas variáveis são combinadas com o **ERGT**, que valoriza o ecossistema completo de uma lâmina (**'cli + model'**) e não apenas os componentes do **LT**, demonstra-se um maior poder preditivo prognóstico para **SPC**.

Em diálogo com as metodologias de **GNN** empregadas no estudo, que são motivadas pela necessidade de superar a análise local (*patch-based*) e captar a compreensão contextual e topológica dos tecidos, o desempenho superior da combinação **'cli + model'** na predição de **SPC** sugere que as características extraídas pelo modelo **M1** proporcionam um ganho preditivo significativo ao incorporar padrões micro-arquiteturais que não são capturados isoladamente. A habilidade das **GNNs** de analisar o tecido além dos componentes estritamente definidos do **LT** — avaliando a composição integral da lâmina e o contexto tecidual — permitiu ao modelo **M1** fornecer características que se associam a desfechos clínicos, complementando as informações de estadiamento e medida de leito tumoral de alta hierarquia.

Em conclusão, a **GNN**, através da sua capacidade de modelar relações espaciais, permitiu que o **M1** destilasse informações prognósticas da imagem histopatológica que, quando integradas ao contexto clínico, criam uma ferramenta preditiva poderosa. O possível *overfitting* observado com uma performance menor em dados externos (**FMRP**) sugere que, embora a arquitetura **GIN** tenha demonstrado maior capacidade de generalização na classificação de categorias do que outras **GNNs**, a complexidade da predição de sobrevida e a limitação dos dados ainda impõem barreiras à generalização do modelo **M1**.

5 CONCLUSÃO

O presente estudo avaliou o modelo **M1**, baseado em **GNN**, para reproduzir o Escore de Regressão Tumoral Gástrico (**ERTG**) e aprimorar a predição de Sobrevida Pós-Tratamento Cirúrgico (**SPC**) em pacientes com adenocarcinoma gástrico localmente avançado. As **GNNs** foram empregadas por sua capacidade de capturar relações microarquiteturais nos tecidos, permitindo integrar padrões histológicos quantificados pelo modelo às variáveis clinicopatológicas. A combinação '**cli + model**' apresentou o melhor desempenho prognóstico com um alto (**CI**) em teste externo, valor substancialmente superior obtido com o **ERTG** isolado ou com categorias morfológicas tradicionais ('**cli**') isoladas.

O modelo **M1** demonstrou um desempenho sólido em todas as etapas: obteve bom desempenho com na identificação de classes histológicas com a arquitetura **GIN** e extrator de *features* Virchow 2, obtendo F1-SCORE superior a 0.87 em todas as etapas de desenvolvimento e teste. Na reprodução do **ERTG**, o modelo **M1** demonstrou uma capacidade robusta de replicar o desempenho observado durante seu treinamento, estabelecendo uma alta congruência com o Patologista **W**, evidenciada pelos mais elevados coeficientes de correlação de *Pearson* e coeficientes *dekendall*. Ainda, o melhor desempenho do conjunto '**cli + model**' na análise de **SPC**, corroborou a premissa original de que a combinação de **M1** com dados anátomo-clínicos melhora o poder preditivo pós tratamento cirúrgico.

Entre as limitações, observou-se viés de anotação, com maior concordância do modelo com o patologista que participou da rotulagem do conjunto de treino, além de confusões entre classes morfológicamente semelhantes (**PIRI**, **FIB**, **NEC**) Detectou-se também um possível *overfitting*, dado o desempenho maior no teste interno em comparação ao teste externo na predição de **SPC**. Como estudo futuro, sugere-se o uso de anotações mais diversas e consensuais para reduzir tais vieses.

Em síntese, o estudo confirma que as **GNNs** podem extrair padrões relevantes da histologia que, combinados a informações clínicas, elevam de forma significativa o poder preditivo de sobrevida pós cirurgia, estabelecendo o modelo **M1** como uma abordagem promissora em patologia computacional aplicada ao câncer gástrico com tratamento neoadjuvante.

REFERÊNCIAS

- Agarwal A, Chang GJ, Hu CY, Taggart M, Rashid A, Park IJ, et al. CancerQuantified pathologic response assessed as residual tumor burden is a predictor of recurrence-free survival in patients with rectal cancer who undergo resection after neoadjuvant chemoradiotherapy. *Cancer*. 2013 Dec; 119:4231–4241.
- Ahmedt-Aristizabal D, Armin MA, Denman S, Fookes C, Petersson L. A survey on graph-based deep learning for computational histopathology. *Computerized Medical Imaging and Graphics*. 2022 jan; 95:102027.
- Aitazaz T, Tubaishat A, Al-Obeidat F, Shah B, Zia T, Tariq A. Transfer learning for histopathology images: an empirical study. *Neural Computing and Applications*. 2023 apr; 35:7963–7974.
- Akbar S, Peikari M, Salama S, Panah AY, Nofech-Mozes S, Martel AL. Automated and manual quantification of tumour cellularity in digital slides for tumour burden assessment. *Scientific Reports*. 2019 oct; 9:14099.
- Alzoubi I, Zhang L, Zheng Y, Loh C, Wang X, Graeber MB. Pathograph: An attention-based graph neural network capable of prognostication based on cd276 labelling of malignant glioma cells. *Cancers*. 2024 feb; 16:750.
- Amaral I. Complex networks. In: Schintler LA, McNeely CL (Ed.). **Encyclopedia of Big Data**. Cham: Springer, 2022. 198–201. Citation on page 34.
- American Joint Committee on Cancer. **AJCC Cancer Staging Manual**. 8th. ed. New York: Springer International Publishing, 2017. ISBN 978-3-319-40617-6.
- Amin MB, Edge SB, Greene FL, Byrd DR, Brookland RK, Washington MK (Ed.). **AJCC Cancer Staging Manual**. Eighth edition. [S.l.]: Springer International Publishing, 2017. ISBN 978-3319406176.
- Arnal MJD, Arenas ÁF, Arbeloa ÁL. Esophageal cancer: Risk factors, screening and endoscopic treatment in western and eastern countries. *World journal of gastroenterology: WJG*. 2015; 21:7933.
- Aswathy M, Jagannath M. An SVM approach towards breast cancer classification from H&E-stained histopathology images based on integrated features. *Medical & Biological Engineering & Computing*. 2021 sep; 59:1773–1783.
- Attaway CC, Fortuna D. Asking the prime questions for grossing: Teaching a framework for grossing and constructing gross descriptions using the prime model. *Academic Pathology*. 2024; 11:100143.
- Baranwal M, Krishnan S, Oneka M, Frankel T, Rao A. CGAT: Cell graph attention network for grading of pancreatic disease histology images. *Frontiers in Immunology*. 2021 oct; 12:727610.
- Bateman AC, Jaynes E, Bateman AR. HistopathologyRectal cancer staging post neoadjuvant therapy—how should the changes be assessed? *Histopathology*. 2009 May; 54:713–721.

Becker K, Mueller JD, Schulmacher C, Ott K, Fink U, Busch R, et al. CancerHistomorphology and grading of regression in gastric carcinoma treated with neoadjuvant chemotherapy. *Cancer*. 2003 Oct; 98:1521–1530.

Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. *Nature Reviews Clinical Oncology*. 2019 nov; 16:703–715.

Bianchi FM, Grattarola D, Livi L, Alippi C. Graph neural networks with convolutional ARMA filters. *CoRR*. 2019;abs/1901.01343:.

Bruna J, Zaremba W, Szlam A, LeCun Y. Spectral networks and locally connected networks on graphs. In: **2014 International Conference on Learning Representations (ICLR)**. [s.n.], 2014. Citations on pages 35, 36, and 40. Disponível em: <https://arxiv.org/abs/1312.6203>.

Brussee S, Buzzanca G, Schrader AMR, Kers J. **G raph Neural Networks in Histopathology: Emerging Trends and Future Directions**. 2024. ArXiv preprint. Disponível em: <https://arxiv.org/abs/2406.12808> .

de Campos-Lobato LF, Stocchi L, da Luz Moreira A, Geisler D, Dietz DW, Lavery IC, et al. Ann Surg OncolPathologic complete response after neoadjuvant treatment for rectal cancer decreases distant recurrence and could eradicate local recurrence. *Ann Surg Oncol*. 2011 Jun; 18:1590–1598.

Chen RJ, Ding T, Lu MY, Williamson DFK, Jaume G, Song AH, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*. 2024 mar; 30:850–862.

Chen RJ, Ding T, Lu MY, Williamson DF, Jaume G, Chen B, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*. 2024;:.

Chen RJ, Ding T, Lu MY, Williamson DFK, Jaume G, Song AH, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*. 2024 mar; 30:850–862.

Chetty R, Gill P, Govender D, Bateman A, Chang HJ, Deshpande V, et al. Hum PatholInternational study group on rectal cancer regression grading: interobserver variability with commonly used regression grading systems. *Hum Pathol*. 2012 Nov; 43: 1917–1923.

Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. *Translational Vision Science & Technology*. 2020 feb; 9:14.

College of American Pathologists. **Protocol for the Examination of Resection Specimens From Patients With Primary Carcinoma of the Colon and Rectum**. 4.4.0.1. ed. [S.l.], 2023. CAP Cancer Protocol Templates. Disponível em: <https://www.cap.org/protocols-and-guidelines/cancer-reporting-protocols/cancer-protocol-templates>.

College of American Pathologists. **Protocol for the Examination of Resection Specimens From Patients With Primary Carcinoma of the Esophagus**. 4.2.0.1.

ed. [S.l.], 2023. CAP Cancer Protocol Templates. Disponível em: <https://www.cap.org/protocols-and-guidelines/cancer-reporting-protocols/cancer-protocol-templates>.

College of American Pathologists. **Protocol for the Examination of Resection Specimens From Patients With Primary Carcinoma of the Stomach**. 5.0.0.0. ed. [S.l.], 2023. CAP Cancer Protocol Templates. Disponível em: <https://www.cap.org/protocols-and-guidelines/cancer-reporting-protocols/cancer-protocol-templates>.

College of American Pathologists. **Cancer Protocol Templates**. 2024. <https://www.cap.org/protocols-and-guidelines/cancer-reporting-protocols/cancer-protocol-templates> Accessed: 2024-05-22.

Corredor G, Wang X, Zhou Y, Lu C, Fu P, Humayun M, et al. Spatial architecture and arrangement of tumor-infiltrating lymphocytes for predicting likelihood of recurrence in early-stage non-small cell lung cancer. *Clinical Cancer Research*. 2019 mar; 25: 1526–1534.

Corso G, Stark H, Jegelka S, et al. Graph neural networks. *Nature Reviews Methods Primers*. 2024; 4:17.

Cunningham D, Allum W, Stenning S, Thompson J, Van de Velde C, Nicolson M, et al. Perioperative chemotherapy versus surgery alone for resectable gastroesophageal. *N Engl J Med*. 2016 Jul; 6:11–20.

Cunningham D, Allum WH, Stenning SP, Thompson JN, Van de Velde CJ, Nicolson M, et al. Perioperative chemotherapy versus surgery alone for resectable gastroesophageal cancer. *The New England Journal of Medicine*. 2006 jul; 355:11–20.

Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. *CoRR*. 2016;abs/1606.09375:.

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: **2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.]: IEEE, 2009. 248–255.

Diamond J, Anderson NH, Bartels PH, Montironi R, Hamilton PW. The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia. *Human Pathology*. 2004 sep; 35:1121–1131.

Dworak O, Keilholz L, Hoffmann A. Int J Colorectal DisPathological features of rectal cancer after preoperative radiochemotherapy. *Int J Colorectal Dis*. 1997; 12:19–23.

Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural message passing for quantum chemistry. In: **Proceedings of the 34th International Conference on Machine Learning (ICML)**. PMLR, 2017. (Proceedings of Machine Learning Research, 70), 1263–1272. Disponível em: <http://proceedings.mlr.press/v70/gilmer17a.html>.

Gonzalez RC, Woods RE. **Digital Image Processing**. 4th. ed. Upper Saddle River, NJ: Pearson, 2018. ISBN 9780133356724.

Hamilton WL, Ying R, Leskovec J. Inductive representation learning on large graphs. *CoRR*. 2017;abs/1706.02216:.

Hanna MG, Pantanowitz L, Dash R, Harrison JH, Deebajah M, Pantanowitz J, et al. Future of artificial intelligence-machine learning trends in pathology and medicine. *Modern Pathology*. 2025 apr; 38:100705.

Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*. 1973 nov;SMC-3:610–621.

Heer EV, Harper AS, Sung H, Jemal A, Fidler-Benaoudia MM. Emerging cancer incidence trends in canada: The growing burden of young adult cancers. *Cancer*. 2020 oct; 126:4553–4562.

Huebner M, Wolff BG, Smyrk TC, Aakre J, Larson DW. World J Surg Partial pathologic response and nodal status as most significant prognostic factors for advanced rectal cancer treated with preoperative chemoradiotherapy. *World J Surg*. 2012 Mar; 36: 675–683.

Kim SH, Chang HJ, Kim DY, Park JW, Baek JY, Kim SY, et al. Cancer Res Treat What Is the Ideal Tumor Regression Grading System in Rectal Cancer Patients after Preoperative Chemoradiotherapy? *Cancer Res Treat*. 2016 Jul; 48:998–1009.

Kim TH, Chang HJ, Kim DY, Jung KH, Hong YS, Kim SY, et al. Int J Radiat Oncol Biol Phys Pathologic nodal classification is the most discriminating prognostic factor for disease-free survival in rectal cancer patients treated with preoperative chemoradiotherapy and curative resection. *Int J Radiat Oncol Biol Phys*. 2010 Jul; 77: 1158–1165.

Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. *CoRR*. 2016;abs/1609.02907:.

Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. In: **International Conference on Learning Representations (ICLR)**. [s.n.], 2017. Published as a conference paper at ICLR 2017. Disponível em: <https://arxiv.org/abs/1609.02907>.

Kuhn M, Vaughan D. **parsnip: A Common API to Modeling and Analysis Functions**. [S.l.], 2025. R package version 1.3.1, <https://parsnip.tidymodels.org/>. Disponível em: <https://github.com/tidymodels/parsnip>.

Lauren P. The two histological main types of gastric carcinoma: Diffuse and so-called intestinal-type carcinoma. an attempt at a histo-clinical classification. *Acta Pathologica et Microbiologica Scandinavica*. 1965; 64:31–49.

LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015 may; 521:436–444.

LeCun Y, Bengio Y, et al. Convolutional networks for images, speech, and time series. In: Arbib MA (Ed.). **The Handbook of Brain Theory and Neural Networks**. [S.l.]: MIT Press, 1995. 3361, n. 10, 255–258. Citations on pages 29 and 30.

Li G, Xiong C, Thabet AK, Ghanem B. Deepergcn: All you need to train deeper gcns. *CoRR*. 2020;abs/2006.07739:.

Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A convnet for the 2020s. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2022. 11976–11986. Citations on pages 59 and 86.

Lu MY, Chen B, Williamson DFK, Chen RJ, Zhao M, Chow AK, et al. A multimodal generative ai copilot for human pathology. *Nature*. 2024 oct; 634:466–473.

Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Advances in neural information processing systems*. 2017; 30:.

Ma J, Shen H, Kapesa L, Zeng S. Lauren classification and individualized chemotherapy in gastric cancer. *Oncology Letters*. 2016 may; 11:2959–2964.

Maas M, Nelemans PJ, Valentini V, Das P, Rödel C, Kuo LJ, et al. *Lancet Oncol* Long-term outcome in patients with a pathological complete response after chemoradiation for rectal cancer: a pooled analysis of individual patient data. *Lancet Oncol*. 2010 Sep; 11:835–844.

Mandard AM, Dalibard F, Mandard JC, Marnay J, Henry-Amar M, Petiot JF, et al. CancerPathologic assessment of tumor regression after preoperative chemoradiotherapy of esophageal carcinoma. Clinicopathologic correlations. *Cancer*. 1994 Jun; 73: 2680–2686.

Martins LVdC, Bueno AP, Defelicibus A, Drummond RD, Valieris R, Yu-Tao Z, et al. Wsi2ml: an open-source whole slide image annotation software for machine learning applications. In: **Brazilian Symposium on Multimedia and the Web - WebMedia**. [S.l.]: ACM, 2023.

Morgan E, Arnold M, Camargo MC, Gini A, Kunzmann AT, Matsuda T, et al. The current and future incidence and mortality of gastric cancer in 185 countries, 2020-40: A population-based modelling study. *EClinicalMedicine*. 2022 apr; 47:101404.

Nagpal K, Holbrook R, Hom N, Peng L, Diggans J, Miller AJ, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *npj Digital Medicine*. 2019 jun; 2:48.

Nagtegaal ID, van de Velde CJH, van der Worp E, Kapiteijn E, Quirke P, van Krieken JH, et al. Macroscopic evaluation of rectal cancer resection specimen: clinical significance of the pathologist in quality control. *Journal of Clinical Oncology*. 2002 apr; 20:1729–34.

Nanni L, Ghidoni S, Brahnam S. Handcrafted vs. non-handcrafted features for computer vision classification. *Pattern Recognition*. 2017 nov; 71:158–172.

Nguyen A, Yosinski J, Clune J. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *arXiv preprint arXiv:160203616*. 2016;:.

Pai RK, Pai RK. Mod PatholPathologic assessment of gastrointestinal tract and pancreatic carcinoma after neoadjuvant therapy. *Mod Pathol*. 2018 01; 31:4–23.

Pantanowitz L, Pearce T, Abukhiran I, Hanna M, Wheeler S, Soong TR, et al. Nongenerative artificial intelligence in medicine: Advancements and applications in supervised and unsupervised machine learning. *Modern Pathology*. 2025 mar; 38:100680.

Pedroso dos Passos TD. **Anatomia funcional do trato gastrointestinal**. 2021. Disponível em: <https://sanarmed.com/anatomia-funcional-do-trato-gastrointestinal-colonistas/> Accessed May 30, 2025.

- Rashidi HH, Pantanowitz J, Hanna MG, Tafti AP, Sanghani P, Buchinsky A, et al. Introduction to artificial intelligence and machine learning in pathology and medicine: Generative and nongenerative artificial intelligence basics. *Modern Pathology*. 2025; 38: 100688.
- Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*. 2015 dec; 115: 211–252.
- Ryan R, Gibbons D, Hyland JM, Treanor D, White A, Mulcahy HE, et al. HistopathologyPathological response following long-course neoadjuvant chemoradiotherapy for locally advanced rectal cancer. *Histopathology*. 2005 Aug; 47: 141–146.
- Rödel C, Martus P, Papadoupoulos T, Füzesi L, Klimpfinger M, Fietkau R, et al. J Clin OncolPrognostic significance of tumor regression after preoperative chemoradiotherapy for rectal cancer. *J Clin Oncol*. 2005 Dec; 23:8688–8696.
- Saltz J, Gupta R, Hou L, Kurc T, Singh P, Nguyen V, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Reports*. 2018 apr; 23:181–193.e7.
- Sanchez-Lengeling B, Reif E, Pearce A, Wiltchko AB. A gentle introduction to graph neural networks. *Distill*. 2021 sep;:.
- Shah SC, Gawron AJ, Mustafa RA, Piazuolo MB. Histologic subtyping of gastric intestinal metaplasia: Overview and considerations for clinical practice. *Gastroenterology*. 2020 feb; 158:745–750.
- Stark AP, Ikoma N, Chiang YJ, Estrella JS, Das P, Minsky BD, et al. Ann Surg OncolCharacteristics and Survival of Gastric Cancer Patients with Pathologic Complete Response to Preoperative Therapy. *Ann Surg Oncol*. 2019 Oct; 26:3602–3610.
- Su P, Yu J. Progress in neoadjuvant therapy for gastric cancer (review). *Oncology Letters*. 2022; 23:172.
- Symmans WF, Peintinger F, Hatzis C, Rajan R, Kuerer H, Valero V, et al. J Clin OncolMeasurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. *J Clin Oncol*. 2007 Oct; 25:4414–4422.
- Tay K, Simon N, Friedman J, Hastie T, Tibshirani R, Narasimhan B. **Regularized Cox Regression • glmnet**. 2025. Disponível em: <https://glmnet.stanford.edu/articles/Coxnet.html> Accessed May 27, 2025.
- The Global Cancer Observatory. **GLOBOCAN 2020**. 2020. Disponível em: <https://gco.iarc.fr/>. Acesso em: 10 jan. 2024.
- Therneau TM. **A Package for Survival Analysis in R**. [S.l.], 2024. R package version 3.8-3. Disponível em: <https://CRAN.R-project.org/package=survival>.
- Tosta TAA, de Faria PR, Neves LA, do Nascimento MZ. Evaluation of statistical and Haralick texture features for lymphoma histological images classification. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*. 2021; 9:613–624.

Valieris R, Martins L, Defelicibus A, Bueno AP, de Toledo Osorio CAB, Carraro D, et al. Weakly-supervised deep learning models enable her2-low prediction from h & e stained slides. *Breast Cancer Research*. 2024 ago.; 26:.

Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. *arXiv preprint arXiv:171010903*. 2017;:.

Wang J, Chen RJ, Lu MY, Baras A, Mahmood F. Weakly supervised prostate TMA classification via graph convolutional networks. In: **2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)**. [S.l.]: IEEE, 2020. 239–243.

Wang X, Zhao J, Marostica E, Yuan W, Jin J, Zhang J, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*. 2024 oct; 634:970–978.

Wu W, Liu X, Hamilton RB, Suriawinata AA, Hassanpour S. Graph convolutional neural networks for histologic classification of pancreatic cancer. *Archives of Pathology & Laboratory Medicine*. 2023 nov; 147:1251–1260.

Xu K, Hu W, Leskovec J, Jegelka S. How powerful are graph neural networks? *CoRR*. 2018;abs/1810.00826:.

Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: **Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I**. [S.l.]: Springer, 2014. (Lecture Notes in Computer Science, 8689), 818–833. Citation on page 30.

Zhou Y, Graham S, Koohbanani NA, Shaban M, Heng PA, Rajpoot N. CGC-Net: Cell graph convolutional network for grading of colorectal cancer histology images. In: **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops**. [S.l.]: IEEE, 2019. 388–397.

Zimmermann E, Vorontsov E, Viret J, Casson A, Zelechowski M, Shaikovski G, et al. **Virchow2: Scaling Self-Supervised Mixed Magnification Models in Pathology**. 2024. Disponível em: <https://arxiv.org/abs/2408.00738> .

Zimmermann E, Vorontsov E, Viret J, Casson A, Zelechowski M, Shaikovski G, et al. **Virchow2: Scaling Self-Supervised Mixed Magnification Models in Pathology**. 2024. Disponível em: <https://arxiv.org/abs/2408.00738> .

APÊNDICE A – Abstract exposto na USCAP

Trabalho exposto como abstract na United States and Canadian academy of Pathology (**USCAP**) na seção de informática.

1235 Who is More Successful in Detecting Cervical High-grade Squamous Intraepithelial Lesions in Atrophic Background? Pathologists or Artificial Intelligence?

Zeynep Bayramoglu¹, Gizem Dursun Demir², Ufuk Ozkaya³, Abdulkemir Capar⁴, Tuba Abacioglu⁵, Fatma Aktepe⁶, Deniz Baycelebi⁷, Emre Karakok⁷, Sezen Kocarslan⁵, Yazgi Koy⁸, Zeynep Pehlivanoglu⁹, Turkan Rezanko¹, Fadime Salman⁷, Serdar Balci¹⁰, Ilknur Turkmen¹⁰

¹Memorial Hospital Group Pathology Laboratory, Istanbul, Turkey, ²Suleyman Demirel University, Turkey, ³Isparta, Turkey, ⁴Sariyer, Turkey, ⁵Istanbul, Turkey, ⁶Memorial Hospital, Turkey, ⁷Memorial Hospitals Group Pathology Department, Istanbul, Turkey, ⁸Memorial Hospital Group Pathology Department, Istanbul, Turkey, ⁹Memorial Hospitals Group Pathology Department, Istanbul, Turkey, ¹⁰Memorial Hospitals Group Pathology Department, Turkey

Disclosures: Zeynep Bayramoglu: None; Gizem Dursun Demir: None; Ufuk Ozkaya: None; Abdulkemir Capar: None; Tuba Abacioglu: None; Fatma Aktepe: None; Deniz Baycelebi: None; Emre Karakok: None; Sezen Kocarslan: None; Yazgi Koy: None; Zeynep Pehlivanoglu: None; Turkan Rezanko: None; Fadime Salman: None; Serdar Balci: None; Ilknur Turkmen: None

Background: Cervical cytology is one of the most important tests in the screening of cervical cancers. Early digital pathology and artificial intelligence (AI) studies were carried out in the cervical cytology era leading to the first commercially available AI products. Diagnosis of dysplasia is challenging in the atrophic background, as the morphology of both entities has similarities. The commercially available automated triage systems also put atrophy cases in the suspicious category. It will be very useful for pathologists and of course for patients if we can set an automatic screening program that has the ability to differentiate atrophy and High-Grade Squamous Intraepithelial Lesion (HSIL).

Design: We created 30 synthetic slide images using a private dataset to detect HSIL in the atrophic background using deep learning networks. The private dataset contains 1832 HSIL and 1238 atrophy images cropped from whole slide images belonging to 9 patients. The synthetic images are created by combining randomly selected 100, 150, 200, 250 and 300 atrophy images with 0, 1, 2, 3, 4 and 5 HSIL images. As a result of this stage, created 30 synthetic slide images are classified by 10 different pathologists and 10 different deep learning networks.

Results: According to synthetic slide images containing HSIL, while the average of sensitivity is of these 10 deep learning networks is 99.79%; this result is calculated as 52.11% for pathologists. Similarly, the average of specificity for these 10 deep learning networks is calculated as 0.9866%; the result is calculated as 97.45% for pathologists. Moreover, for synthetic slide images containing HSIL, the rate of correctly detection at least one HSIL is 100% for 10 deep learning networks, this rate is 83.2% for pathologists. As a result, deep learning networks used in this work, classified HSIL and atrophy cells more successfully than pathologists.

Conclusions: Our study is the preliminary study for an automated WSI screening program to increase the accuracy of HSIL detection in atrophic background. A system categorizing HSIL correctly, can aid pathologists in the final diagnosis. The next step is making the study on patient slides, instead of our synthetic images.

1236 Replication of Gastrointestinal Tumor Regression Scores by Estimation of Tumor Bed Histopathological Components Through Artificial Intelligence

Adriana Bueno¹, Luan Martins², Evandro De Mello³, Mariangela Brunaldi⁴, Gabriel Santos¹, Alexandre Defelicibus⁵, Felipe Coimbra¹, Renan Valieris⁵, Israel da Silva⁵

¹Instituto de Anatomia Patológica, AC Camargo Cancer Center, São Paulo, Brazil, ²Universidade de São Paulo, São Paulo, Brazil, ³University of São Paulo, São Paulo, Brazil, ⁴Ribeirão Preto Medical School - University of São Paulo, São Paulo, Brazil, ⁵A.C. Camargo Cancer Center, São Paulo, Brazil

Disclosures: Adriana Bueno: None; Luan Martins: None; Evandro De Mello: None; Mariangela Brunaldi: None; Gabriel Santos: None; Alexandre Defelicibus: None; Felipe Coimbra: None; Renan Valieris: None; Israel da Silva: None

Background: Tumor regression scores (TRS) for gastrointestinal cancers are laborious, with high interobserver variability and fail to provide precise prognostic projection. This study aims to reproduce TRS for gastrointestinal adenocarcinomas through artificial intelligence (AI) by developing an algorithm capable of recognizing and estimating the area of the components within the tumor bed in histopathological images.

Design: Whole-slide images were obtained from 46 surgical specimens of patients with esophagus, stomach, and rectal adenocarcinoma post treatment. It was defined 17 histological categories considered to represent areas of interest in a tumor bed (Table 1). Representative regions of these classes were hand-delineated using a self-developed platform (Figure 1). HE image-patches were generated to train a Graph Neural Network (GNN), aiming to predict the components within the gastrointestinal tumor bed. We then tested the model performance in partially annotated slides, obtaining the prediction of all classes (Table 1; Figure 2).

Results: The hand-delineated regions made in 262 slides summed a total of 22308 annotations, which rendered 471.186, 1.966.856 and 4.571.554 tiles with 0%, 50% and 67% intersection respectively. Upon prediction in the testing set we obtained an accuracy up to 99.96% for each class (Table 1) and an F1 score that ranged from 26.54% to 97.54% per slide.

Accuracy of the GNN prediction per class	
Class	GNN Accuracy
Fibrosis	0.84
Invasive tumor	0.92
Acellular mucin	0.97
Necrosis	0.99
Dystrophic calcification	0.43
Inflammatory/reparative inspecific process	0.69
Lymphocytic infiltrate	0.67
Lymphoid aggregate	0.35
Lymph node	0.28
Squamous mucosa	0.99
Gastric Mucosa	0.95
Colonic mucosa	0.97
Adipose tissue	0.99
Stroma	0.77
Muscular tissue	0.83
Debri	0.33
Artefact	0.57

Figure 1 - 1236

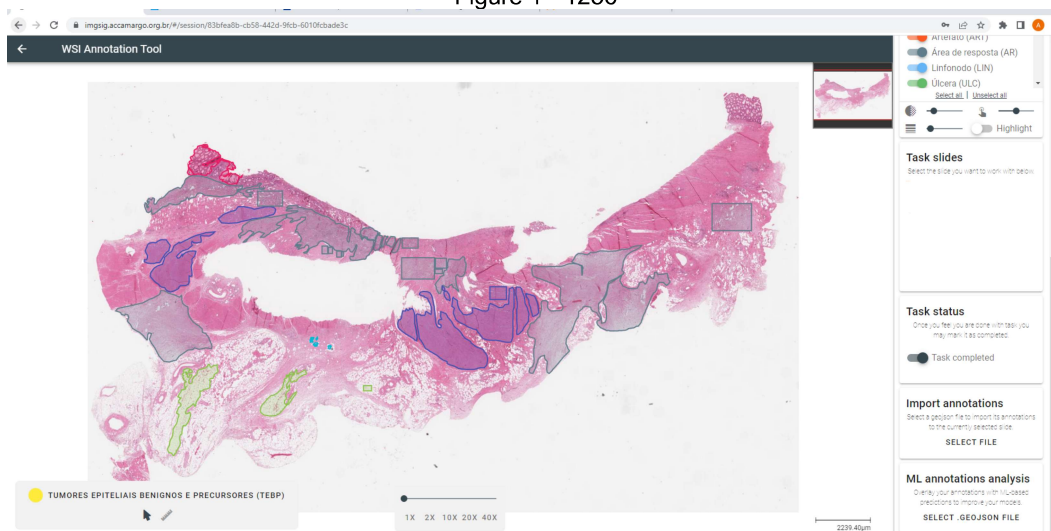
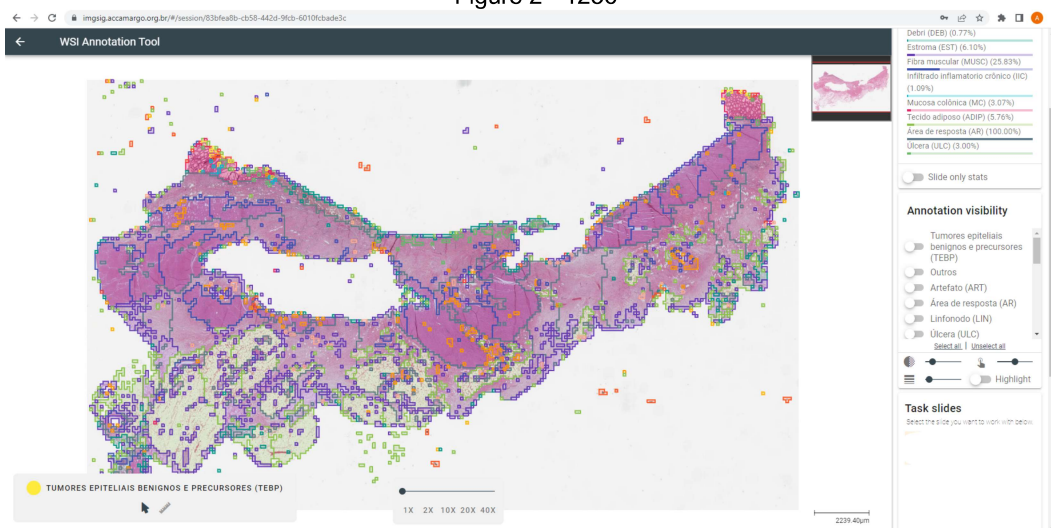


Figure 2 - 1236



Conclusions: So far, we have generated the largest data collection with regions composed exclusively of gastrointestinal tissue from post neoadjuvant patients. Also, we developed an online platform designed not only to facilitate hand delineating of regions, but also provide visualization of the prediction result, which allows precise modeling of the algorithm (Figure 2). We are now at further accuracy evaluation and model optimization phase. The next step is to predict TRS via AI and validate our findings with prospective and multi-institutional data. Also, we plan to build a digital signature to correlate the estimated proportion of classes with disease free and global survival, aiming to find the best combination of classes that better predict prognosis.

1237 High Concordance Between Manual and AI-assisted Evaluation of Ki-67 in Breast Cancer with a Ki-67 Containing Antibody Cocktail

Xin Chen¹, Yihui He¹, Chenyang Xia², Qinghai Yang³

¹Fujian Provincial Hospital, Fuzhou, China, ²Ningbo Sunny Instruments Co., Ltd., Ningbo, China, ³Fuzhou Maixin Biotech. Co., Ltd., Fuzhou, China

Disclosures: Xin Chen: None; Yihui He: None; Chenyang Xia: None; Qinghai Yang: None

Background: Proliferating marker Ki-67 is not only prognostic for invasive breast cancer, but also predictive for the effectiveness of chemotherapy or endocrine therapy. It can be utilized as a dynamic indicator for the efficacy pre-, during, and post neoadjuvant therapy. In daily practice, however, the in situ carcinoma component and the mesenchymal cells/lymphocytes challenge the reliability of Ki67 evaluation. As a result, the interpretation is significantly impacted by the expertise and subjectivity of a pathologist.

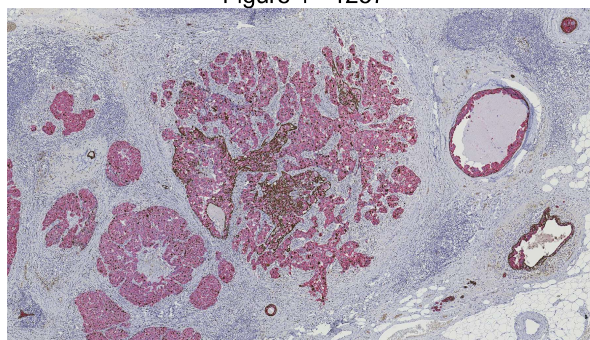
Design: We designed an antibody cocktail consisting of myoepithelial marker (cytoplasmic brown staining), tumor cell marker (cytoplasmic red staining), and Ki-67 (nuclear brown staining) to help pathologists evaluate the Ki-67 index in invasive breast carcinoma (Table 1). This cocktail was tested with manual analysis by pathologists (initial) and with aid of artificial intelligent (AI) on a set of 100 cases of invasive carcinoma of no specific type from Fujian Provincial Hospital.

Results: With this newly developed cocktail, we could easily (manual or AI) separate Ki-67 staining within carcinoma in situ, invasive carcinoma, and stromal cells (Figure 1). In 97% (100 cases in tissue microarray (TMA) cases, Ki67 index in invasive breast cancers was easily evaluated by manual counting the Ki-67 positive cells in a background of pink staining cancer cells that lack myoepithelial cells. Three cases of triple negative breast cancer (TNBC) tumor cells showed overlapping brown staining of both cytoplasm and nucleus in tumor cells (mixed brown and red staining), indicating that this cocktail might not be helpful in basal-like TNBCs. The Ki-67 scoring algorithm was developed with AI. The invasive breast cancer areas were precisely identified by the Ki-67 scoring algorithm, and the Ki-67 indexes within those areas were calculated (see Figure 2). The ICC (intraclass correlation coefficient) for this cocktail between the AI-scoring algorithm and the manual evaluation was 0.989 (P<0.001), indicating high concordance.

Table 1 The staining pattern of cocktail in breast cancer tissues

Type of tissue	Myoepithelial markers	Tumor cell markers
Carcinoma in situ of the breast	Positive (+)	Positive (+)
Invasive carcinoma of the breast	Negative (-)	Positive (+)
stromal cells and lymphocytes	Negative (-)	Negative (-)

Figure 1 - 1237



APÊNDICE B – WSI2ML

Plataforma de anotação e visualização de imagens de desenvolvimento próprio

WSI2ML– An Open-Source Whole Slide Image Annotation Software for Machine Learning Applications

Luan V. C. Martins

Institute of Mathematics and
Computer Science
University of São Paulo (USP)
São Carlos, Brazil
luan.martins@usp.br

Adriana Passos Bueno

Laboratory of Computational Biology
Bioinformatics
CIPE/A.C. Camargo Cancer Center
São Paulo, Brazil
adriana.passos@accamargo.org.br

Alexandre Defelicibus

Laboratory of Computational Biology
Bioinformatics
CIPE/A.C. Camargo Cancer Center
São Paulo, Brazil
alexandre.defelicibus@accamargo.org.br

Rodrigo D. Drummond

Laboratory of Computational Biology
Bioinformatics
CIPE/A.C. Camargo Cancer Center
São Paulo, Brazil
rddrummond@gmail.com

Renan Valieris

Laboratory of Computational Biology
Bioinformatics
CIPE/A.C. Camargo Cancer Center
São Paulo, Brazil
renan.valieris@accamargo.org.br

Yu-tao Zhu

China Branch of BRICS Institute of
Future Networks
ShenZhen, China
zhuyutao@bifnc.cn

Israel Tojal da Silva*

Laboratory of Computational Biology
Bioinformatics
CIPE/A.C. Camargo Cancer Center
São Paulo, Brazil
itojal@accamargo.org.br

Liang Zhao*

Faculty of Philosophy, Science, and
Letters
University of São Paulo (USP)
Ribeirão Preto, Brazil
zhao@usp.br

ABSTRACT

Machine learning (ML) has emerged as a powerful tool for improving the clinical pathology routine; however, developing novel ML research requires a complex multidisciplinary team effort. Therefore, a software for tackling the challenges of effectively annotating, building, and validating ML models is desirable. In this work we present WSI2ML, a web-based platform that provides a friendly interface suited for ML computational pathology research to be easily performed. Compared to similar tools currently available, the proposed software provides a complete toolset for each stage of ML workflow. We demonstrate the usefulness and functionality of WSI2ML by analyzing the performance results obtained in a tissue recognition task using a novel gastric cancer research dataset that is currently being developed with the tool. The software, documentation, installation instructions and related annotation handling library is freely available at <https://luanvcmartins.github.io/WSI2ML/>.

CCS CONCEPTS

• **Applied computing** → **Imaging**.

*Both authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WebMedia '23, October 23–27, 2023, Ribeirão Preto, Brazil

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0908-1/23/10...\$15.00

<https://doi.org/10.1145/3617023.3617038>

KEYWORDS

bioinformatics, image tagging, whole slide image

ACM Reference Format:

Luan V. C. Martins, Adriana Passos Bueno, Alexandre Defelicibus, Rodrigo D. Drummond, Renan Valieris, Yu-tao Zhu, Israel Tojal da Silva, and Liang Zhao. 2023. WSI2ML– An Open-Source Whole Slide Image Annotation Software for Machine Learning Applications. In *Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia '23)*, October 23–27, 2023, Ribeirão Preto, Brazil. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3617023.3617038>

1 INTRODUCTION

With higher computational power and storage and increasingly lower costs, several healthcare-related fields have benefited from the new technology-based approaches to patient care. A prominent example is the field of digital pathology [8], in which novel exciting Machine Learning (ML) techniques, especially deep learning, have been proposed for improving clinical practice by detecting clinically actionable biomarkers directly from histological slide images [8]. Further, AI-based methods are used for automation, grading, and immunohistochemistry scoring and thus can help pathologists in their clinical routine [23]. At the forefront of new ML research in digital pathology, high-resolution Whole Slide Images (WSI) are commonly used to assess histological features [10, 18].

Often, the first step in proposing novel ML techniques is to obtain a representative dataset for training, which requires large amounts of WSIs to be collected and annotated. As such, developing new ML algorithms is a multidisciplinary effort made only possible by the collaboration between ML experts and domain specialists (such

as pathologists), who can provide the data for training such models and the technical expertise to analyze the ML predictions. To achieve success, being able to coordinate the people is of foremost importance, requiring carefully created tools and procedures to make the process easier. In this context, although some software for annotating WSI has been developed, such as Cytomine [16], EXACT [17], and OpenHI [19], they generally lack certain important features for a complete ML workflow, such as annotation revision and model validation functionalities. Furthermore, they are often an ensemble of different analysis tools packed into one software, lacking the adaptability and user-friendliness required for handling the already complex and multifaceted tasks of data annotation, generating a high entry cost and making it challenging to integrate pathologists and ML experts to effectively use the software, which, in turn, hinders the interplay of the interdisciplinary team.

Herein, we propose an end-to-end workflow to assist in this multidisciplinary collaboration by removing the burden of project organization. The proposed software allows the creation and validation of datasets by experts effortlessly and facilitates ML experts to build and apply their computational models. The software and the proposed workflow it enables has the potential to help in key data-related challenges in the adoption of AI in pathology [20] when compared to the toolset of currently available alternatives, as described in Section 2. Overall, the tool allows: i) project managers to assign and overview, and direct the WSI tagging efforts; ii) pathologists to annotate WSIs with a variety of tools; iii) domain-specialist to revise previously made annotations; iv) ML specialists to export the data for training machine learning models and v) pathologists to validate and revise model made predictions, in turn, allowing them to improve the dataset over time. We demonstrate the proposed software's utility as a research tool by reporting preliminary results obtained in tissue recognition tasks using a novel gastric cancer research dataset currently being developed with the software.

2 RELATED WORKS

Whole-Slide Images (WSI) are special image files that allow the exploration of high-definition images scanned at up to 40x zoom levels [10], capturing much detail that classic image files are not capable of. In practice, viewing a WSI file is similar to exploring an online map: the user can move the viewed location and zoom in on regions of interest in the image, loading more details on demand. While there are some libraries available that can read WSI files, due to the complexity of dealing with high-resolution zoomable WSI data, few libraries can handle creating and displaying annotations for images of this type. The proposed tool uses the open-source library OpenSlide [9] for loading a wide range of compatible WSI images and has a custom annotation functionality implemented using web technologies (e.g., JavaScript and VueJS).

Over the years, various software tools have been proposed for various challenges in digital pathology. QuPath is one of the first popular open-source solutions designed for WSI analysis [2]. Developed in Java as a desktop application, the software can be extended with plugins and scripting functionality and has been crucial in bringing forth several prominent research projects [12]. Like QuPath, Cytomine [16] is also a comprehensive toolkit for digital pathology tasks, focusing on collaborative effort in image analysis.

In contrast to QuPath, however, it is an online environment, but it can also be extended via plugins and scripting functionality.

More recently, new software tools have been proposed to deal specifically with WSI annotations; prominent examples are EXACT [17], and OpenHI [19]. EXACT offers a comprehensive toolset for annotating images. In addition to WSI files, it can also be used to annotate a wider range of files, including sound histograms or video frames. OpenHI, on the other hand, focuses only on WSI files by employing a pre-segmentation approach where the software segments the image based on its colors using the SLIC technique, and annotators must select which segments they want in their annotation.

While each of the mentioned platforms makes a valuable contribution to the field of digital pathology as a comprehensive toolset for displaying and analyzing images, in our experiments, we found that they still do not have essential features to build new ML-based research datasets, such as annotation review functionality, task-based approach to team management, and model validation tools for measuring progress and improving the data. For example, we could not easily track the progress annotators were making nor evaluate, quantify and improve the performance of ML models using the data the annotators were creating. In conclusion, they posed challenges for us in using them to effectively train and coordinate a team of domain-specialists annotators and to customize their functionality to fit all the required stages of a complete ML workflow.

	Tasks	Revision	Data export	Model validation
QuPath	No	No	Partially	Partially
Cytomine	No	No	Partially	Partially
OpenHI	No	No	Yes	No
EXACT	No	Yes	Yes	No
WSI2ML	Yes	Yes	Yes	Yes

Table 1: Key feature comparison between WSI2ML and other available annotations software focusing on Machine Learning tasks.

In our proposal, we streamline the focus of the application for image annotating, revision, and ML model validation. As such, the software implements a complete whole slide image for machine learning workflow, as illustrated in Figure 1. A comparison of the described systems focusing on features for WSI to ML research is available in the Table 1.

3 SYSTEM AND METHODS

In this section we will detail the features of the proposed software and their importance in the enabled ML workflow, illustrated in Figure 1. We developed WSI2ML using modern web-based technologies. For the front-end, the software uses *Vue.js* with the *Vuetify* framework, enabling the usage of Material Design guidelines for user interface. The backend is implemented in Python ≥ 3.6 , with the following libraries: *Flask* for the web-server functionality, *openslide* for opening and handling WSI images and *SQLAlchemy* for abstracting database management, allowing users to use to any database supported by the library.

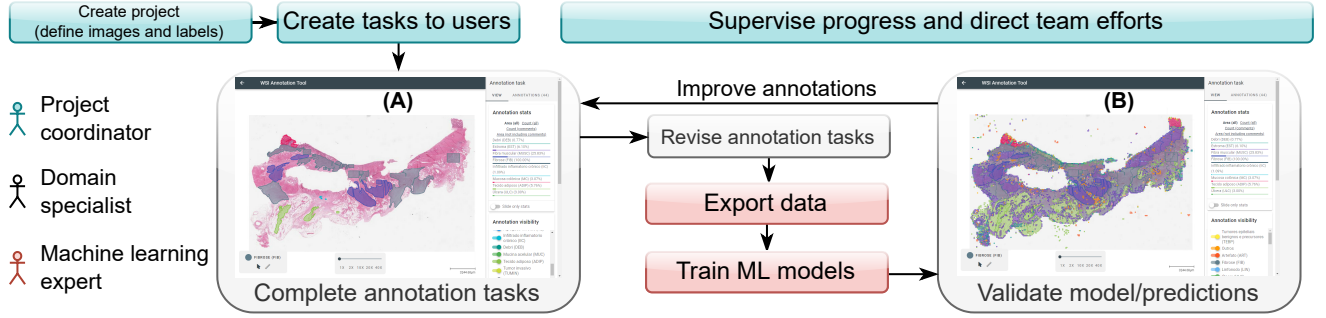


Figure 1: ML-based workflow for building dataset with the proposed tool. The user roles are defined by their permissions, which dictates if they can manage the tasks, projects, users and apps

3.1 Coordinating teams and managing the platform

One of the most complex aspects of interdisciplinary research is coordinating people with varying degrees of computer skills. Therefore, the proposed platform takes a permission-based approach to user roles. Once a user is registered in the system, they can access different screens by the assigned permissions, see Figure 1. The task-based approach, which assigns users to tasks, allows project coordinators to introduce notions of accountability in the project's progress.

3.2 Annotating slides

The software employs a task-based approach to slide annotation, where tasks are created and assigned to annotators. The goal is to streamline and simplify the experience for annotators as much as possible: once a user logs on through the web interface, they are greeted by the task screen, which contains the status of their work and remaining assignment. After the user chooses a task to continue, the annotation interface opens with the WSI visualizer and annotation tools. The user is free to explore the image through the visualizer, see Figure 1 (A). Finally, the user can annotate the slide by selecting a label (defined when creating the project) and the tool most appropriate to delimit the region. The following annotation tools are available: polygon/free drawing, rectangle, and circle.

3.3 Reviewing annotations

Pathology image annotation is subjective and the results will largely depend on the domain-specialist's experience and expectations of the task. As a result, a common challenge to build high-quality datasets is handling diverging annotators' perspectives effectively. The proposed software offers a set of revision tools that allow users to review each other's annotations during the course of project. The revising domain-specialist can mark each annotation as "correct" or "incorrect". Finally, when exporting the data, ML experts can filter out unreliable annotations based on the reviewers' agreement. This process ensures only trustworthy annotations are used, depending on the exporting options.

3.4 Exporting annotations

Once tasks are completed, ML experts can export the annotations using the *Export screen*, which is available for users with the *Can export annotations*. On the screen, the tool allows to customize which slides, annotators, and revision options to generate the data. One of the platform's goals is to be agnostic and non-opinionated in how the ML experts will use or handle the annotations. Therefore, the platform exports data in the *geojson* [3] format, an industry-standard format supported by many open-source libraries. It allows developers to implement their logic for extracting tiles or preparing the data in any other novel format (such as graphs).

3.5 Validating models and improving datasets

An essential step in conducting ML-related research is to validate the ML model and improve the dataset based on its performance. Once a model is trained, the ML experts can generate slide predictions, which can be fed back into the platform for annotators to visualize the results and, if needed, improve the annotations. This process enables a feedback loop where the dataset can be enhanced by increasingly improving the annotations based on the prediction performance of the ML model. Researchers can also use the interactive interface to visualize their results, see Figure 1 (B).

4 PROOF OF CONCEPT

For the purpose of illustrating the usefulness and functionality of the proposed software and ML workflow, we report and analyze our progress and preliminary results in creating a novel Hematoxylin and Eosin-stained (H&E) histopathological image dataset and employ it for tissue recognition [4]. This dataset was created using the proposed software over the course of 1 year. In this section, we demonstrate how the software was used and the results we obtained in the task of tissue recognition (image classification task).

4.1 Context and goals

Machine Learning can assist clinical practice analysis by providing tools to enhance the work of pathologists, helping them to provide diagnostics. Novel ML approaches may offer several benefits to clinical practice, such as an increase in accuracy, efficiency, and cost-saving [18]; indeed, one of the main challenges ML-based tools can improve is the interobserver variability, in which the experience

and care of the pathologist may play a vital role in the fairness of their analysis, this is the case of gastric cancer. Gastric cancer is an uncontrolled growth of cells in the stomach region, being the 5th most common cancer worldwide and the 4th leading cause of death [21]; yet, there is still a lack of a large enough open-source image dataset for deep learning research in this particular region. Indeed, a crucial step in novel research is to identify the tissue types that compose the specimen, which is an image classification task that requires annotated tissue samples for training. For this purpose, the software was employed in creating a novel gastric cancer tissue recognition dataset.

4.2 Image collection and annotation

To start this research project, we retrospectively analyzed archived H&E-stained pathology slides obtained from gastric cancer patients (N=46). First, H&E, a contrast technique used both during clinical practice and ML applications [7], is applied to the specimen to highlight their structures [7]. The contrast-activated samples are then scanned with a high-resolution glass scanner capable of capturing the image at 40x zoom, generating a WSI file. The images are then exported as SVS files, which must be anonymized by converting them to the TIF format and removing all personal information from their metadata. The file is then placed in a project's folder for further analysis by the proposed software.

Once the images are ready to be annotated, we registered a new project using the software by providing the project name, description, images folder, and labels for ten histological categories, which are both the tissues type and the dataset labels (see Figure 2 (B) for the list). Next, the tasks were created and assigned to the domain specialists, who will analyze and annotate the images. Then, throughout the annotation phase of the process, the project coordinator monitored the progress and directed the efforts. When required, the ML specialists exported the data periodically and trained machine learning models to analyze the performance and identify errors. Finally, once a new version of the dataset is created and used to train an ML model, the model predictions were fed back into the software for validation – the domain specialists visualized the predictive results of the model, provided feedback, and improve the dataset for the next iteration. This process can be repeated until the dataset can match quality expectations, controlled by the project coordinator.

4.3 Dataset preparation

There are many ways to work with histological images; as a result, the software exports the annotations as *geojson* files containing the area annotated and the associated metadata. This approach allows researchers to employ their preferred ML approach, such as working directly with the image pixels or other novel representation methods, like cell or tissue graphs [1]. In this section, we will focus on analyzing the tissue recognition performance of the dataset by employing well-known Deep Learning CNN models.

Due to the massive size of Whole Slide Images, they cannot be used directly in classic image classification settings. Instead, we select and classify small areas of 224x224 pixels, called tiles. By classifying the tiles, the model can understand the composition of the

whole surgical specimens, which may serve as features for downstream tasks, such as making prognostics or creating personalized treatment strategies [13]. The training procedure of CNN requires extracting tiles from the annotated WSIs. This is accomplished by iterating over each annotation and listing the tile coordinates available in the annotated area, ensuring that they stayed within the region's bounds, as illustrated in Figure 2 (A). For this purpose, we also developed an open-source library capable of handling the public geojson file format, loading the annotations, and generating tiles with several customization options. The library is available on GitHub¹.

4.4 Experimental setup and results

We train and test the performance of the dataset on two prominent Deep Convolutional Neural Networks on different depth configurations: the ResNet [11] model (on versions of 18, 34, 50 and 101 layers) and the ConvNext [14] model (on versions tiny, small, base and large). For our analysis, we selected a set of 7 independent patients containing a combined total of 21 annotated WSI files – this set and their annotations are fixed throughout the different dataset versions.

The software was used to create multiple versions of the dataset by training the models on the current version, generating predictions, and loading them back into the software, where the domain specialists were able to analyze the model's mistakes and improve their annotations. The improvements were primarily increasing the sample areas that domain specialists found have commonly been mistaken by the models. This process was repeated several times over the course of a few months. This section analyzes 4 dataset versions created throughout this process.

The process of training and testing was standardized among dataset versions. First, all CNN models were initialized with ImageNet weights [6]. Then, for training, we employ a sampling-based approach for creating each batch to avoid biasing the model towards better-represented classes: instead of presenting all images on the dataset, we provide one label at a time. The goal is to ensure that each batch contains samples of all labels of the dataset at the same time – this training procedure is also facilitated by the open-source library we provide, which automates this sampling approach. Given a label, a tile of that class is randomly selected to be included in the batch. It's important to notice that due to color differences in contrast created by different laboratories, working with H&E stained images requires the tiles to be color normalized to allow it to generalize to new WSIs created by other lab techniques or procedures [22]. In this project, we apply the Macenko color normalization method [15], a well-established normalizer for H&E stained images (see [5]). Next, we apply augmentation techniques to create varying samples of the same tissue type by rotating, flipping, and changing the saturation and brightness of the tiles. Finally, we trained each model over 7 epochs with a 64-image batch size.

4.4.1 Tile overlap parameter impact. One of the most essential tile generator parameters is the overlap percentage, which dictates how many pixels are skipped when creating the next tile. In practice, a tile overlap of 1/2 indicates each tile shares roughly half of the

¹<https://github.com/luanvcmartins/WSI2ML-Tools>

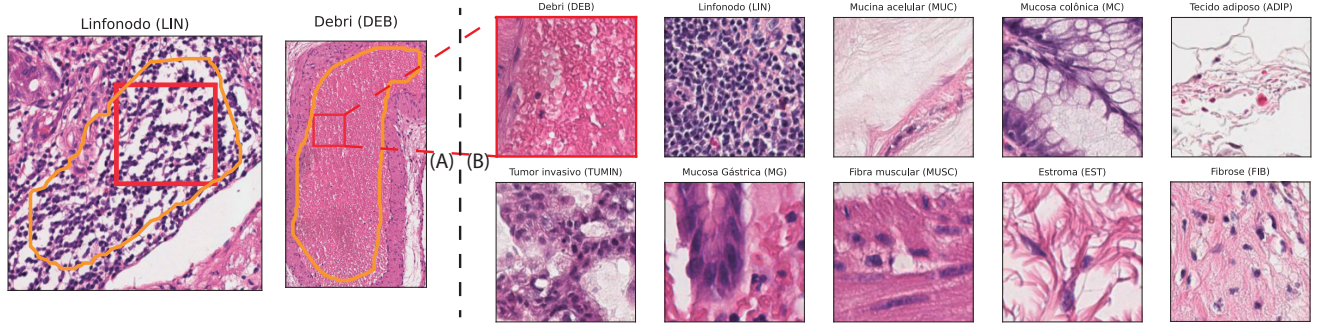


Figure 2: Dataset generation example from the exported data. (A) The annotations are exported using the JSON format, containing the geometry and the metadata. The ML specialist is free use any approach they want. (B) For our approach, tiles must be extracted from the annotations by sampling valid regions from it. The additional library is able to automate this process of generating tiles from the image.

same pixel region with the next tile. By controlling the overlap parameter, we can generate more tiles from the same annotation, allowing for a diverse sampling of the annotated area; on the other hand, a too-high overlap value will generate a high amount of tiles with lower patterns or texture diversity, which may decrease the predictive performance of the model.

	No overlap	1/2	2/3	3/4
Version #1	0,650	0,702	0,701	0,686
Version #2	0,850	0,850	0,856	0,853
Version #3	0,890	0,893	0,891	0,890
Version #4	0,889	0,893	0,895	0,889

Table 2: F1 score of the test set obtained for each overlap parameter on the ConvNext-base CNN. The highest predictive performance for each dataset version is highlighted in bold.

Table 2 shows the predictive performance obtained on the ConvNext-base model for each dataset version using varying levels of overlap. The results demonstrate that using an overlap level between 1/2 and 2/3 of pixels improves the model’s accuracy compared to using no overlap at all, and interestingly, setting a too-high value of overlap results in diminishing returns, as bigger regions of the same pixels containing the same patterns and textures are reused in increasingly more tiles.

4.4.2 Deep CNN predictive performance. Table 3 shows the F1 score obtained on the fixed test set. The largest ResNet model with 101 layers can better fit the data, producing the most accurate tissue recognition results with the latest dataset version with a 90,2 F1 score. However, interestingly, the ConvNext model obtained a higher predictive performance on smaller models and with early dataset versions, with its deepest version staying short of reaching the overall best performance of the ResNet 101 layers. This indicates that the ConvNext model is a better choice for low memory contexts where training larger models is not feasible; additionally, the ConvNext representation was better able to distinguish the different tissue types when training with fewer data, as indicated by its performance on early versions of the dataset. It’s worth noticing that the

deepest version of the ResNet only reaches 1.3% higher performance than the shallowest version of ConvNext on the same dataset version. Finally, the performance improvement over dataset iterations also demonstrates the effectiveness of the workflow enabled by the proposed software in building and validating high-quality datasets using whole slide images, illustrated in Figure 1. The models were trained using an A100 GPU with 40GB of memory.

5 CONCLUSION

This paper presented WSI2ML, an open-source tool for annotating, building, and validating machine learning models applied to Whole Slide Images, a high-resolution zoomable format commonly used in medical image applications. We demonstrated how WSI2ML provides a streamlined workflow for WSI analysis employed in modern digital pathology for novel ML-based research, with tools and functionality for each step of the process. We also illustrated the platform’s usefulness by providing a preliminary overview of a new histopathology dataset created by A. C. Camargo Cancer Center pathologists and ML researchers that will be employed in novel pathology ML-based research. Our work contributes to advancing digital pathology and ML research by facilitating the collaboration between domain specialists and ML experts and enabling the creation and validation of high-quality datasets for various ML tasks. Future work will focus on expanding the functionality and usability of WSI2ML and applying it to other types of medical images and ML problems.

ACKNOWLEDGMENTS

his work was carried out at the Center for Artificial Intelligence (C4AI-USP), with support by the São Paulo Research Foundation (FAPESP grant #2019/07665-4 and #2020/15129-2) and by the IBM Corporation. The WSI Annotation Tool was developed in collaboration with A. C. Camargo Cancer Center.

REFERENCES

- [1] David Ahméd-Aristizabal, Mohammad Ali Armin, Simon Denman, Clinton Fookes, and Lars Petersson. 2022. A survey on graph-based deep learning for computational histopathology. *Computerized Medical Imaging and Graphics* 95 (2022), 102027.

	ResNet				ConvNext			
	18 layers	34 layers	50 layers	101 layers	Tiny	Small	Base	Large
Version #1	0,618	0,636	0,658	0,666	0,679	0,702	0,701	0,693
Version #2	0,835	0,836	0,848	0,853	0,853	0,852	0,850	0,858
Version #3	0,875	0,871	0,882	0,900	0,884	0,883	0,893	0,892
Version #4	0,864	0,876	0,887	0,902	0,889	0,877	0,896	0,891

Table 3: F1 score of several versions of the ResNet and ConvNext models over the different versions of the dataset created with the software using validating functionality.

- [2] Peter Bankhead, Maurice B Loughrey, José A Fernández, Yvonne Dombrowski, Darragh G McArt, Philip D Dunne, Stephen McQuaid, Ronan T Gray, Liam J Murray, Helen G Coleman, et al. 2017. QuPath: Open source software for digital pathology image analysis. *Scientific reports* 7, 1 (2017), 1–7.
- [3] Howard Butler, Martin Daly, Allan Doyle, Sean Gillies, Stefan Hagen, and Tim Schaub. 2016. *The geojson format*. Technical Report.
- [4] Mingyu Chen, Bin Zhang, Win Topatana, Jiasheng Cao, Hepan Zhu, Sarun Juengpanich, Qijiang Mao, Hong Yu, and Xijun Cai. 2020. Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. *NPJ precision oncology* 4, 1 (2020), 14.
- [5] Francesco Ciompi, Oscar Geessink, Babak Ehteshami Bejnordi, Gabriel Silva De Souza, Alexi Baidoshvili, Geert Litjens, Bram Van Ginneken, Iris Nagtegaal, and Jeroen Van Der Laak. 2017. The importance of stain normalization in colorectal tissue classification with convolutional networks. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE, 160–163.
- [6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.
- [7] Shujian Deng, Xin Zhang, Wen Yan, Eric I-Chao Chang, Yubo Fan, Maode Lai, and Yan Xu. 2020. Deep learning in digital pathology image analysis: a survey. *Frontiers of medicine* 14 (2020), 470–487.
- [8] Amelie Echle, Niklas Timon Rindtorff, Titus Josef Brinker, Tom Luedde, Alexander Thomas Pearson, and Jakob Nikolas Kather. 2021. Deep learning in cancer pathology: a new generation of clinical biomarkers. *British journal of cancer* 124, 4 (2021), 686–696.
- [9] Adam Goode, Benjamin Gilbert, Jan Harkes, Drazen Jukic, and Mahadev Satyanarayanan. 2013. OpenSlide: A vendor-neutral software foundation for digital pathology. *Journal of pathology informatics* 4, 1 (2013), 27.
- [10] Matthew G Hanna, Anil Parwani, and Sahuasapont Joseph Sirintrapun. 2020. Whole slide imaging: technology and applications. *Advances in Anatomic Pathology* 27, 4 (2020), 251–259.
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [12] MP Humphries, P Maxwell, and M Salto-Tellez. 2021. QuPath: The global impact of an open source digital pathology system. *Computational and Structural Biotechnology Journal* 19 (2021), 852–859.
- [13] Cesare Lancellotti, Pierandrea Cancian, Victor Savevski, Soumya Rupa Reddy Kotha, Filippo Fraggetta, Paolo Graziano, and Luca Di Tommaso. 2021. Artificial intelligence & tissue biomarkers: advantages, risks and perspectives for pathology. *Cells* 10, 4 (2021), 787.
- [14] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11976–11986.
- [15] Marc Macenko, Marc Niethammer, James S Marron, David Borland, John T Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E Thomas. 2009. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE international symposium on biomedical imaging: from nano to macro*. IEEE, 1107–1110.
- [16] Raphaël Marée, Loïc Rollus, Benjamin Stévens, Renaud Hoyoux, Gilles Louppe, Rémy Vandaele, Jean-Michel Begon, Philipp Kainz, Pierre Geurts, and Louis Wehenkel. 2016. Collaborative analysis of multi-gigapixel imaging data using Cytomine. *Bioinformatics* 32, 9 (2016), 1395–1401.
- [17] Christian Marzahl, Marc Aubreville, Christof A Bertram, Jennifer Maier, Christian Bergler, Christine Kröger, Jörn Voigt, Katharina Breininger, Robert Klopffleisch, and Andreas Maier. 2021. EXACT: a collaboration toolset for algorithm-aided annotation of images with annotation version control. *Scientific reports* 11, 1 (2021), 1–11.
- [18] Sambit K Mohanty and Anil V Parwani. 2022. Whole slide imaging: applications. *Whole Slide Imaging: Current Applications and Future Directions* (2022), 57–79.
- [19] Pargorn Puttapirat, Haichuan Zhang, Yuchen Lian, Chunbao Wang, Xiangrong Zhang, Lixia Yao, and Chen Li. 2018. OpenHI-An open source framework for annotating histopathological image. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 1076–1082.
- [20] Jorge S Reis-Filho and Jakob Nikolas Kather. 2023. Overcoming the challenges to implementation of artificial intelligence in pathology. *JNCI: Journal of the National Cancer Institute* (2023), djad048.
- [21] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. 2021. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* 71, 3 (2021), 209–249.
- [22] Thaina A Azevedo Tosta, Paulo Rogério de Faria, Leandro Alves Neves, and Marcelo Zanchetta do Nascimento. 2019. Computational normalization of H&E-stained histological images: Progress, challenges and future potential. *Artificial intelligence in medicine* 95 (2019), 118–132.
- [23] Jeroen Van der Laak, Geert Litjens, and Francesco Ciompi. 2021. Deep learning in histopathology: the path to the clinic. *Nature medicine* 27, 5 (2021), 775–784.