UNIVERSIDADE FEDERAL DE PERNAMBUCO
DEPARTAMENTO DE GENÉTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA
E BIOLOGIA MOLECULAR

JULIO ANTUNES BARRETO LINS

PREDIÇÃO MULTIMODAL DO SUBTIPO MOLECULAR DO ADENOCARCINOMA GÁSTRICO: INTEGRANDO VISÃO COMPUTACIONAL HISTOPATOLÓGICA E ALTERAÇÕES GENÔMICAS NA PRÁTICA MÉDICA

Recife
2025

JULIO ANTUNES BARRETO LINS

PREDIÇÃO MULTIMODAL DO SUBTIPO MOLECULAR DO ADENOCARCINOMA GÁSTRICO: INTEGRANDO VISÃO COMPUTACIONAL HISTOPATOLÓGICA E ALTERAÇÕES GENÔMICAS NA PRÁTICA MÉDICA

Tese apresentada ao Programa de Pós-Graduação em Genética e Biologia Molecular da Universidade Federal de Pernambuco como parte dos requisitos exigidos para obtenção do título de Doutor em Genética e Biologia Molecular.

Orientador(a): Lucas André Cavalcanti Brandão

Coorientador(a): Fernando Maciano de Paula Neto

Recife
2025

JULIO ANTUNES BARRETO LINS

PREDIÇÃO MULTIMODAL DO SUBTIPO MOLECULAR DO ADENOCARCINOMA GÁSTRICO: INTEGRANDO VISÃO COMPUTACIONAL HISTOPATOLÓGICA E ALTERAÇÕES GENÔMICAS NA PRÁTICA MÉDICA

Área de Concentração: Biologia Molecular

Aprovado em 29/ 09 / 2025

Banca Examinadora

Dr. Lucas André Cavalcanti Brandão
Universidade Federal de Pernambuco

Dr. Ronaldo Celerino da Silva
Universidade Federal de Pernambuco

Dr. Paulo Salgado Gomes de Mattos Neto
Universidade Federal de Pernambuco

Dr. Luiz Alberto Reis Mattos Júnior
Universidade Federal de Pernambuco

Dra. Isabelle Freire Tabosa Viana
Universidade Federal de Pernambuco

Recife
2025

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Lins, Julio Antunes Barreto.
Predição multimodal do subtipo molecular do adenocarcinoma gástrico: integrando visão computacional histopatológica e alterações genômicas na prática médica / Julio Antunes Barreto Lins. - Recife, 2025.
181f.: il.

Tese (Doutorado)- Universidade Federal de Pernambuco, Centro de Biociências - CB, Programa de Pós-Graduação em Genética e Biologia Molecular - PPGGBM, 2025.
Orientação: Lucas André Brandão.
Coorientação: Fernando Maciano de Paula Neto.

1. Adenocarcinoma gástrico; 2. Classificação molecular; 3. Redes neurais convolucionais; 4. Ensemble multimodal; 5. Variações genéticas somáticas. I. Brandão, Lucas André. II. Paula Neto, Fernando Maciano de. III. Título.

UFPE-Biblioteca Central

AGRADECIMENTOS

Quero expressar minha profunda gratidão a todos que contribuíram para a realização desta tese. Aqui represento-os todos por aqueles que mais diretamente contribuíram. Meus mais sinceros agradecimentos à minha família pelo apoio essencial. Represento-os todos na gratidão especial à minha mãe, Consuelo Antunes, médica patologista, pioneira em imuno-histoquímica no norte-nordeste que contribuiu grandemente com o presente trabalho. Além de sempre ter sido a inspiração para essa tese, dedicou longas horas na revisão dos diagnósticos das imagens histopatológicas e à marcação de campos com tumores. Destaco o apoio incondicional de minha esposa Paula Lira, seu cuidado com nossos 3 filhos nos momentos de necessário afastamento para escrita da tese e seu fundamental apoio na revisão final. Agradecimentos especiais a Tasso Moraes que participou na implementação dos algoritmos e acompanhou todos os experimentos. Sem seu apoio não teria sido possível realizar essa tese. A Rodrigo Santos que implementou e acompanhou parte dos experimentos com comitês de redes neurais convolucionais. A Álvaro Mello e José Mendonça, médicos patologistas, que orientaram a marcação de campos e contribuíram na reflexão das aplicações práticas. Agradeço também a Victor Soares que nos meses de finalização da tese e do projeto de inovação atuou como moderador de scrum. A Fernando Barbosa Filho e Gustavo Freitas que foram fundamentais em todo o desenvolvimento do escâner de Lâminas. Agradeço ao Prof. Lucas Brandão, meu orientador, por ter-me concedido toda a liberdade na busca dos caminhos dessa tese e, nos momentos necessários, por me fazer persistir nas trilhas já abertas e eliminar outras que eram distrações. Agradeço ao Prof. Fernando de Paula Neto, meu co-orientador, pelas muitas horas dedicadas em reuniões na exploração dos melhores caminhos para os resultados, em especial na implementação das métricas. Agradeço sinceramente aos membros da banca de qualificação, cujos apontamentos foram fundamentais para o aprimoramento do trabalho: Dr. Tsang Ren, Dr. Ronald Moura e Dra. Isabelle Viana. Agradeço também à Financiadora de Estudos e Projetos (Finep) vinculada ao Ministério da Ciência, Tecnologia e Inovação – MCTI, pelo fomento ao projeto de inovação "Sistema de detecção precoce do câncer", aprovado na chamada pública MCTI/FINEP/FNDCT - Tecnologias 4.0.

RESUMO

O adenocarcinoma gástrico é uma neoplasia altamente heterogênea, classificada em subtipos moleculares pelo The Cancer Genome Atlas (TCGA): instabilidade cromossômica CIN, instabilidade de microssatélites MSI, genômica estável GS e associado ao vírus Epstein-Barr EBV. Esses subtipos influenciam prognóstico e terapia, mas ainda não há métodos diagnósticos de rotina convencionados. A presente tese desenvolveu comitês de sistemas preditivos para os subtipos moleculares por meio de inteligência artificial, concatenou modelos de visão computacional em imagens histopatológicas e modelos de aprendizado de máquina em dados genômicos, demonstrando resultados melhores que a literatura. Organizada em quatro capítulos, a tese utiliza dados do TCGA-STAD para treinamento e validação. No primeiro capítulo, comitês de redes neurais convolucionais CNN treinados em recortes de imagens de lâminas inteiras alcançaram precisão macro de 0,79-0,81 e precisão de 1,00 para EBV e MSI, superando abordagens anteriores em classes minoritárias. O segundo capítulo testa se CNN capturam padrões histopatológicos novos para CIN em dataset controlado por morfologia (apenas adenocarcinomas tubulares, OMS-2019), com NASNet-Mobile obtendo AUC-ROC médio >0,70, confirmando predição independente de tipologias conhecidas. O terceiro capítulo identifica painéis genéticos com Florestas Aleatórias (Random Forest) e avaliação de influência cooperativa dos genes para predição. Identifica genes não previamente associados ao câncer gástrico e organiza dois painéis com 18 e 9 genes. O quarto capítulo propõe comitê multimodal G.SUBTGENOVISION concatenando MobileNetV2 com Random Forest em painel de genes influentes. O modelo apresentou média macro AUC-ROC de 0.95, obtendo AUC-ROC CIN (0,91), EBV (0,98), GS (0,90), MSI (0,99) superior à literatura. A presente tese contribuiu ao demonstrar que modelos de aprendizado profundo revelam padrões histológicos subjacentes a genótipos e portanto denominados fenótipos profundos que podem ser concatenados com dados genômicos em comitês multimodais eficientes. Contribui ainda com a prática médica ao desenvolver em paralelo sistema de inovação e comitê multimodal demonstrando a superioridade dessa abordagem na predição de subtipos moleculares.

Palavras-chaves: adenocarcinoma gástrico; classificação molecular; redes neurais convolucionais; ensemble multimodal; variações genéticas somáticas.

ABSTRACT

Gastric adenocarcinoma is a highly heterogeneous neoplasm classified into molecular subtypes by The Cancer Genome Atlas (TCGA): chromosomal instability (CIN), microsatellite instability (MSI), genomically stable (GS), and Epstein–Barr virus-associated (EBV). These subtypes influence prognosis and therapy, but no standardized diagnostic methods are yet available in clinical practice. This thesis developed predictive system ensembles for molecular subtypes through artificial intelligence, integrating computer vision models on histopathological images with machine learning models on genomic data, achieving results superior to those reported in the literature. Organized into four chapters, the thesis uses TCGA-STAD data for training and validation. In the first chapter, convolutional neural network (CNN) ensembles were trained on whole-slide image tiles, achieving macro accuracy between 0.79 and 0.81 and perfect accuracy (1.00) for EBV and MSI, surpassing previous approaches for minority classes. The second chapter tests whether CNNs capture novel histopathological patterns for CIN in a morphology-controlled dataset (only tubular adenocarcinomas, WHO-2019), with NASNet-Mobile obtaining a mean AUC-ROC > 0.70, confirming prediction independent of known histological types. The third chapter identifies genetic panels using Random Forests and evaluates cooperative gene influence for prediction. It identifies genes not previously associated with gastric cancer and organizes two panels with 18 and 9 genes. The fourth chapter proposes the multimodal ensemble G.SUBTGENOVISION, integrating MobileNetV2 with Random Forests trained on influential gene panels. The model achieved a mean macro AUC of 0.95, with AUC-ROC values for CIN (0.91), EBV (0.98), GS (0.90), and MSI (0.99), all higher than those reported in the literature. This thesis contributes by demonstrating that deep learning models can reveal histological patterns underlying genotypes—thus termed deep phenotypes—that can be integrated with genomic data in efficient multimodal ensembles. It also contributes to medical practice by developing, in parallel, an innovation system and multimodal ensemble demonstrating the superiority of this approach in predicting molecular subtypes.

Keywords: gastric adenocarcinoma; molecular classification; convolutional neural networks; multimodal ensemble; somatic genetic variations.

LISTA DE ILUSTRAÇÕES

TESE

CAPÍTULO 1: G.SUBTVISION – SUBTIPAGEM MOLECULAR DO CÂNCER GÁSTRICO COM MÉTODOS DE ENSEMBLE DE REDES NEURAIS CONVOLUCIONAIS (CNNS)

CAPÍTULO 2: REDES NEURAIS CONVOLUCIONAIS CLASSIFICAM SUBTIPO MOLECULAR DO CÂNCER GÁSTRICO EM DATASET TUBULAR-CONTROLADO

LISTA DE TABELAS

TESE

CAPÍTULO 1: G.SUBTVISION – SUBTIPAGEM MOLECULAR DO CÂNCER GÁSTRICO COM MÉTODOS DE ENSEMBLE DE REDES NEURAIS CONVOLUCIONAIS (CNNS)

CAPÍTULO 2: REDES NEURAIS CONVOLUCIONAIS CLASSIFICAM SUBTIPO MOLECULAR DO CÂNCER GÁSTRICO EM DATASET TUBULAR-CONTROLADO

CAPÍTULO 3 – G.SUBTFOREST: CLASSIFICADOR DE SUBTIPOS MOLECULARES DO CA GÁSTRICO COM TCGA VIA RANDOM FOREST E PAINÉIS OTIMIZADOS

CAPÍTULO 4 - G.SUBTGENOVISION: SISTEMA ENSEMBLE MULTIMODAL PARA CLASSIFICAÇÃO DOS SUBTIPOS MOLECULARES DO ADENOCARCINOMA GÁSTRICO COM IMAGENS HISTOPATOLÓGICAS E PAINEL DE MUTAÇÕES

DISCUSSÃO GERAL

LISTA DE ABREVIATURAS, SIGLAS E SÍMBOLOS

Ac	Anticorpo
AUC-ROC	Área sob a curva *Receiver Operating Characteristic* - Características do Receptor
CA	Câncer
CAP	College of American Pathologists [Colégio Americano de Patologistas]
CG	Câncer Gástrico
CNN	Convolutional Neural Network [Rede Neural Convolucional]
IA	Inteligência Artificial
IARC	International Agency for Research on Cancer [Agência Internacional para Pesquisa sobre o Câncer]
IDH	Índice de Desenvolvimento Humano
INCA	Instituto Nacional do Câncer
OLGA	Operative Link on Gastritis Assessment
OLGIM	Operative Link on Gastric Intestinal Metaplasia Assessment
OMS	Organização Mundial de Saúde
SNV	Single Nucleotide Variant [Variante de nucleotídeo único]
UICC	Union for International Cancer Control [União Internacional para Controle do Câncer]
VC	Visão Computacional
WSI	Whole Slide Image [Imagem de Lâmina Inteira]
CA	Câncer
CAP	College of American Pathologists [Colégio Americano de Patologistas]
CG	Câncer Gástrico
CNN	Convolutional Neural Network [Rede Neural Convolucional]
IA	Inteligência Artificial
IARC	International Agency for Research on Cancer [Agência Internacional para Pesquisa sobre o Câncer]
IDH	Índice de Desenvolvimento Humano
INCA	Instituto Nacional do Câncer
OLGA	Operative Link on Gastritis Assessment
OLGIM	Operative Link on Gastric Intestinal Metaplasia Assessment

OMS	Organização Mundial de Saúde
SNV	Single Nucleotide Variant [Variante de nucleotídeo único]
UICC	Union for International Cancer Control [União Internacional para Controle do Câncer]
VC	Visão Computacional
WSI	Whole Slide Image [Imagem de Lâmina Inteira]

SUMÁRIO

1 INTRODUÇÃO

A Patologia é o tronco central da árvore da medicina moderna e estuda os processos de adoecimento. Ela revela as transformações estruturais e funcionais patogênicas, sendo o elo entre as ciências básicas e os variados ramos da medicina, tendo destaque particular no diagnóstico do câncer. A presente tese foca particularmente no câncer gástrico CG. A escolha do objeto de estudo se deu pela vivência do autor dirigindo o laboratório Ampliar Patologia, principalmente pela quantidade de casos desse tipo de câncer e pelo frequente envio de materiais resultantes de gastrectomias radicais ao laboratório. Tradicionalmente, a classificação dos tipos de câncer gástrico foi baseada em características morfológicas e histopatológicas, o que, embora útil, limita a capacidade de identificar de maneira precisa as nuances moleculares que podem impactar diretamente nas opções terapêuticas e no prognóstico dos pacientes. O diagnóstico do CG é realizado por médicos especialistas em anatomopatologia através de microscopia ótica em campo claro de lâminas histopatológicas de biópsias extraídas do paciente. Os médicos patologistas correlacionam as suas observações nas lâminas com dados clínicos.

Com o avanço das técnicas de sequenciamento molecular e da bioinformática, surgiram novas abordagens para compreender melhor a biologia subjacente dos diferentes tipos de câncer gástrico. Entre elas, destaca-se o trabalho do projeto STAD (stomach adenocarcinoma) do The Cancer Genome Atlas (TCGA) que estabeleceu uma classificação molecular para o CG baseada em dados multiômicos, como genômicos e transcriptômicos.

Essa classificação tem o potencial de revolucionar a prática clínica ao permitir tratamentos personalizados e direcionados, mas enfrenta um obstáculo importante: as tecnologias de sequenciamento multiômicas empregadas na classificação. Essas têm acesso restrito na maioria dos serviços de saúde e apresentam custos elevados. Esse cenário reforça a necessidade de desenvolver métodos alternativos que permitam predizer os perfis moleculares do CG de forma precisa e acessível. Um primeiro método é o estabelecimento de painéis de imuno-histoquímica como em (KIM et al., 2016). Essa abordagem, embora clássica para outros tipos de câncer, na classificação molecular do CG ainda está em desenvolvimento inicial.

Outro método investigado são as Redes Neurais Convolucionais (convolutional neuroneworks CNN´s) treinadas por supervisão molecular (WANG et al., 2022; FLINNER et al., 2022). Isso porque com o desenvolvimento recente da patologia digital, as lâminas passaram a poder serem digitalizadas inteiras. Esse fato oportunizou o uso de redes neurais para a identificação de padrões nas imagens que auxiliem no diagnóstico. As CNN 's vem demonstrando resultados com grande acurácia com treinamento supervisionado por rótulos atribuídos por médicos especialistas em uma quantidade crescente de problemas. Permitindo análises quantitativas onde antes só era possível análises qualitativas. Esse sucesso, no entanto, tem o viés de tomar os rótulos manualmente atribuídos como verdade de base.

A ideia de supervisão molecular no treinamento de redes neurais foi proposta como solução disruptiva. Ou seja, do ponto de vista computacional é um treinamento supervisionado, porém, os rótulos são advindos diretamente de dados moleculares, portanto sem viés humano (MONJO et al., 2022). Essa abordagem, utilizada na presente tese, propicia o uso de algoritmos de aprendizado de máquina e redes neurais como ferramenta de investigação científica a partir de rótulos moleculares.

Considerando a natureza interdisciplinar da presente tese, foram incluídos três apêndices inéditos destinados a uma breve introdução às três principais áreas correlacionadas na tese como apêndices A,B e C. Escritos no estilo de divulgação científica e sugere-se que os leitores iniciem a leitura por eles conforme suas eventuais necessidades de familiarização dependendo da sua área de origem. O apêndice A intitulado "Sistemas preditivos na medicina do século XXI" apresenta a necessidade e oportunidade da aplicação de sistemas preditivos computadorizados com o aumento exponencial de informação sobre os pacientes. O apêndice B intitulado "Breve Apresentação à Multiômica" apresenta os conceitos fundamentais dos dados multiômicos para os leitores que não são da área. O apêndice C intitulado "Breve Apresentação ao Aprendizado de Máquina e à Visão Computacional" apresenta os conceitos fundamentais dos métodos de aprendizado de máquina e redes neurais para os leitores que não são da área. Esses três apêndices foram escritos para promover o tipo de colaboração interdisciplinar que foi necessária para a realização do presente trabalho. São o resultado das explicações dadas pelo autor para o alinhamento de sua equipe. Importante notar que nesses textos a fluidez e riqueza de metáforas foram preferidas ao rigor acadêmico, assim, embora busquem

estar corretos em sua apresentação das ideias principais, foi dada preferência à didática da apresentação e da formação do imaginário do leitor. Pretende-se futura publicação desses apêndices como divulgação científica.

A presente tese segue a apresentação por Artigos. Quatro artigos escritos para serem publicados como artigos científicos em revista indexada, com foco na Bioinformatics. O primeiro artigo intitulado "G.SUBTVISION – SUBTIPAGEM MOLECULAR DO CÂNCER GÁSTRICO COM MÉTODOS DE ENSEMBLE DE REDES NEURAIS CONVOLUCIONAIS (CNNS)" trata da classificação de imagens histopatológicas de câncer gástrico em subtipos moleculares por CNN. Desenvolve ensemble de múltiplas arquiteturas demonstrando resultados melhores que a literatura.

O segundo artigo intitulado: "REDES NEURAIS CONVOLUCIONAIS CLASSIFICAM SUBTIPO MOLECULAR DO CÂNCER GÁSTRICO EM DATASET TUBULAR-CONTROLADO" avalia descoberta de atributos com a organização de novo conjunto de dados controlado para a tipo histopatológico. Esse artigo busca avaliar uma possível refutação à aplicação de CNN para a classificação de subtipos moleculares em imagens histológicas. O terceiro artigo intitulado: "G.SUBTFOREST – CLASSIFICADOR DE SUBTIPOS MOLECULARES DO CA GÁSTRICO COM TCGA VIA RANDOM FOREST E PAINÉIS OTIMIZADOS" utiliza algoritmo floresta aleatória (Random Forest) em variantes genéticas somáticas, identifica genes mais influentes e painéis diagnósticos.

O quarto artigo intitulado: "G.SUBTFOREST – CLASSIFICADOR DE SUBTIPOS MOLECULARES DO CA GÁSTRICO COM TCGA VIA RANDOM FOREST E PAINÉIS OTIMIZADOS"desenvolve um sistema para a predição do subtipo molecular do CG por comitê (Ensemble) multimodal.

É digno de nota que a tecnologia necessária para levar os resultados da presente tese à prática médica foi desenvolvida em paralelo à tese. Isso, pois o autor, durante o seu doutorado, escreveu e dirigiu o projeto "Sistema de detecção precoce do câncer"aprovado na chamada pública MCTI/FINEP/FNDCT - Tecnologias 4.0. O sistema decorrente foi chamado Pathoscope e já possibilita o acesso aos modelos computacionais em imagens histopatológicas à prática médica com solução integrada de escaneamento de lâminas inteiras e visualizador em nuvem. A inovação é brevemente apresentada na sessão "Outras Produções durante o vínculo com o PPGGBM".

1.1 OBJETIVOS

1.1.1 Objetivo geral

Desenvolver modelos de predição diagnóstica para os subtipos moleculares do adenocarcinoma gástrico em imagens histopatológica e em dados genômicos.

1.1.2 Objetivos Específicos

- Desenvolver modelos *ensemble* com múltiplas arquiteturas de redes neurais convolucionais CNN´s para predição dos subtipos moleculares do adenocarcinoma gástrico em imagens histopatológicas. (Capítulo 1)
- Investigar se redes neurais convolucionais treinadas com imagens histopatológicas mantêm desempenho preditivo significativo em subtipagem molecular quando aplicadas a um conjunto de dados histologicamente controlado, composto apenas por adenocarcinomas gástricos tubulares, avaliando sua robustez frente à redução da heterogeneidade morfológica. (Capítulo 2)
- Identificar genes influentes na classificação do subtipo molecular e construir painéis com alto poder preditivo. (Capítulo 3)
- Desenvolver modelo preditivo por comitê multimodal integrando redes neurais com imagens histopatológicas e florestas aleatórias com painel de genes. (Capítulo 4)

2 REVISÃO DA LITERATURA

2.1 O CÂNCER GÁSTRICO

2.1.1 Incidência e mortalidade do Câncer Gástrico

A estimativa para o triênio de 2023 a 2025, a mais atualizada disponível, aponta que ocorrerão **704 mil casos novos de câncer** no Brasil, dos quais **483 mil** excluem os casos de *câncer de pele não melanoma*. Este último, apesar de ser o mais numeroso, com **220 mil casos novos (31,3% do total bruto)**, é frequentemente excluído das análises comparativas por apresentar *alto índice de cura, baixa letalidade e comportamento clínico menos agressivo*. Assim, as comparações entre os cânceres mais relevantes em termos de morbimortalidade são realizadas com base nos **483 mil casos restantes** Tabela 1. (Instituto Nacional de Câncer José Alencar Gomes da Silva, 2022) Considerando este número como base, os cânceres mais incidentes no Brasil no período estimado:

Tabela 1 – Cânceres mais incidentes no Brasil (Triênio 2023-2025), exceto pele não melanoma.

Tipo de Câncer	Nº de Casos Novos (Estimativa)	Incidência Relativa (%)
Mama	74 mil	15,3%
Próstata	72 mil	14,9%
Cólon e reto	46 mil	9,5%
Pulmão	32 mil	6,6%
Estômago	21 mil	4,3%

Base de cálculo: 483 mil casos novos (total exceto câncer de pele não melanoma). Fonte: (Instituto Nacional de Câncer José Alencar Gomes da Silva, 2022).

Ao se considerar somente os tipos de câncer com maior impacto clínico e epidemiológico, excluindo o *câncer de pele não melanoma*, os tipos mais frequentes entre os homens, no triênio de 2023 a 2025, totalizam aproximadamente **136 mil casos novos** Tabela 2. Os principais tipos e suas proporções relativas são:

A taxa ajustada de incidência, segundo o INCA, é 17% maior em homens (185,61) do que em mulheres (154,08).

Existe grande variação na incidência entre as diferentes Regiões do Brasil. As Regiões Nordeste e Norte, possuem os menores IDH e apresentam uma distribuição diferente das regiões de maior IDH. Em homens, o câncer de próstata é predominante em todas as Regiões, mas, para as de maior IDH, os de cólon e reto ocupam a segunda ou a terceira

Tabela 2 – Cânceres mais incidentes em homens no Brasil (Triênio 2023-2025), exceto pele não melanoma.

Tipo de Câncer (Homens)	Nº de Casos Novos (Estimativa)	Incidência Relativa (%)
Próstata	72 mil	52,9%
Cólon e reto	22 mil	16,2%
Pulmão	18 mil	13,2%
Estômago	13 mil	9,6%
Cavidade oral	11 mil	8,1%

Base de cálculo: Aprox. 136 mil casos novos em homens (total exceto câncer de pele não melanoma).

posição, enquanto, para as regiões de menor IDH, o câncer de estômago é o segundo ou o terceiro mais frequente. (Instituto Nacional de Câncer José Alencar Gomes da Silva, 2022)

Foram estimados 21.480 casos novos de câncer gástrico. Ocupando a quinta posição entre os tipos de câncer de maior morbidade. Nos homens, o CG é o segundo mais frequente na Região Norte (12,55 por 100 mil). Na Região Nordeste (12,17 por 100 mil), ocupa o terceiro lugar.

Dentre mulheres, é o quinto mais frequente nas Regiões Sul (8,41 por 100 mil) e Norte (6,53 por 100 mil). Nas Regiões Nordeste (7,46 por 100 mil) e Centro-oeste (6,68 por 100 mil), ocupa a sexta posição.

Segundo o Observatório Global do Câncer em 2020 o CG foi responsável por mais de um milhão de novos casos de CA. As taxas são duas vezes mais alta entre homens que entre mulheres. É o câncer de maior incidência dentre homens em vários países do sul da Asia Central — como Irã, Afeganistão, Turcomenistão e Quirguistão. Globalmente representa 5,6% de todos os CA, entre homens representa 7,1% de todos os CA (SUNG et al., 2021)

No Brasil foram 13.850 óbitos por câncer de estômago em 2020, ocupando a quinta colocação entre os CA que mais matam. Dentre essas pessoas que faleceram, 5.078 foram mulheres. Já entre homens foram 8.772 mortes. (Instituto Nacional de Câncer José Alencar Gomes da Silva, 2022) No mundo todo em 2020 foram 769.000 mortes, ocupando a quarta colocação global entre os CA que mais mataram. É, portanto, um tipo de câncer de importante impacto na população, justificando a atenção científica ao tema.

2.1.2 Topografias e vulnerabilidades associadas

A primeira maneira de classificar os CA é conforme a localização do tumor primário, ou em outras palavras, a topografia. Quando alguém se refere a um CA como "gástrico" esse está fazendo menção ao endereço do tumor, ou seja, que surgiu no estômago. Ao longo da tese diferentes abordagens serão utilizadas para aprofundar a compreensão dos CA, já que tumores de uma mesma topografia muitas vezes são diferentes uns dos outros. Por outro lado, tumores de topografias diferentes podem ser muito semelhantes do ponto de vista histopatológico e molecular. Podendo responder ao mesmo tratamento.

Ou seja, não é necessariamente a localização do surgimento no corpo a maneira mais eficaz de classificação. No entanto, ela é importante e muito útil do ponto de vista epidemiológico e de geração de hipóteses, por existirem indubitavelmente associações estatisticamente significativas entre topografia e diversos outros fatores.

Do ponto de vista topográfico o CG pode acorrer em qualquer compartimento do estômago (Cárdia, Corpo ou Antro). A cárdia é a região do estômago imediatamente após a junção esôfago-gástrica. A oitava edição da UICC considera que deve ser considerado CG o tumor cujo epicentro estiver a mais de 2 cm da junção esôfago-gástrica, mesmo que a acometa(FUKAYAMA; RUGGE; WASHINGTON, 2019).

Sung e colaboradores (SUNG et al., 2021) reforçam a importância da classificação em dois sub-sítios principais: *cárdia* e *não-cárdia*. Cada um desses é associado a diferentes fatores de risco, epidemiologia e carcinogênese. A infecção pelo *Helicobacter pylori* é considerada o principal fator causador do CG *não-cárdia*. Embora a prevalência da infecção pela bactéria seja muito alta, acometendo até 50% da população, apenas 5% dos infectados desenvolverão CA, fatores como alimentação, ingestão de álcool, tabagismo, diferentes cepas e outros diversos fatores que possam ser responsáveis pelas grandes diferenças regionais.

Com o avanço das condições sanitárias, a ampla disseminação da refrigeração de alimentos e a diminuição da prevalência da infecção por *Helicobacter pylori*, observou-se uma redução significativa na incidência dos adenocarcinomas gástricos não-cardia em diversos países desenvolvidos. Essa tendência tem sido interpretada como um “triunfo não planejado” da modernização e das mudanças alimentares, conforme argumentado por Howson et al. (HOWSON; HIYAMA; WYNDER, 1986), que associaram a queda global do câncer gástrico a melhorias ambientais e de higiene, mesmo na ausência de programas de prevenção es-

pecíficos.

Martel e Parsonnet (MARTEL; PARSONNET, 2018) corroboram essa interpretação ao identificarem a infecção por *H. pylori* como o principal fator de risco para os tumores localizados no corpo e antro gástrico (não-cardia), destacando o impacto da erradicação da bactéria na redução desses subtipos tumorais.

Contudo, esse processo foi acompanhado por uma elevação proporcional — e, em determinadas populações, também absoluta — dos adenocarcinomas da cárdia. Powell e McConkey (POWELL; MCCONKEY, 1990) já haviam descrito esse fenômeno, observando um aumento na incidência de tumores situados na junção gastroesofágica, enquanto os tumores distais apresentavam tendência de queda.

Kamangar et al. (KAMANGAR; DAWSEY; BLASER, 2006) demonstraram que a associação entre infecção pelo *H. pylori* e risco de câncer gástrico varia de maneira oposta entre os subtipos cárdico e não-cárdico, reforçando a distinção etiológica entre essas localizações anatômicas. Assim, as estratégias de prevenção centradas na detecção e erradicação da infecção bacteriana parecem exercer maior impacto sobre os tumores distais, enquanto os tumores proximais permanecem menos afetados por essas intervenções.

Dessa forma, o aumento relativo dos tumores da cárdia pode ser interpretado não apenas como um reflexo estatístico da redução dos demais subtipos, mas também como expressão de uma transição epidemiológica, marcada por etiologias distintas e ainda não completamente mitigadas pelas atuais políticas de prevenção.

Com o avanço das condições sanitárias, a ampla disseminação da refrigeração de alimentos e a queda na prevalência da infecção por Helicobacter pylori, observou-se uma redução significativa na incidência dos adenocarcinomas gástricos não-cardia em diversos países desenvolvidos. Essa tendência vem sendo interpretada como um “triunfo não planejado” da modernização e das mudanças alimentares, conforme argumentam (HOWSON; HIYAMA; WYNDER, 1986) que associaram a queda global do câncer gástrico às melhorias ambientais e de higiene, mesmo na ausência de programas de prevenção específicos.

No entanto, essa mesma transição epidemiológica revelou um aumento proporcional — e, em algumas populações, absoluto — dos adenocarcinomas da cárdia. (POWELL; MCCONKEY, 1990) Foi observada uma elevação na incidência dos tumores situados na junção gastroesofágica, em contraste com a queda dos demais tumores gástricos.

Esse padrão, longe de ser paradoxal, encontra respaldo biológico. Kamangar e colaboradores demonstraram que a infecção por *H.pylori* está inversamente associada ao risco

de adenocarcinoma da cárdia (KAMANGAR; DAWSEY; BLASER, 2006).

Portanto, as mudanças de comportamento populacional e as ações preventivas, embora eficazes na prevenção do câncer gástrico distal, não alcançam o mesmo impacto nos tumores da cárdia, cuja etiologia está estatisticamente associada a obesidade e refluxo gastroesofágico, muito provavelmente por condições de inflamação. Assim, o aumento relativo dos tumores da cárdia pode ser interpretado não apenas como uma consequência estatística da queda dos demais subtipos, mas também como reflexo de uma transição etiológica ainda pouco afetada por políticas de saúde pública (OLEFSON; MOSS, 2015).

2.1.3 Prevenção primária

A International Agency for Research on Cancer -IARC completou uma revisão de 40 anos de suas monografias a respeito de carcinogênicos (COGLIANO et al., 2011). Essas monografias são centrais na orientação das ações de prevenção primárias. Foram levantados mais de 100 carcinógenos entre químicos e agentes biológicos, com diferentes níveis de evidências. Embora esses sejam pesquisados para o perigo carcinogênico em uma topografia, por exemplo no estômago, podem provocar câncer em múltiplas topografias, havendo mais concordância com o mecanismo molecular que com a topografia de manifestação (BAAN; STEWART; STRAIF, 2019).

A IARC (PEARCE et al., 2015) estabeleceu critérios de evidência com base em estudos epidemiológicos e de mecanismos moleculares organizando as substâncias químicas, agentes biológicos, comportamentos e predisposições em Grupos (International Agency for Research on Cancer, 2019).

Entre eles, destacam-se a infecção por *Helicobacter pylori*, a exposição ocupacional na indústria de borracha, o tabagismo e a exposição à radiação X e gama. Esses fatores possuem evidência suficiente para serem reconhecidos como causadores do câncer gástrico em humanos (FUKAYAMA; RUGGE; WASHINGTON, 2019).Outros que igualmente tem forte evidência são o amianto (em todas as formas), a infecção pelo vírus Epstein-Barr (EBV), compostos inorgânicos de chumbo, ingestão de nitratos ou nitritos sob condições que favoreçam a nitrosação endógena, consumo de vegetais em conserva (tradicionalmente asiáticos), peixe salgado à moda chinesa e carnes processadas. Esses fatores devem ser considerados em estratégias de prevenção primária, especialmente em populações de risco (FUKAYAMA; RUGGE; WASHINGTON, 2019).

Tabela 3 – Principais características dos carcinógenos descritas por Smith et al. (2016).

#	Características-chave de carcinógenos
1	É eletrofílico ou pode ser metabolicamente ativado para formar um eletrofílico.
2	É genotóxico.
3	Altera a reparação do DNA ou causa instabilidade genômica.
4	Induz alterações epigenéticas.
5	Induz estresse oxidativo.
6	Induz inflamação crônica.
7	É imunossupressor.
8	Modula efeitos mediados por receptores.
9	Causa imortalização celular.
10	Altera a proliferação celular, morte celular ou suprimento de nutrientes.

Os principais fatores carcinogênicos conhecidos para o câncer gástrico são a infecção por Helicobacter pylori, o consumo de alimentos ricos em nitritos e nitratos, o tabagismo, o consumo excessivo de sal e álcool, e determinadas exposições ocupacionais e ambientais. A infecção por H. pylori, especialmente por cepas CagA+, leva à ativação de vias inflamatórias e desregulação de genes supressores tumorais por hipermetilação, como CDH1, p16 e MLH1. Tais alterações epigenéticas contribuem para a progressão da inflamação crônica para metaplasia intestinal e, posteriormente, displasia e adenocarcinoma gástrico (MITHANY et al., 2024; HE et al., 2025).

Dietas ricas em carnes processadas, alimentos defumados e vegetais em conserva expõem a mucosa gástrica a compostos N-nitrosos, que são potentes agentes alquilantes. Esses compostos geram adição de grupos etil e metil ao DNA, provocando mutações somáticas em genes como TP53, frequentemente mutado nos adenocarcinomas gástricos (HE et al., 2025; SHAH; BENTREM, 2022). Além disso, níveis elevados de sal exacerbam o dano à mucosa gástrica e favorecem a colonização por H. pylori, gerando um ciclo pró-inflamatório que acelera a carcinogênese.

O tabagismo é outro fator de risco estabelecido, associado à liberação de nitrosaminas e hidrocarbonetos policíclicos aromáticos, que atuam como genotóxicos diretos. O efeito do fumo está associado ao aumento de mutações pontuais em genes supressores de tu-

mor, além da hipermetilação de promotores gênicos envolvidos na reparação do DNA. Já o álcool, particularmente em grandes quantidades, é metabolizado em acetaldeído, um composto com potencial genotóxico que gera radicais livres e favorece instabilidade genômica (SHAH; BENTREM, 2022; MAZUREK et al., 2024).

Além dos fatores de estilo de vida, exposições ocupacionais (como na indústria de borracha, mineração e agricultura) e ambientais (como a poluição atmosférica por PM2.5) também têm sido implicadas na carcinogênese gástrica. Essas exposições induzem estresse oxidativo persistente, promovendo mutações somáticas e alterações no microambiente tumoral, incluindo desregulação da imunidade local. Estudos recentes reforçam que indivíduos com variantes genéticas em genes como GSTM1, XRCC1 e NAT2 podem apresentar susceptibilidade aumentada frente a esses agentes ambientais, demonstrando a interação entre predisposição genética e exposições externas (HE et al., 2025; MITHANY et al., 2024).

2.1.4 Sinais, Sintomas e Diagnóstico

O câncer gástrico apresenta-se como uma neoplasia de curso insidioso, frequentemente assintomática nas fases iniciais. Quando o paciente apresenta sintomas, são inespecíficos como dor epigástrica leve e dispepsia. Tal padrão clínico contribui para o diagnóstico tardio, sendo um dos principais fatores relacionados à elevada taxa de mortalidade associada à doença. Apenas com o avanço da doença outros sinais mais evidentes como perda de peso corporal e sinais de massa abdominal. O clínico responsável precisa estar atento à epidemiologia do CG na região onde atua. Como demonstrado acima, a incidência do CG varia conforme a região e ao IDH. Assim, a tomada de decisão de qual paciente com sintomas inespecíficos enviar para investigação endoscópica não é amplamente padronizada e existem diferentes padrões internacionais com espaço para ajustes individualizados pelos médicos assistentes. O objetivo é claro e comum, identificar o CG antes de passar a ser invasivo. A identificação precoce permite a cura endoscópica da doença e constitui, portanto, a prevenção secundária.

O consenso é o entendimento geral que pacientes mais velhos devem ser encaminhados para investigação endoscópica ao referirem quaisquer sintomas, já os mais jovens devem ser encaminhados caso apresentem sintomas persistentes à terapia inicial ou se apresentarem sinais de alerta, como perda de peso, ou fadiga. Também é consenso que

sinais de alerta são tardios de não se deve esperar até seu aparecimento. A diferença entre autores e associações de especialistas está no ponto de corte etário para a investigação precoce. Por exemplo, as diretrizes brasileiras não estabelecem uma faixa etária fixa para endoscopia em pacientes assintomáticos, recomendando sua indicação com base em sintomas clínicos e fatores de risco individuais. Todos os pacientes com mais de 40 anos com quaisquer sintomas. Também devem ser submetidos à investigação endoscópica todos os pacientes até 40 anos que apresentem sintomas sem resposta aos tratamentos iniciais e claro que tenham sinais de alerta (BARCHI et al., 2020). Em contraste, no Japão, onde há alta incidência de câncer gástrico, a estratégia populacional de rastreamento populacional endoscópico em indivíduos assintomáticos é implementada nacionalmente a partir dos 50 anos, com intervalos de 2 a 3 anos, como forma de promover diagnóstico precoce e ressecção endoscópica curativa (HAMASHIMA, 2018). Nos Estados Unidos, a *American Society for Gastrointestinal Endoscopy* recomenda que pacientes com dispepsia e idade superior a 50 anos sejam encaminhados à investigação endoscópica (SHAUKAT et al., 2015).

2.2 VIGILÂNCIA ENDOSCÓPICA: PROTOCOLOS OLGA E OLGIM

Durante a realização da endoscopia digestiva alta, toda anormalidade da mucosa deve ser biopsiada. A investigação histopatológica é mandatória nas áreas de metaplasia, atrofia, ulcerações, nódulos, erosões elevadas ou depressões, especialmente no antro e incisura angular. Na ausência de sinais identificáveis à endoscopia de lesões precursoras devem ser realizadas pelo menos duas biópsias, pelo menos uma no corpo e uma no antro. Quando porém há sinais de gastrite atrófica ou metaplasia intestinal outros protocolos devem ser aplicados.

As lesões precursoras são identificadas pelos médicos patologistas e estão diretamente associados ao risco de câncer gástrico. As transformações teciduais acontecem devido processo inflamatório crônico, podendo ser de dois tipos principais: Gastrite Atrófica e Metaplasia Intestinal. Na gastrite atrófica os principais achados são atrofia das glândulas nativas da mucosa gástrica e fibrose da lâmina própria (em substituição à perda glandular). Já na metaplasia intestinal é definida como a substituição do epitélio gástrico nativo por epitélio de tipo intestinal (FUKAYAMA; RUGGE; WASHINGTON, 2019).

É importante avaliar o risco individual de cada paciente para indicar uma periodicidade adequada de vigilância endoscópica para diagnosticar o câncer gástrico em estágio pre-

coce. Para isso o *Operative Link on Gastritis Assessment* OLGA propôs uma classificação com estágios que permite classificar os pacientes em conformidade com o grau de atrofia de mucosa e distinguir os de maior risco dos de menor (RUGGE et al., 2008). O grau da gastrite é determinado histologicamente em cinco estágios (0 a IV) com risco progressivo para o câncer gástrico. A classificação OLGA e sua variante OLGIM (*Operative Link on Gastric Intestinal Metaplasia Assessment*) são sistemas de estadiamento com base em achados de lesões precursoras que estratificam o risco de progressão para CG. A categoria OLGA utiliza a topografia e extensão da gastrite atrófica, enquanto o OLGIM substitui pela metaplasia intestinal como marcador histopatológico (RUGGE et al., 2008; BENITES-GOÑI et al., 2025).

Para uma avaliação diagnóstica adequada da gastrite atrófica e seus riscos associados, recomenda-se a adoção sistemática do Protocolo de Sydney Modificado de padronização de coleta de biopsias, com coleta de cinco fragmentos: dois do antro (menor e maior curvatura), dois do corpo gástrico (menor e maior curvatura) e um da incisura angular (PIMENTEL-NUNES et al., 2019).

O Protocolo de Sydney Modificado e o sistema OLGA são abordagens complementares na avaliação histopatológica da gastrite atrófica. Enquanto o Protocolo de Sydney Modificado padroniza a amostragem endoscópica e a análise histológica da mucosa gástrica, com a coleta de cinco fragmentos (antro, incisura angular e corpo), seu foco está na descrição qualitativa das alterações inflamatórias, presença de *Helicobacter pylori*, metaplasia intestinal e grau de atrofia glandular (STOLTE; MEINING, 2001). Por outro lado, o sistema OLGA utiliza os achados obtidos segundo o Protocolo de Sydney para estabelecer um estadiamento topográfico da atrofia gástrica, classificando os pacientes em estágios de 0 a IV, com base na extensão e severidade da atrofia nas diferentes regiões do estômago. Dessa forma, o sistema OLGA fornece uma estratificação prognóstica do risco e é, portanto, especialmente útil na definição de estratégias de vigilância endoscópica individualizadas.

A decisão de aplicar esses protocolos em casos cuja atrofia ou metaplasia eram previamente desconhecidas deve ser tomada mediante alterações endoscópicas sugestivas de atrofia da mucosa, tais como: visualização acentuada da vasculatura subepitelial, indicando afilamento da mucosa; áreas com descoloração esbranquiçada ou acinzentada, sugerindo metaplasia intestinal; redução das pregas gástricas, padrão mucoso irregular ou com brilho anormal.

Com base no estadiamento histológico obtido por meio dos sistemas OLGA ou OL-

GIM, a conduta clínica deve ser individualizada conforme o grau de risco para neoplasia gástrica. Para os pacientes classificados nos estágios I e II, que representam baixo risco, não há necessidade de vigilância endoscópica programada, devendo ser seguido o protocolo geral, em outras palavras, endoscopia no caso de quaisquer sintomas dispépticos novos. Em contraste, os estágios III e IV estão associados a um risco significativamente aumentado de desenvolvimento de adenocarcinoma gástrico e, por isso, recomendam-se exames endoscópicos periódicos com biópsias em intervalos periódicos. A periodicidade deve levar em conta o risco epidemiológico e pode variar em conformidade com a decisão médica podendo variar a cada 1 ou 2 anos no caso de estágios III ou IV (PIMENTEL-NUNES et al., 2019).

A erradicação do *Helicobacter pylori* é recomendada em todos os pacientes com alterações precursoras. Essa medida pode interromper ou mesmo reverter a progressão de lesões atróficas iniciais, embora a metaplasia intestinal avançada possa persistir. Portanto, o reconhecimento sistemático e o seguimento individualizado das alterações histológicas da mucosa gástrica são fundamentais na prevenção do adenocarcinoma gástrico intestinal. (PIMENTEL-NUNES et al., 2019)

2.3 PROCEDIMENTOS PRÉ-ANALÍTICOS

As biópsias são encaminhadas ao laboratório de patologia devendo-se observar os procedimentos pré-analíticos. O laboratório deve seguir as melhores práticas internacionais ao controlar as amostras e informações dos pacientes e preparar as lâminas histológicas da biópsias encaminhadas. Os médicos patologistas então examinam ao microscópio óptico de campo claro e correlacionam com as informações prestadas pelo médico assistente.

As biópsias devem ser imersas em solução pré-fixadora, formol tamponado a 10%, tão logo quanto seja possível, minimizando o tempo de isquemia fria (COMPTON et al., 2019). Isquemia fria é o intervalo de tempo entre a remoção da amostra do paciente (seja por biópsia ou cirurgia) e sua imersão efetiva no líquido pré-fixador, formalina tamponada a 10%. Durante esse período, o tecido permanece fora do corpo e ainda não está quimicamente estabilizado, tornando-se vulnerável à degradação enzimática e alterações moleculares que podem comprometer a preservação morfológica e molecular. As diretrizes recomendam que o tempo de isquemia fria seja igual ou inferior a 60 minutos, sendo ideal mantê-lo o mais curto possível. É importante notar que o formol penetra a um ritmo de até 1mm/h

podendo ser mais lento em tecidos gordurosos. Assim, caso o material a ser analisado for espesso, mesmo se colocada imediatamente no formol ainda ultrapassaria o tempo de necrose fria no sítio de interesse. Daí a recomendação é que as amostram tenham até 4 mm de espessura (COMPTON et al., 2019). Do ponto de vista prático, nas biopsias as dimensões das amostras já permitem a penetração, já em peças cirúrgicas maiores é importante que o cirurgião realize incisões profundas até a região tumoral, a fim de assegurar a adequada e rápida penetração do líquido pré-fixador.

Embora seja comum, inclusive na literatura científica, referir-se à solução de formaldeído tamponado como “fixador”, essa denominação é tecnicamente imprecisa. O formaldeído exerce uma ação pré-fixadora ao estabilizar parcialmente as estruturas teciduais por meio de ligações cruzadas covalentes entre cadeias laterais das proteínas, mas a fixação definitiva ocorre apenas após o processamento completo do tecido e sua inclusão em bloco de parafina. Este processamento envolve etapas sequenciais de desidratação em banhos de álcool, diafanização em solventes orgânicos (como xilol ou substitutos) e, por fim, a impregnação em parafina aquecida. Somente após esse ciclo completo o tecido encontra-se efetivamente fixado de maneira estável, apto para corte histológico e análises subsequentes.

Após a imersão da amostra na solução de formalina tamponada a 10%, inicia-se a etapa de pré-fixação. No entanto, a exposição prolongada do tecido à formalina pode gerar artefatos morfológicos e comprometer análises moleculares, especialmente de ácidos nucleicos e epítopos proteicos. O aldeído do formaldeído reage inicialmente com grupos amino formando um intermediário instável chamado hidroximetileno, que, com o tempo, pode reagir com outro grupo amino e formar uma ponte de metileno (-CH_2-), estabilizando a estrutura tridimensional das proteínas e da matriz extracelular.

Esse processo é o que preserva a arquitetura histológica das células e tecidos, impedindo a autólise (degradação enzimática) e a putrefação bacteriana. No entanto, esse tipo de reticulação dificulta a extração de DNA e proteínas intactas para análises moleculares subsequentes, já que altera as conformações naturais dos ácidos nucleicos e epítopos antigênicos.

Apesar da estrutura geral do DNA ser preservada com exposição controlada à formalina tamponada 10%, há demonstrada interferência com o aparecimento de mutações artefatuais.

Segundo as diretrizes do CAP, recomenda-se que o tempo total de permanência do

tecido em formalina situe-se idealmente entre 6 e 72 horas. Abaixo desse intervalo, há risco de fixação incompleta; acima, há maior degradação molecular. Assim, tanto a minimização da isquemia fria quanto o controle rigoroso do tempo de exposição ao formaldeído são etapas críticas da fase pré-analítica, com impacto direto na qualidade diagnóstica e na reprodutibilidade de análises histológicas, imunoistoquímicas e moleculares. (COMPTON et al., 2019)

Importante comentar que os procedimentos citados acima se referem ao diagnóstico histopatológico de rotina sendo, portanto, suficientes para histopatologia, imunoistoquímica e hibridização *in situ* validadas. Para os estudos genômicos e transcriptômicos procedimentos mais rigorosos são necessários. Isso porque a exposição prolongada ao formaldeído induz assinaturas mutacionais artefatuais específicas no DNA, predominantemente transições C>T, causadas por desaminação de citosina, e essas alterações podem ser indistinguíveis de mutações reais associadas ao tumor. Essas mutações artefatuais ocorrem mesmo em condições padronizadas de fixação, sugerindo que o material biológico pré-fixado em formalina e embebido em parafina deve ser utilizado com cautela em estudos genômicos. Para minimizar erros, recomenda-se o uso de controles negativos, replicação de amostras e algoritmos de correção de artefatos. Portanto, embora o tecido fixado em formalina ainda seja uma fonte viável para análise molecular, especialmente em contextos clínicos onde material fresco é escasso, deve-se ter plena consciência das limitações e cuidados técnicos exigidos para garantir a fidelidade dos achados genéticos (WILLIAMS et al., 1999; GUO, 2022; SRINIVASAN; SEDMAK; JEWELL, 2002). Por outro lado, para a extração de RNA e, portanto, para a realização de estudos transcriptômicos mencionados na revisão de literatura e na discussão geral, são necessários procedimentos bem mais rigorosos e específicos. A preservação adequada do RNA exige a minimização extrema do tempo de isquemia fria, preferencialmente, o congelamento imediato em nitrogênio líquido.

Cabe destacar ainda que as amostras que participam do banco de dados do TCGA, cujos dados foram utilizados na presente tese, foram obtidas a partir de tecidos frescos congelados pareados a tecidos normais do paciente igualmente obtidos frescos congelados(Cancer Genome Atlas Research Network, 2014).

2.4 CLASSIFICAÇÃO HISTOPATOLÓGICA

Do ponto de vista histopatológico lesões proliferativas epiteliais do estômago podem ser pólipos não-neplásicos, lesões neoplásicas não-invasivas (displasia e adenoma) e adenocarcinomas (FUKAYAMA; RUGGE; WASHINGTON, 2019). Os adenocarcinomas são mais de 95% dos CG (Cancer Genome Atlas Research Network, 2014), sendo propriamente o tipo que advêm do epitélio do estômago. Por isso o adenocarcinoma do estômago é frequentemente referido apenas com câncer de estômago embora outros tipos de neoplasias invasivas também ocorram mais raramente no estômago, como: linfomas, carcinomas sarcomatóides, carcinomas neuroendócrinos, dentre outros (FUKAYAMA; RUGGE; WASHINGTON, 2019).

Sobre as lesões neoplásicas não-invasiva é importante destacar que a Displasia Gástrica é definida como uma lesão precursora do CG com três mudanças histológicas que podem o não estarem todas presentes: Atipia epitelial, diferenciação anormal ou arquitetura mucosa desorganizada (FUKAYAMA; RUGGE; WASHINGTON, 2019). Podendo ser categorizadas como de baixo grau ou alto grau. Na displasia de baixo grau as células neoplásicas apresentam dentre outras características: aberrações arquiteturais, núcleos hipercromáticos alongados e a atividade mitótica é baixa a moderada. Na displasia de alto grau apresentam mudanças arquiteturais com conformações cuboidais ou colunares, perda da polaridade do núcleo, com aumento da razão do tamanho do núcleo em relação ao do citoplasma e mitoses são frequentemente identificadas (FUKAYAMA; RUGGE; WASHINGTON, 2019).

A distinção entre a displasia de alto grau e o carcinoma intraepitelial (*in situ*) é conceitual, isso ocorre pois há diferenças relevantes entre países quanto aos critérios diagnósticos de adenocarcinoma gástrico. Por um lado a maioria dos patologistas na América do Norte, Europa e Coreia exige a demonstração de invasão estromal — ou seja, penetração na membrana basal — para diagnosticar carcinoma. Por outro os patologistas japoneses frequentemente classificam como carcinoma não invasivo lesões com atipia citológica e arquitetural de alto grau, mesmo na ausência de invasão da membrana basal. Em resumo, tais lesões são chamadas de carcinoma não invasivo no Japão, mas seriam classificadas como displasia de alto grau em outras regiões (FUKAYAMA; RUGGE; WASHINGTON, 2019).

O adenocarcinoma origina-se do epitélio glandular (por isso o prefixo *adeno*) da mucosa do estômago, apresentando grande diversidade de apresentações histopatológicas. A classificação histopatológica do câncer gástrico tem evoluído ao longo das décadas,

refletindo avanços no entendimento da biologia tumoral, da epidemiologia e das características moleculares. Iniciando com a classificação proposta por Lauren em 1965, que estabeleceu uma divisão fundamental baseada em padrões morfológicos, o sistema progrediu para abordagens mais detalhadas e integradas, culminando na edição de 2019 da Organização Mundial da Saúde OMS. Essa evolução não apenas refinou a categorização dos tumores, mas também incorporou elementos prognósticos e terapêuticos, auxiliando na personalização do tratamento (LAURÉN, 1965).

2.4.1 Classificação de Lauren (1965)

A classificação de Lauren representa um marco inicial na histopatologia do adenocarcinoma gástrico. Proposta pelo patologista finlandês Pekka Lauren, divide os tumores em dois tipos principais com base em características morfológicas e epidemiológicas:

- **Tipo Intestinal**: Caracterizado por estruturas glandulares bem diferenciadas, semelhantes ao epitélio intestinal. Associado a fatores ambientais, como infecção por *Helicobacter pylori*, e prevalente em regiões de alta incidência de câncer gástrico. Apresenta melhor prognóstico em comparação ao tipo difuso.

- **Tipo Difuso**: Composto por células pouco coesas, frequentemente com morfologia em anel de sinete, infiltrando o estroma de forma dispersa. Relacionado a fatores genéticos e hereditários, com pior prognóstico devido à maior agressividade e tendência a metástases precoces.

- **Tipo Misto ou Indeterminado**: Casos que exibem características de ambos os tipos ou não se enquadram claramente em um deles.

Essa classificação simples, mas robusta, influenciou estudos subsequentes e continua sendo utilizada por sua correlação com padrões clínicos e moleculares (LAURÉN, 1965).

2.4.2 Classificação da OMS de 2019

A edição de 2019 da Classificação de Tumores do Sistema Digestivo da OMS (5ª Edição) é a mais recente, ela refina as categorias e inclui subtipos emergentes:

Tabela 4 – Principais Subtipos de Adenocarcinoma Gástrico na Classificação OMS 2019.

Subtipo	**Características Principais**
Tubular	Estruturas glandulares tubulares.
Papilífero	Projeções papilares com eixo fibrovascular.
Mucinoso	Acúmulo extracelular de mucina (>50% do tumor).
Pouco Coeso (incluindo em anel de sinete)	Células dispersas, pouca adesão, alinhado ao tipo difuso de Lauren.
Misto	Combinação de componentes tubulares e pouco coesos.
Hepatóide	Morfologia semelhante a hepatócitos, com produção de alfa-fetoproteína.
Com Estroma Linfóide	Infiltração linfocítica densa, frequentemente associado a EBV ou MSI.
Micropapilar	Padrões micropapilares invasivos, prognóstico pior.

Fonte: Adaptado de (NAGTEGAAL et al., 2020).

Essa edição enfatiza a correlação com perfis moleculares, como tumores EBV-positivos, MSI-altos, genomicamente estáveis e com instabilidade cromossômica. Além disso, atualiza a classificação de lesões precursoras, como displasia de baixo e alto grau, e integra códigos ICD-O atualizados para melhor padronização global (FUKAYAMA; RUGGE; WASHINGTON, 2019).

A classificação histopatológica da OMS (WHO) de 2019 (FUKAYAMA; RUGGE; WASHINGTON, 2019) para o câncer gástrico representa um avanço significativo na estratificação morfológica desta neoplasia, sendo fundamental para o diagnóstico preciso e orientação do tratamento adequado. Enquanto a classificação de Lauren permanece útil por sua simplicidade, a OMS 2019 oferece uma visão mais abrangente, facilitando a integração com terapias direcionadas baseadas em biomarcadores moleculares.

Desenvolvimento da compreensão dos fenótipos histopatológicos

Tabela 5 – Evolução das Classificações histopatológicas.

Laurén (1965)	**Nakamura et al. (1968)**	**JGCA (2017)**	**OMS (2019)**
Intestinal	Diferenciado	Papilar: pap	Papilífero
-	-	Tubular 1, bem diferenciado: tub1	Tubular, bem diferenciado
Indeterminado	-	Tubular 2, moderadamente diferenciado: tub2	Tubular, moderadamente diferenciado
-	Indiferenciado	Pouco diferenciado 1 (tipo sólido): por1	Tubular (sólido), pouco diferenciado
difuso	em anel de sinete	-	Pouco coeso, fenótipo de célula em anel de sinete
-	-	Pouco diferenciado 2 (tipo não sólido): por2	Pouco coeso, outros tipos de células
Intestinal difuso indeterminado	Diferenciado indiferenciado	Mucinoso	Mucinoso
Misto	-	Descrição de acordo com a proporção (ex.: por2 > sig > tub2)	Misto
Não definido	Não definido	Tipo especial:	Outros subtipos histológicos:
-	-	Carcinoma adenoescamoso	Carcinoma adenoescamoso
-	-	Carcinoma de células escamosas	Carcinoma de células escamosas
-	-	Carcinoma indiferenciado	Carcinoma indiferenciado

Laurén (1965)	Nakamura et al. (1968)	JGCA (2017)	OMS (2019)
-	-	Carcinoma com estroma linfoide	Carcinoma com estroma linfoide
-	-	Adenocarcinoma hepatóide	Adenocarcinoma hepatóide
-	-	Adenocarcinoma com diferenciação enteroblástica	Adenocarcinoma com diferenciação enteroblástica
-	-	Adenocarcinoma do tipo glândula fúndica	Adenocarcinoma do tipo glândula fúndica
-	-	-	Adenocarcinoma micropapilar

Fonte: Adaptado de (FUKAYAMA; RUGGE; WASHINGTON, 2019).

- Adenocarcinoma tubular: caracterizado por estruturas glandulares bem formadas, frequentemente associadas a um estroma desmoplásico. A diferenciação tubular pode variar de bem a pobremente diferenciada, influenciando o prognóstico.
- Adenocarcinoma papilar: apresenta projeções epiteliais digitiformes com eixo fibrovascular. Este subtipo é frequentemente associado a um melhor prognóstico, especialmente quando bem diferenciado.
- Adenocarcinoma mucinoso: definido pela presença de mais de 50% de mucina extracelular. A presença de células em anel de sinete flutuando na mucina é comum, mas não deve exceder 50% das células tumorais.
- Carcinoma pouco coeso: esta categoria inclui o carcinoma de células em anel de sinete e outras variantes com baixa coesão celular. Caracteriza-se por células tumorais isoladas ou em pequenos grupos, frequentemente com morfologia discohesiva. Este subtipo está frequentemente associado a mutações no gene CDH1 e tem implicações prognósticas significativas.

- Carcinoma misto: apresenta uma mistura de pelo menos dois dos subtipos acima mencionados, cada um compreendendo pelo menos 10% do tumor.

Além destas categorias principais, a classificação da OMS 2019 reconhece variantes mais raras e subtipos específicos, incluindo:

- Adenocarcinoma hepatoide: morfologicamente semelhante ao carcinoma hepatocelular, frequentemente associado à produção de alfa-fetoproteína.
- Carcinoma com estroma linfoide: caracterizado por um estroma com proeminente infiltrado linfocítico, frequentemente associado à infecção pelo vírus Epstein-Barr (EBV).

É crucial notar que esta classificação histopatológica tem associação heterogênea com os subtipos moleculares. Por exemplo:

2.4.3 Estadiamento TNM de Carcinomas do Estômago

O estadiamento TNM é um sistema padronizado utilizado para classificar a extensão anatômica dos tumores malignos, permitindo uma avaliação precisa do prognóstico e orientação terapêutica. Em outras palavras, trata-se de uma ferramenta essencial na oncologia que categoriza o tumor com base em seu tamanho e invasão local (T), envolvimento de linfonodos regionais (N) e presença de metástases distantes (M), facilitando a comparação de casos clínicos e a escolha de tratamentos adequados, como cirurgia, quimioterapia ou radioterapia, de forma prática no dia a dia médico (SOBIN; GOSPODAROWICZ; WITTEKIND, 2017a). Esse sistema aplica-se exclusivamente a carcinomas confirmados histologicamente, com considerações específicas para tumores na junção esofagogástrica. Carcinomas cujo epicentro esteja a até 2 cm da junção e sem extensão esofágica são estadiados como gástricos. Em termos práticos, isso significa que o estadiamento ajuda a determinar se o tumor é operável ou requer abordagens neoadjuvantes, influenciando diretamente a sobrevida do paciente (SOBIN; GOSPODAROWICZ; WITTEKIND, 2017b).

As categorias TNM fornecem uma descrição detalhada da progressão tumoral. Em outras palavras, elas dividem o câncer em componentes mensuráveis, permitindo uma avaliação objetiva que guia o planejamento terapêutico e o acompanhamento clínico.

A categoria T avalia a extensão da invasão do tumor primário nas camadas da parede gástrica. Em outras palavras, ela indica quão profundamente o tumor penetrou no estômago, o que é crucial para decidir sobre ressecções endoscópicas em estágios iniciais ou cirurgias mais extensas em casos avançados.

A categoria N quantifica o envolvimento de linfonodos regionais por metástases. Em outras palavras, ela reflete o quanto o câncer se espalhou para gânglios linfáticos próximos, um fator chave para prever o risco de recorrência e indicar linfadenectomia durante a cirurgia. A categoria M indica a presença de metástases em sítios distantes. Em outras palavras, ela identifica se o câncer já se disseminou para órgãos remotos, como fígado ou pulmões, o que geralmente altera o foco do tratamento de curativo para paliativo.

Os estágios combinam as categorias TNM para fornecer uma classificação global da doença. Em outras palavras, eles sintetizam a gravidade do câncer em níveis progressivos, auxiliando na comunicação entre equipes multidisciplinares e na estimativa de sobrevida, como a taxa de 5 anos que varia de quase 100% em estágio 0 para menos de 5% em estágio IV.

O estadiamento TNM apresenta uma abordagem dinâmica, com versões clínicas (cTNM), baseadas em exames pré-tratamento como endoscopias e imagens, e versões patológicas (pTNM), definidas após a ressecção cirúrgica do tumor. Em outras palavras, o cTNM atua como um guia inicial, ajudando médicos a traçar o primeiro plano de tratamento, enquanto o pTNM oferece uma confirmação mais precisa do estágio da doença, permitindo ajustes em terapias adjuvantes para maximizar os resultados clínicos (FUKAYAMA; RUGGE; WASHINGTON, 2019). Compreender o TNM é essencial, pois ele fornece uma estrutura para prever a progressão tumoral e personalizar intervenções. Esse sistema traz uma dimensão humana ao oferecer esperança: o diagnóstico precoce, especialmente nos estágios iniciais (como 0 ou IA), está diretamente associado a taxas de sobrevida significativamente mais altas, que podem ultrapassar 90% em cinco anos, em contraste com menos de 5% nos estágios IV (AMIN et al., 2017).

2.5 APRESENTAÇÃO À MULTIÔMICA

2.5.1 Conceitos fundamentais e genômica

O termo “multiômica” se origina da combinação do prefixo “multi-”, derivado do latim *multus* (muitos), com o sufixo “-ômica”, uma adaptação de “ômica” (do grego *ome*, que significa “todo” ou “conjunto”). Esse sufixo é comumente usado para descrever campos de estudo que se dedicam a explorar de forma abrangente os componentes integrais de sistemas biológicos, como genes, proteínas, e metabólitos. Em um contexto mais amplo, multiômica refere-se à análise integrada de várias dessas camadas biológicas (genômica, transcriptômica, proteômica, metabolômica e epigenômica) em um organismo, visando uma visão mais holística de seus processos moleculares.

O conceito começou a tomar forma nos anos 1990, quando o Projeto Genoma Humano (PGH) avançava rapidamente e revelava, de maneira inédita, a sequência completa do DNA humano. Esse projeto foi o marco inicial de uma era em que o prefixo “-ômica” passou a ser usado com frequência, representando diferentes áreas que analisam coletivamente os componentes biológicos. Embora o PGH fosse inicialmente focado em genômica, o vasto volume de dados e a necessidade de compreensão integrada dos sistemas levaram ao surgimento e popularização da multiômica.

Cronologicamente, a adoção do termo “multiômica” pode ser considerada uma consequência natural do PGH, que impulsionou novas tecnologias e plataformas analíticas capazes de explorar não apenas o DNA, mas outros componentes, como o RNA e proteínas, de maneira integrada. Esse termo começou a ser amplamente utilizado após o PGH, quando surgiu a necessidade de uma abordagem mais complexa para a interpretação dos dados biológicos e suas interações. Esse movimento foi fundamental para o avanço de áreas como a biologia de sistemas, que utiliza a multiômica para mapear e compreender a complexidade das redes biológicas.

Além de “multiômica”, outros termos menos conhecidos fora da comunidade de pesquisadores emergiram. Termos como “epigenômica”, que se refere ao estudo das modificações epigenéticas (alterações químicas que regulam a expressão gênica sem alterar a sequência de DNA), e “interatômica”, que analisa as interações entre proteínas, surgiram na esteira desse avanço, representando a especialização e sofisticação crescente das abordagens ômicas (DURAN, 2023). A epigenômica, por exemplo, “tem permitido o

entendimento dos processos que controlam o envelhecimento e a diferenciação celular, fundamentais para doenças degenerativas" (DURAN, 2023)

Recentemente, os avanços nas tecnologias multiômicas têm promovido grandes descobertas no estudo do envelhecimento celular. A integração de dados transcriptômicos e metabolômicos tem permitido identificar assinaturas moleculares associadas ao envelhecimento celular e à senescência. Esses estudos mostram que "a instabilidade genômica e a acumulação de mutações, anteriormente vistas apenas como consequência do envelhecimento, podem, na verdade, desempenhar papéis ativos nos mecanismos que promovem a senescência celular" (LOPES, 2024). Essa integração de dados, que também considera informações epigenômicas, tem sido crucial para identificar padrões moleculares complexos que delineiam o envelhecimento celular, permitindo uma visão mais detalhada das alterações bioquímicas que ocorrem ao longo do tempo.

No contexto da patogênese do câncer, a multiômica tem se mostrado igualmente revolucionária, especialmente no que diz respeito à compreensão de como diferentes tipos de câncer se desenvolvem e progridem. A aplicação da multiômica possibilita identificar as vias e mutações específicas associadas a tipos específicos de câncer, tendo impacto direto na personalização dos tratamentos. Como observam (PEZZOTTI, 2022), "a multiômica permite uma caracterização mais precisa das células tumorais e facilita o desenvolvimento de estratégias terapêuticas que atacam vulnerabilidades específicas de cada tipo de tumor" (PEZZOTTI, 2022). Esses avanços são fundamentais para o desenvolvimento da medicina de precisão, que visa individualizar o tratamento oncológico com base nas características biológicas e moleculares de cada paciente.

Em suma, a multiômica representa uma abordagem integrativa e essencial para a biologia moderna, viabilizando a análise conjunta de diferentes camadas de dados moleculares e promovendo uma compreensão mais profunda e sistêmica de processos complexos, como o envelhecimento celular e a patogênese do câncer. Esses estudos, que evoluem rapidamente com o apoio de novas tecnologias e métodos computacionais, tornam-se centrais para a medicina personalizada e para estratégias terapêuticas cada vez mais eficazes e direcionadas.

2.5.2 Compreendendo o genoma como código digital

Podemos pensar no genoma como um vasto “código digital” constituído por quatro letras – as bases nitrogenadas, timina (T) citosina (C), guanina (G) e adenina (A) que, ao longo de bilhões de pares, codificam todas as informações necessárias para o funcionamento celular e o desenvolvimento de um organismo. Essa sequência de bases é equivalente a linguagem digital binária, mas em vez de zeros e uns, opera em quatro nucleotídeos que se combinam para formar genes, as “instruções” básicas para as funções celulares. Esse código é composto de DNA e é o ponto de partida da biologia molecular, seu conjunto é o genoma. Nas células humanas há dois genomas, o nuclear e o mitocondrial, organela produtora de energia aeróbica que tem seu próprio DNA independente do nuclear. Quando o termo genoma humano é referido geralmente se refere ao genoma nuclear, caso se faça referência ao genoma mitocondrial se fará menção específica.

Em cada gene humano (como em todos os eucariontes) há regiões chamadas exons e outras chamdas introns. Os exons são por definição participarão da codificação das proteínas, os introns não. Os introns tem muitas vezes funções regulatórias ainda pouco compreendidas durante o splicing altenartivo, pois um mesmo gene pode levar à sintese de duas proteínas diferentes dependendo de quais exons são ligados ou alternativamente cortados. Portanto o conjuto total de todos o material genético que orientará a síntese proteica é chamado de exoma. O exoma tem grande importância prática por representar grande quantidade de informação sobre um individuo a um custo de sequenciamento muito inferior que o genoma e ainda assim representando o código genético de todas as proteínas de um indivíduo.

A primeira camada de ativação desse código ocorre através da transcrição, um processo onde um gene específico no DNA é copiado em uma outras versões da mesma linguaguem quaternária de nucleotídeos, em diferentes tipos de RNA. O RNA mensageiro (mRNA), uma cadeia simples de nucleotídeos, leva uma cópia do código para fora do núcleo da célula, direcionando o maquinário celular para produzir proteínas. Esse conjunto completo dos vários tipos de RNAs expressos em uma célula ou tecido específico é conhecido o transcriptoma. O transcriptoma, portanto, representa a primeira camada dinâmica de expressão do código genético, refletindo quais genes estão sendo transcritos em determinado momento e ambiente celular.

A etapa seguinte é a tradução, onde o código do RNA é novamente traduzido para

uma nova linguagem, essa com 20 aminoácidos. Por isso se chama tradução, pois há a passagem de uma lingua para outra. Cada conjunto de três bases nitrogenadas no RNA (um códon) corresponde a um aminoácido específico. Essa tradução resulta na produção de proteínas, que são sequências de aminoácidos que desempenham funções estruturais, regulatórias e catalíticas essenciais nas células. A sequência final de proteínas expressas em uma célula ou tecido é chamada de proteoma e representa uma camada funcional do código, onde os processos celulares de fato ocorrem e são controlados.

Além dessas camadas genômica e transcriptômica, a epigenômica traz uma camada de regulação adicional ao código. Ela não altera a sequência de DNA em si, mas envolve mecanismos de modificação química, como a metilação do DNA e a modificação de histonas (proteínas ao redor das quais o DNA se enrola). Esses mecanismos regulam quais genes serão transcritos, permitindo ou impedindo o acesso do maquinário de transcrição aos genes. A epigenômica, assim, representa o conjunto das “instruções de uso” que modulam a ativação e a inibição de genes, adaptando a expressão genética ao ambiente e às necessidades específicas da célula em diferentes condições.

Retornando à discussão sobre a multiômica e seu impacto na compreensão da patogênese do câncer, esses avanços permitiram uma visão extraordinariamente detalhada dos mecanismos moleculares subjacentes à transformação celular e à evolução tumoral. A integração multiômica – combinando dados genômicos, transcriptômicos, proteômicos, epigenômicos e metabolômicos — permite que cientistas acompanhem, em alta resolução, como células normais podem se transformar em células malignas. Isso ocorre, em parte, devido à instabilidade genômica, que gera um acúmulo de mutações e rearranjos cromossômicos, alterando o “código” genético e levando à ativação de oncogenes e à inativação de genes supressores de tumor (PEZZOTTI, 2022)

O avanço na tecnologia de multiômica possibilitou identificar assinaturas moleculares específicas de diferentes tipos de câncer, facilitando o desenvolvimento de abordagens terapêuticas personalizadas. Como destacam (LOPES, 2024), “a multiômica permitiu mapear padrões moleculares e metabólicos que são característicos de subtipos tumorais, promovendo um direcionamento mais preciso das terapias” (LOPES, 2024). Por exemplo, no adenocarcinoma gástrico, a análise multiômica revelou diferentes perfis epigenômicos que modulam a expressão gênica de modo a favorecer a proliferação descontrolada e a resistência ao tratamento, possibilitando a criação de terapias que atacam especificamente essas alterações.

Além disso, a integração de dados multiômicos revelou o papel central da instabilidade epigenômica e da remodelação do transcriptoma na evolução de células cancerígenas, tornando-se uma área essencial na pesquisa oncológica atual. Estudos recentes demonstram que alterações epigenéticas podem ocorrer precocemente em células pré-malignas, direcionando-as para estados que favorecem a adaptação clonal e a heterogeneidade celular — uma característica fundamental do câncer avançado (DURAN, 2023). Isso implica que terapias que visam reverter ou modular o epigenoma podem ser promissoras, interrompendo o ciclo de progressão tumoral antes que ele atinja estágios mais agressivos.

Portanto, a multiômica, com sua capacidade de integrar diferentes camadas de dados moleculares, está revolucionando a biologia molecular e a medicina oncológica, promovendo uma compreensão mais profunda e detalhada dos processos que governam a patogênese do câncer. Ao estudar o genoma como um código digital e suas traduções em níveis de expressão gênica, síntese proteica e regulação epigenética, a pesquisa multiômica nos fornece ferramentas fundamentais para mapear e interferir nos processos celulares que levam à transformação maligna, promovendo novas possibilidades para a medicina de precisão.

2.5.3 Cancer e perda da estabilidade do código

A distinção entre SNPs (*Single Nucleotide Polymorphisms*) e CNVs (*Copy Number Variations*) é central para entender como o genoma se comporta e como as variações podem afetar diretamente a estabilidade do material genético ao longo do tempo. Assim como em um sistema de programação, onde uma linha de código corrompida pode levar a falhas no software, pequenas alterações no DNA — seja por SNPs ou CNVs — podem afetar o funcionamento normal das células e, cumulativamente, contribuir para o envelhecimento celular e a progressão de doenças.

Os *SNPs* são variações em nucleotídeos individuais ao longo do genoma. Eles representam uma substituição pontual de uma “letra” do código genético e são a forma mais comum de variação genética entre os indivíduos. Em termos de programação, um *SNP* é como a troca de um único caractere em uma linha de código. Se bem posicionado, esse erro pode ser insignificante, alterando apenas um detalhe menor da função celular. Porém, caso ocorra em uma região crucial — como num trecho que codifica uma proteína ou regula um gene — pode comprometer a função, predispondo a célula a falhas e desregulações.

Já as *CNVs* (Variações no Número de Cópias) são variações mais amplas que envolvem a duplicação ou a exclusão de grandes segmentos do DNA. Em analogia com a programação, uma *CNV* seria equivalente a copiar e colar um bloco de código ou apagar uma seção inteira de uma função. Esse tipo de alteração afeta de forma mais profunda o genoma, pois pode resultar no ganho ou na perda de várias "linhas" de código genético, alterando a dosagem de genes e gerando um desequilíbrio funcional na célula. Quando essas variações incluem genes inteiros ou conjuntos de genes, o impacto é significativo: genes duplicados podem levar à produção excessiva de proteínas, enquanto a ausência de genes pode comprometer funções essenciais.

A *instabilidade cromossômica* se manifesta a partir de uma sucessão de alterações como *SNP*s e CNVs, contribuindo para a "corrupção" progressiva do código genético de uma célula. À medida que as células replicam seu DNA, a presença dessas variações pode gerar erros de leitura ou até impedir que partes do código sejam executadas corretamente, aumentando a probabilidade de defeitos. Com o passar do tempo, especialmente em organismos que envelhecem, esses erros acumulam-se, promovendo a desregulação dos sistemas de controle celular. Esse processo de "envelhecimento molecular" é, assim, um reflexo de pequenas variações e grandes falhas estruturais que comprometem o funcionamento correto do genoma e levam à senescência celular — um estado em que as células param de se dividir, mas permanecem metabolicamente ativas, muitas vezes contribuindo para inflamações e outros distúrbios celulares associados ao envelhecimento.

No contexto do câncer, a instabilidade cromossômica amplificada pela acumulação de *SNP*s e *CNV*s gera uma diversidade clonal dentro dos tumores, fornecendo às células malignas uma "biblioteca" de variações que podem ser exploradas para resistir a tratamentos e prosperar em diferentes ambientes do organismo. Analogamente, seria como um código de software com trechos duplicados e corrompidos que, em vez de travar o sistema, o tornam imprevisível e resistente a tentativas de correção. Essa imprevisibilidade, alimentada pela instabilidade cromossômica, faz com que tumores se tornem geneticamente diversos e mais agressivos.

Estudos recentes têm o entendimento sobre como essas variações, particularmente as CNVs, contribuem para o envelhecimento celular e a patogênese do câncer. Como observado por (DURAN, 2023), "o acúmulo de CNVs ao longo do tempo desempenha um papel chave na instabilidade genômica associada ao envelhecimento celular e à transformação maligna" (DURAN, 2023). Esse processo de acúmulo, uma vez iniciado, se torna difícil de

reverter, pois as células perdem progressivamente a capacidade de corrigir suas próprias falhas, levando a um ciclo de decadência funcional que se torna ainda mais pronunciado em tecidos envelhecidos e tumores.

Assim, compreender a distinção entre SNPs e CNVs e seu papel na instabilidade cromossômica é crucial para desvendar os mecanismos que sustentam o envelhecimento celular e a carcinogênese. A crescente capacidade da biologia molecular de identificar e monitorar essas variações genômicas nos dá uma visão detalhada dos “códigos corrompidos” que moldam tanto o envelhecimento como a progressão do câncer, permitindo o desenvolvimento de intervenções mais precisas que busquem estabilizar ou corrigir essas falhas antes que seus efeitos se tornem irreversíveis.

2.6 APRESENTAÇÃO AO APRENDIZADO DE MÁQUINA E DA VISÃO COMPUTACIONAL

2.6.1 Aprendizado de máquina

O aprendizado de máquina (do inglês *machine learning*) é uma área ciência da computação que busca desenvolver sistemas capazes de aprender. Para compreender o que se quer dizer com aprender nesse contexto primeiro é importante destacar que as pessoas apresentam um aprendizado muito singular. Os animais em geral podem aprender por condicionamento e observação, porém a capacidade de adaptação a novas situações ambientais é extremamente limitada, pois a maior parte dos comportamentos complexos são instintivos. Um pássaro joão-de-barro *Furnarius rufus* demonstra a habilidade peculiar de construir ninhos de barro. Seu ninho é construído com barro e palha e apresenta um design altamente funcional e robusto. É, no entanto, uma expressão clara de instintos genéticos que não se alteram diante de mudanças no ambiente, esses pássaros constroem suas casas sempre iguais. A adaptação de padrões instintivos se dá apenas pela seleção da variabilidade aleatória em cada geração. O organismos vivos tem padrões complexos epigenéticos que podem afetar esse processo, mas para os objetivos da explicação do aprendizado de máquina o leitor se beneficiará desse exemplo do joão-de-barro.

A inteligência humana distingue-se pela sua complexidade cognitiva, social e cultural, que supera amplamente as habilidades observadas em outras espécies. Enquanto os animais adaptam seus comportamentos principalmente com base em instintos e experiências

acumuladas, os seres humanos possuem a capacidade única de transcender essas limitações ao criar ferramentas e sistemas que ampliam suas possibilidades de adaptação. Nesse contexto, o aprendizado de máquina emerge como uma extensão dessa habilidade exclusivamente humana, concebido para replicar, de forma controlada, a capacidade de aprender, formular abstrações e modificar estratégias diante de novas situações.

Inspirado nos processos cognitivos humanos, o aprendizado de máquina busca reproduzir, a capacidade de lidar com incertezas e adaptar-se a cenários diversos. A inteligência de máquina é definida pela habilidade de processar dados e utilizá-los para tomar decisões informadas, algo que se manifesta na análise de conjuntos de dados, na identificação de padrões subjacentes e na geração de previsões ou classificações com base nesses padrões. A partir dessas análises, as máquinas aplicam modelos para determinar a melhor ação em situações variadas, adaptando-se às circunstâncias.

Além disso, o aprendizado de máquina simula aspectos do raciocínio humano, como a identificação de padrões complexos em dados visuais, sonoros ou textuais, essenciais para tarefas como reconhecimento de voz ou imagens. Também inclui a capacidade de resolver problemas e planejar ações por meio de raciocínio lógico estruturado em regras ou informações disponíveis. Outro componente importante é a habilidade de lidar com incertezas, avaliando probabilidades e tomando decisões em cenários onde os dados são incompletos ou ambíguos. Essas características colocam a inteligência de máquina como uma extensão adaptativa da cognição humana, ampliando as capacidades de análise e tomada de decisão em contextos cada vez mais diversos e desafiadores.

Em vez de seguir regras pré-definidas, esses sistemas utilizam dados para identificar padrões, fazer previsões e tomar decisões. Ao contrário dos métodos tradicionais de programação, onde um programador fornece instruções detalhadas para cada tarefa, o aprendizado de máquina permite que o sistema descubra automaticamente a solução. Isso é feito treinando um modelo com um grande volume de dados relevantes.

Os dados são o núcleo do aprendizado de máquina. Eles podem ser estruturados, como tabelas de informações numéricas e categóricas, ou não estruturados, como imagens. Antes de serem utilizados, os dados passam por etapas de pré-processamento, treinamento e avaliação. As técnicas de aprendizado podem ser classificadas em supervisionadas, não-supervisionadas e por reforço, diferenciando-se pela forma como os dados são apresentados ao modelo. Esses tipos de aprendizados serão apresentados adiante.

Do ponto de vista técnico, o aprendizado de máquina baseia-se na construção de mo-

delos matemáticos que processam dados para realizar previsões ou tomar decisões. Esses modelos utilizam métodos estatísticos e computacionais para identificar padrões nos dados e generalizar esses padrões para novos exemplos. O objetivo é minimizar o erro, geralmente representado por uma função de perda, usada para quantificar a discrepância entre as previsões feitas por um modelo e os valores reais observados nos dados. Ajustando os parâmetros ou pesos do modelo durante o treinamento.

Um aspecto técnico importante é a validação do modelo. Para garantir que ele funcione bem em novos dados, divide-se o conjunto de dados em partes, como treino, validação e teste. Ou seja, uma parte dos dados é salva apenas para o teste final. Outra parte é usada para validar os diferentes parâmetros de treinamento, como um teste intermediário. E outra parte é utilizada no treinamento. Isso é fundamental para evitar o problema do *overfitting*, que acontece quando o modelo tem boa acurácia nos dados de treino, mas não generaliza bem para novos exemplos, ou seja a acurácia caí quando o modelo é testado com novos dados. Assim, quando se refere à acurácia de um modelo de aprendizado de máquina se refere a sua acurácia em dados teste, que não devem ter feito parte dos dados usados para o treinamento. Em outras palavras, dado novos para o modelo.

Todos com formação científica tem conhecimento da estatística analítica, muitos menos conhecida é o aprendizado de máquina. Ambas compartilham os algorítimos matemáticos e buscam interpretar dados e produzir insights, mas diferem profundamente em seus enfoques e aplicações. A estatística analítica concentra-se em compreender relações entre variáveis, frequentemente buscando identificar associações ou inferir causalidades. Esse ramo da estatística utiliza métodos como regressões, testes de hipóteses e análise de variância para explicar fenômenos observados. A ênfase está em interpretar os dados disponíveis.

O aprendizado de máquina, por sua vez, adota algoritmos frequentemente originados da estatística analítica, como regressões e árvores de decisão, mas os utiliza de forma inovadora, com foco no aprendizado contínuo e no poder preditivo. Em vez de se limitar à análise estática de dados, como na estatística analítica, o aprendizado de máquina aplica esses mesmos métodos para identificar padrões em grandes volumes de dados e utilizá-los para realizar previsões com maior eficiência e adaptabilidade. Por exemplo, uma regressão linear simples, amplamente usada na estatística analítica para modelar relações entre variáveis, pode ser incorporada em um sistema de aprendizado de máquina para ajustar continuamente seus parâmetros à medida que novos dados são recebidos, melhorando

assim a precisão das previsões.

Dessa forma, o aprendizado de máquina pode ser visto como uma extensão avançada da estatística analítica, que, ao incorporar a entrada contínua de dados com o poder computacional, apresenta maior flexibilidade e escalabilidade. Ao combinar o rigor analítico da estatística com a adaptabilidade computacional, o aprendizado de máquina transforma a maneira como os dados são usados.

2.6.1.1 Aprendizado supervisionado

No aprendizado supervisionado o modelo é treinado utilizando um conjunto de dados rotulados, onde cada entrada é associada a uma saída esperada (*label*). Essa abordagem visa construir uma relação mapeada entre as variáveis de entrada e as saídas, sendo amplamente empregada em tarefas como classificação, onde se deseja prever categorias, e regressão, onde o objetivo é prever valores contínuos. O aprendizado supervisionado pode ser compreendido por meio de uma analogia com o processo de ensino de uma criança a identificar objetos, como uma maçã. Inicialmente, apresenta-se o objeto à criança, informando-lhe o nome correspondente. Após repetidas exposições a exemplos variados, ela aprende a associar características como formato e cor ao nome previamente dado (*label*). Esse processo permite que a criança diferencie uma maçã de outros objetos, como uma laranja. De maneira similar, no aprendizado supervisionado, o modelo computacional é treinado com dados rotulados, isto é, exemplos de entrada acompanhados de suas respectivas saídas esperadas. Ao observar múltiplos exemplos em diferentes contextos, o modelo desenvolve a capacidade de realizar associações precisas entre entradas e *labels*, aprimorando sua habilidade de identificar padrões e realizar predições em cenários futuros.

2.6.1.2 Aprendizado Não Supervisionado

O aprendizado não supervisionado trabalha com dados não rotulados, permitindo que o modelo identifique padrões ou estruturas ocultas nos dados sem a necessidade de instruções explícitas. É frequentemente utilizado em tarefas como agrupamento (*clustering*), que organiza dados em grupos com características semelhantes, e redução de dimensionalidade, onde informações redundantes são eliminadas para simplificar a análise. Esse tipo

de aprendizado pode ser ilustrado pelo processo de uma criança explorando objetos desconhecidos sem instruções explícitas. Imagine que uma criança tenha acesso a diferentes frutas, como maçãs e laranjas, mas sem que ninguém lhe diga os nomes ou características específicas de cada uma. Por conta própria, a criança começa a observar semelhanças e diferenças entre os objetos, agrupando-os com base em características como cor, formato ou textura. Apesar de não saber os nomes das frutas, ela pode distinguir que algumas têm casca lisa e alaranjada, enquanto outras são vermelhas e arredondadas. De forma análoga, no aprendizado não supervisionado, o modelo não recebe dados rotulados, mas busca identificar padrões ocultos e agrupar ou organizar as informações com base em semelhanças entre os dados, permitindo descobrir estruturas subjacentes sem a necessidade de associações explícitas previamente fornecidas.

2.6.2 Floresta Aleatória - *Random Forest*

O *Random Forest* foi apresentado por Leo Breiman no início dos anos 2000, especificamente em 2001, como uma evolução das árvores de decisão e do método de *bagging Bootstrap Aggregating*, que ele mesmo havia proposto anteriormente. A motivação para seu desenvolvimento surgiu das limitações inerentes às árvores de decisão individuais, que, embora sejam modelos intuitivos e poderosos, sofrem com a tendência ao *overfitting*. Essa limitação as tornava pouco confiáveis em cenários de alta variabilidade ou com conjuntos de dados ruidosos.

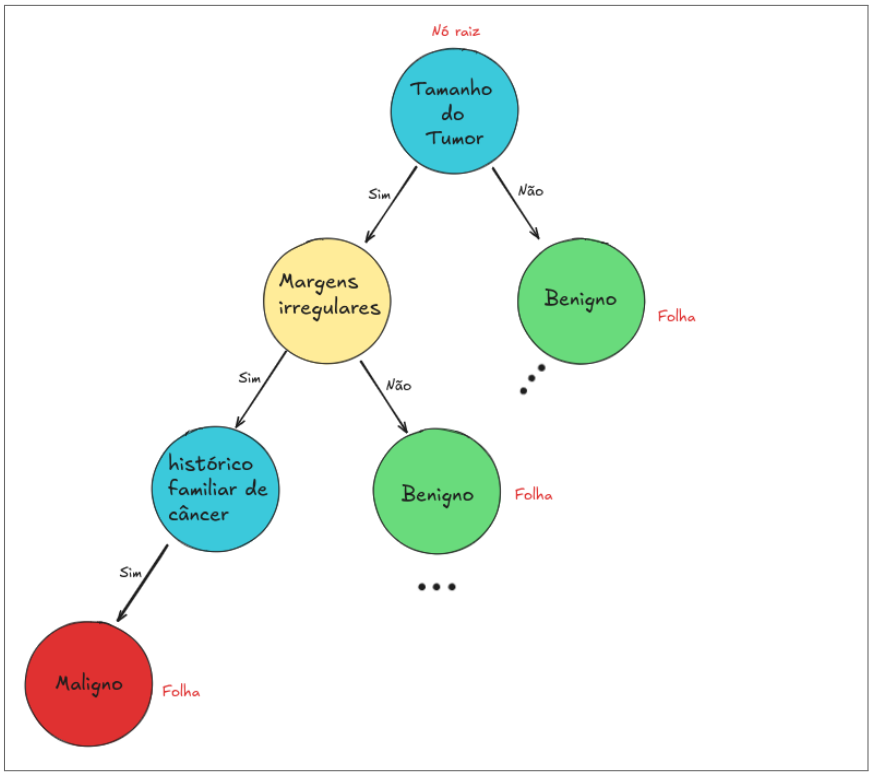
Breiman, com sua abordagem pragmática e engenhosa, combinou a ideia de criar múltiplas árvores de decisão independentes a partir de amostras aleatórias dos dados e introduziu um elemento de aleatoriedade adicional: a seleção aleatória de subconjuntos de características para cada divisão da árvore. Essa combinação estratégica transformou o Random Forest em um modelo capaz de capturar padrões robustos e generalizáveis, enquanto minimizava os riscos associados às peculiaridades dos dados de treinamento.

A árvore de decisão é uma das estruturas mais intuitivas e amplamente utilizadas no aprendizado de máquina, possui uma história que remonta a diversos campos de estudo, incluindo estatística, ciência da computação e psicologia. Essa metodologia foi desenvolvida a partir de conceitos teóricos e evoluiu ao longo do tempo para se tornar um dos pilares da análise de dados, devido à sua simplicidade e capacidade de interpretação. Para compreender sua história, é necessário observar sua origem multidisciplinar e os

avanços que transformaram essa técnica em uma ferramenta fundamental. Os primeiros fundamentos podem ser encontrados em trabalhos de estatísticos como Francis Galton e Ronald Fisher, que estudaram métodos de classificação e análise de variância. Esses estudos estabeleceram bases matemáticas para dividir dados em grupos com base em suas características.

O *Random Forest* é um algoritmo usado no aprendizado de máquina para resolver problemas de regressão e classificação. Combina múltiplas árvores de decisão, cada uma criada com diferentes subconjuntos dos dados e características. No final, ele combina os resultados dessas árvores para gerar uma resposta mais precisa. Uma árvore de decisão é um modelo simples que organiza as decisões em forma de um diagrama ramificado. Ela divide os dados em partes menores com base em perguntas "sim/não", criando uma hierarquia que leva a uma previsão final. Uma árvore de decisão pode ser comparada a um processo de tomada de decisão lógica.

Figura 5 – Esquemático de uma árvore de decisão



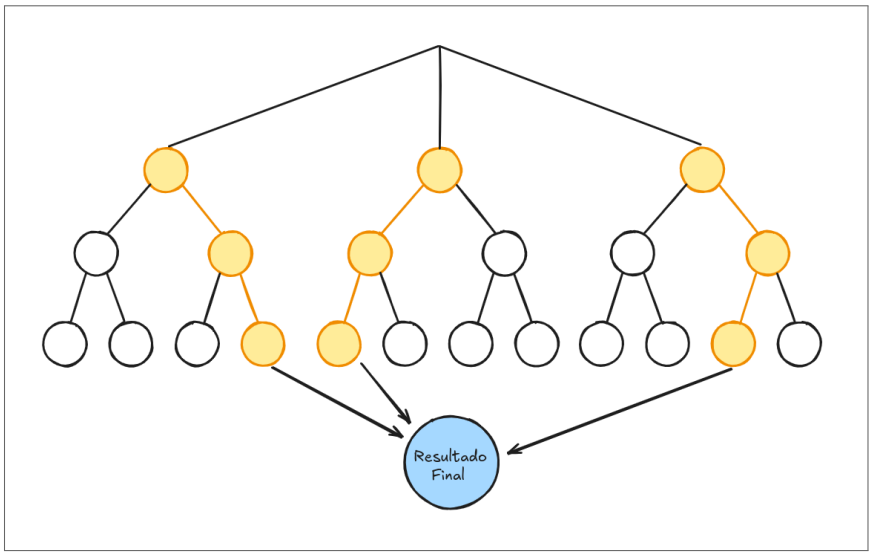
Fonte: Autor

A diferença principal entre uma árvore de decisão e uma floresta aleatória é que uma árvore de decisão funciona como um único caminho lógico: ela começa calculando qual característica tabulada divide os casos em dois grupos com base no maior ganho de in-

formação e na redução da entropia dos dados (outras métricas também eventualmente participam como o índice de pureza dos dados) é o chamado nó raiz. A escolha do nó raiz é baseada na avaliação de todas as características disponíveis, buscando a divisão que mais reduz a incerteza ou aumenta a pureza das classes. Por exemplo, o nó raiz pode ser representado como uma pergunta "há mutação do P53?"dividindo os casos em sim ou não. Segue dividindo os dados em dois grupos. Essa abordagem considera todas as características disponíveis nos dados e tenta encontrar as divisões mais eficazes para classificar cada caso.

Por outro lado, a floresta aleatória não depende de apenas uma árvore. Em vez disso, ela cria várias árvores de decisão, cada uma construída com partes diferentes dos dados e com perguntas baseadas em subconjuntos aleatórios de características. Por exemplo, enquanto uma árvore pode começar analisando "Há mutação do P53?", outra árvore pode começar com "há mutação do MLH1?". Essa aleatoriedade garante que as árvores sejam independentes e avaliem os dados sob diferentes perspectivas. No final, a floresta combina as decisões de todas as árvores o que torna o modelo mais robusto, confiável e menos propenso a erros que poderiam surgir de uma única árvore.

Figura 6 – Esquemático de uma Floresta Aleatória



Fonte: Autor

O Random Forest é frequentemente classificado como um algorítimo clássico porque combina simplicidade e eficácia de forma exemplar. Ele não exige ajustes complexos e é naturalmente robusto, funcionando bem em uma ampla variedade de problemas e tipos de dados, especialmente estruturados. Além disso, sua capacidade de avaliar a importância

das características o torna uma ferramenta valiosa em contextos onde a interpretabilidade é essencial, como na medicina e na genética.

Apesar do advento de arquiteturas mais complexas, como redes neurais profundas, o Random Forest continua sendo amplamente utilizado. Sua relevância se dá não apenas por sua eficiência computacional em comparação com técnicas mais avançadas, mas também porque fornece soluções interpretáveis e confiáveis.

2.6.3 Redes Neurais profundas

A descoberta de que o neurônio cerebral humano, assim como o de outros mamíferos, opera com um mecanismo binário — disparar ou não disparar, transmitir ou não transmitir informação — foi um marco fundamental para a neurociência e para a inspiração computacional no desenvolvimento de redes neurais artificiais. Cada neurônio, com base em estímulos recebidos, se passará o impulso elétrico gerado pelas rápidas e ativas trocas de potássio e sódio pela membrana plasmática ou se o neurônio não disparará e portanto não passará a descarga elétrica a frente. Caso os neurônios disparem aquela via é fortalecida por neuroplaticidade, caso contrário é enfraquecida. Em suma o nurônio humano é binário. Na literatura científica, o termo correto para descrever o "disparo"de um neurônio é "potencial de ação"(*action potential*). Esse processo descreve a rápida mudança de voltagem na membrana celular de um neurônio, que ocorre quando ele transmite um sinal elétrico ao longo de seu axônio. O potencial de ação é um fenômeno "tudo ou nada"(all-or-none), o que significa que o neurônio ou gera o sinal completamente ou não o faz, dependendo se o limiar de excitação foi alcançado.

O fenômeno do potencial de ação, amplamente estudado na neurofisiologia, é um processo eletroquímico que define a base funcional da comunicação entre neurônios. Ele ocorre de maneira "tudo ou nada", o que significa que um neurônio ou dispara completamente, transmitindo o sinal, ou não dispara, caso o estímulo recebido não atinja o limiar necessário. Esse fenômeno inicia-se com a despolarização da membrana celular, quando canais de sódio se abrem e permitem a entrada rápida desses íons, invertendo temporariamente a polaridade da membrana. Após atingir o pico do potencial de ação, os canais de sódio se fecham e os canais de potássio se abrem, promovendo a saída desses íons e restaurando a carga negativa interna no processo conhecido como repolarização. Eventualmente, ocorre uma hiperpolarização transitória, onde o potencial da membrana se torna

ainda mais negativo antes de retornar ao estado de repouso, regulado pela bomba de sódio-potássio. Esse ciclo assegura a propagação unidirecional do sinal ao longo do axônio e permite que o neurônio retome seu estado inicial para disparar novamente quando necessário. A simplicidade do mecanismo contrasta com a complexidade das redes neurais formadas, demonstrando como interações precisas entre bilhões de neurônios podem dar origem a processos cognitivos e comportamentais sofisticados.

Os potenciais graduados são fenômenos regulatórios fundamentais da atividade neuronal, representam pequenas alterações localizadas no potencial de membrana. Esses resultam da interação com estímulos recebidos na superfície celular. Esses potenciais, excitatórios ou inibitórios são acumulativos, permitindo ao neurônio integrar informações provenientes de múltiplas sinapses antes de atingir o limiar necessário para disparar. Essa modulação sináptica, essencial para a regulação da sensibilidade ao estímulo, permite que o neurônio ajuste sua resposta a contextos variados, refinando a transmissão da informação através da rede neural. A inibição, por sua vez, é um mecanismo crucial pelo qual um neurônio pode ser impedido de disparar, frequentemente devido à ação de outros neurônios que, através de sinapses inibitórias, dificultam a despolarização suficiente para o início do potencial de ação. Quando o neurônio efetivamente dispara.

Apesar da complexidade das interações que regulam a atividade neuronal a ação final de cada neurônio é binária. Ele dispara ou não dispara; responde ou permanece inerte; um ou zero. Esse mecanismo "tudo ou nada", sustentado por uma intrincada teia de conexões e processos regulatórios, revela a beleza e a precisão do sistema nervoso desde reflexos básicos até os mais altos processos cognitivos. O cérebro humano contém aproximadamente 100 bilhões de neurônios em rede, mais outros 500 milhões espalhados pelo sistema nervoso periférico, coração e intestino. A rede é densamente conectada pois cada neurônio é capaz de estabelecer conexões com até 10 mil outros neurônios, formando um emaranhado de sinapses cuja contagem total é estimada em 100 trilhões de conexões. Essa vasta rede neural é o que permite a manifestação de propriedades cognitivas complexas, como memória, aprendizado e raciocínio.

No entanto, o que singulariza o cérebro humano em relação ao de outros mamíferos é a organização dessas conexões, especialmente nas regiões associadas a funções cognitivas de alta ordem. O córtex pré-frontal conta com 20 bilhões de neurônios dedicados densamente conectados e hierarquizados. Responsável por funções como planejamento, tomada de decisão, controle de impulsos e raciocínio abstrato, ele ocupa uma proporção

maior no cérebro humano do que em qualquer outra espécie. Esse córtex, ao integrar informações provenientes de áreas especializadas em percepção sensorial, memória e emoções, permite associações de alta ordem que constituem a base da linguagem, criatividade e pensamento crítico.

O conhecimento neurofisiológico dessa estrutura inspirou os cientistas da computação a imaginar sistemas artificiais que pudessem imitar, ainda que de forma rudimentar, essa habilidade de conectar, processar e aprender. A invenção do neurônio artificial foi, portanto, uma tentativa de capturar essa lógica simples, mas potente, de redes interconectadas que definem tanto o cérebro humano quanto as máquinas que aspiram imitá-lo. A ideia de modelar computacionalmente a capacidade adaptativa do cérebro humano foi, talvez, um dos mais audaciosos empreendimentos intelectuais do século XX. Inspirados pela complexidade da mente e pela elegância das conexões neurais, cientistas e pensadores deram início a uma jornada que uniria biologia, matemática e computação em um esforço para criar máquinas que pudessem aprender.

2.6.3.1 A invenção do neurônio

A invenção do neurônio artificial como parte de uma rede integrada, foi o marco inicial de uma transformação que moldaria o futuro da inteligência artificial. Foi na década de 1940 que Warren McCulloch, um neurofisiologista, e Walter Pitts, um lógico matemático, deram o primeiro passo significativo nessa direção. Em seu artigo seminal, eles propuseram um modelo matemático que descrevia um neurônio artificial como uma unidade binária capaz de processar informações e produzir uma saída com base em entradas recebidas. Para McCulloch e Pitts, a chave estava na simplicidade: um neurônio artificial deveria seguir regras lógicas elementares, mas, quando conectado a outros neurônios, poderia realizar cálculos complexos, aproximando-se daquilo que entendemos como aprendizado ou cognição. (MCCULLOCH; PITTS, 1943)

Esse modelo matemático não apenas capturava a essência funcional de um neurônio biológico, mas também introduzia a ideia de que redes de neurônios poderiam ser projetadas para resolver problemas. Essa visão deu origem ao paradigma das redes neurais, onde o comportamento emergente do sistema depende da interação de seus componentes. Entretanto, o impacto desse insight não se restringiu ao campo da neurociência computacional. Ele abriu portas para reflexões filosóficas profundas sobre o que significa aprender,

adaptar-se e tomar decisões. A analogia com o cérebro humano, embora limitada, trouxe consigo questões sobre a natureza da inteligência: seria a inteligência uma propriedade emergente de conexões suficientemente complexas? Ou dependeria de algo mais fundamental, talvez inatingível por máquinas?

À medida que essa ideia evoluía, novas abordagens começaram a surgir, impulsionadas por avanços teóricos e tecnológicos. Em 1958, Frank Rosenblatt introduziu o perceptron, um modelo computacional inspirado no neurônio de McCulloch e Pitts, mas com a capacidade adicional de "aprender"a partir de exemplos. Rosenblatt acreditava que, ao ajustar os pesos das conexões entre os neurônios, uma máquina poderia melhorar seu desempenho ao longo do tempo, reproduzindo, ainda que de forma rudimentar, a plasticidade do cérebro humano. (ROSENBLATT, 1958)

Essa concepção inicial foi recebida com entusiasmo, mas também com ceticismo. Durante décadas, as redes neurais enfrentaram limitações matemáticas e tecnológicas que restringiam sua aplicabilidade. Simplesmente ainda não havia o poder computacional para demonstrar seu potencial. Ainda assim, a ideia de que sistemas computacionais poderiam, um dia, reproduzir aspectos da inteligência humana progrediu academicamente, com a rápida introdução de praticamente todos os novos desenvolvimentos no campo da matemática e com a evolução teórica das arquiteturas. O salto imenso salto recente na capacidade das inteligências artificiais se deve muito mais ao aumento do poder computacional, pois as bases científica já foram há décadas bem fudamentadas.

2.6.3.2 Apresentação à Arquiteturas de redes neurais

A invenção do neurônio artificial é um avanço tecnológico relativamente simples. Imagine o leitor que faça uso de um dos muitos algorítimos da estatística analítica para analisar um conjunto de dados e resultar em uma resposta binária. Agora imagine-se fazendo o mesmo para todos os algorítimos que conheça ou que tenham sido computacionalmente implementados por colegas com mais habilidade matemática de maneira a que cada um desses neurônios está em paralelo avaliando os dados e passando ou não seu "disparo"a frente. Agora imagine-se conectando uma série de neurõnios em camadas subsequentes de dimenções crescentes e ajustando camadas finais para reduzir as dimensões das camadas ao formato necessário a uma tarefa específica.

Entre os neurônios das diferentes camadas o leitor imagine-se atribuindo variáveis de

importância que atuam como os potenciais graduados da biologia, facilitando ou dificultando o disparo daquela conexão essas variáveis são os chamados pesos. Esses pesos são um conceito muito importante a saber. Os pesos vão estabelecer a qual dos neurônios da camada anterior deve ter a informação mais valorizada. Os melhores pesos são os que valorizam a informação mais relevante para o sucesso. Imagine-se atribuíndo pesos aleatoriamente a esses muitos algorítimos em múltiplas camadas e automatizando esse processo de atribuição mais ou menos aleatória de pesos um algorítimo de otimização. Pronto, o leitor já tem sua arquiteura estabelecida. Ainda mais, essa arquitetura com o conjunto de pesos é um modelo a ser testado nos dados.

2.6.3.3 Adaptação e Seleção artificial

Na década de 1990, a ideia de otimização evolutiva começou a ganhar popularidade na inteligência artificial como uma abordagem para resolver problemas complexos por meio de simulações inspiradas em processos biológicos. Esses métodos utilizam conceitos derivados da evolução natural, como seleção, variação e adaptação. O algoritmo genético, uma das principais ferramentas desse período, simulava a evolução de populações de "organismos digitais"— que na verdade são modelos (arquiteruras com conjuntos de pesos) representando possíveis soluções para um problema. Cada organismo, ou modelo, é avaliado com base em sua capacidade de executar uma tarefa, como a maximização de uma função matemática ou a resolução de um problema específico em um ambiente simulado.

Após cada iteração, os melhores modelos são selecionados, suas características são reporduzidas ligeiramente alteradas — em um processo análogo à recombinação genética, é testada. Assim subsequentemente por múltiplas gerações de modelos. Esse é o processo de treinamento é linhas gerais o processo de treinamento nos dados. Cada uma das características é um parâmetro que pode ser alterado pelo cientista de dados, como o número de interações (gerações) qual o grau de variabilidade dos descendentes, quanto maior o salto mais rápido o treino é menos computacionalemnte custoso, porém o risco de alcançar um pico intermédiário de sucesso aumenta.

Essa abordagem, embora rudimentar comparada aos padrões atuais, permitiu grandes avanços na época. No contexto das tecnologias disponíveis na década de 1990, o número de gerações que podia ser simulado dependia da capacidade de processamento dos computadores, que era significativamente limitada. Um experimento típico envolvendo

otimização evolutiva poderia levar dias ou semanas para treinar algumas centenas de gerações, dependendo da complexidade do problema e do tamanho da população de modelos digitais. Em termos de redes neurais, os modelos da época eram relativamente simples, com apenas algumas centenas ou milhares de parâmetros — valores ajustáveis que determinam o comportamento do modelo. Essa simplicidade era uma consequência direta das limitações computacionais e da disponibilidade de dados, que restringiam tanto a profundidade quanto a largura das redes.

Avançando para 2023, o cenário transformou-se do ponto de vista de aumento geométrico de poder computacional, as arquiteturas possuem bilhões de peso. Esses modelos não apenas superam em muito as capacidades das redes neurais dos anos 1990, mas também são treinados em escalas de dados e tempos de computação que seriam impensáveis naquela época.

Além disso, as abordagens atuais frequentemente utilizam treinamento competitivo entre modelos, conhecido como aprendizado adversarial ou aprendizado por competição. Nesse paradigma, dois ou mais modelos são treinados simultaneamente, competindo entre si para aprimorar seu desempenho. Um exemplo clássico são as Generative Adversarial Networks (GANs), onde um modelo gerador tenta criar dados convincentes enquanto outro modelo discriminador tenta identificar falhas nos dados gerados. Esse tipo de abordagem acelera o aprendizado, ao mesmo tempo em que aumenta a robustez do modelo resultante.

Portanto, a metáfora da otimização evolutiva, na qual cada geração de modelos é aprimorada com base nos melhores desempenhos, permanece relevante para entender o desenvolvimento atual da inteligência artificial. O que começou como um processo lento e limitado por hardware modesto transformou-se em uma prática altamente sofisticada, que utiliza recursos computacionais massivos para gerar modelos capazes de resolver problemas com níveis de complexidade impossíveis há poucos anos.

2.6.3.4 Breve histórico das arquiteturas e a introdução das camadas profundas

A primeira aruitetura com diversas aplicações práticas hoje é a perceptron. O perceptron pode ser entendido como um algoritmo que recebe várias entradas e retorna uma saída binária (0 ou 1). Apesar do entusiasmo inicial, a falta de poder computacional e de dados limitou o progresso. Somente na década de 1980, com a introdução do algoritmo de retro propagação (*backpropagation*), foi possível o treinamento de redes neurais multica-

madas, vário perceptrons aninhados em camadas. Esse avanço permitiu a modelagem de problemas mais complexos ao mesmo tempo que métodos estatísticos, como máquinas de vetor de suporte (SVMs) (VAPNIK, 2013) e árvores de decisão (BREIMAN et al., 1984), começaram a ser amplamente utilizados em tarefas de classificação, como classificar e-mails como *span* ou não, e regressão, estimar o preço de uma casa com base em suas características.

Na década de 2000, dois fatores principais impulsionaram o aprendizado de máquina para o mainstream: a explosão de dados digitais, impulsionada pela internet e redes sociais, e a disponibilidade de maior poder computacional, principalmente por meio de unidades de processamento gráfico (GPUs). Esses avanços tornaram viável a aplicação de algoritmos mais sofisticados em problemas reais, como tradução de idiomas, reconhecimento de fala e análise de imagens.

O termo “profundo” no contexto da ciência de dados se refere às camadas das arquiteturas das redes neurais cujos padrões, ou “pesos”, descobertos durante o treinamento nos *datasets*, não são diretamente visíveis ou compreensíveis para o programador. Esses padrões emergem da interação entre os dados e o modelo, e não de regras explícitas programadas. Aqui, utiliza-se o termo “fenótipo profundo” para designar esses padrões identificados pela rede e refletem características complexas que não são evidentes à primeira vista.

As redes neurais profundas, ou *Deep Neural Networks* (DNNs), são um tipo avançado de modelo computacional inspirado no funcionamento do cérebro humano. Elas são formadas por muitas camadas de unidades computacionais chamadas “neurônios artificiais”. Cada camada processa as informações recebidas, simplificando ou combinando padrões, até chegar a um resultado, como reconhecer uma imagem ou prever uma tendência. O “profundo” no nome refere-se à grande quantidade de camadas que essas redes possuem.

Um exemplo simples ajuda a entender: imagine ensinar uma criança a reconhecer fotos de cães. Primeiramente, ela aprende características básicas, como o formato geral do corpo ou o focinho. Depois, começa a reconhecer detalhes, como o pelo ou as orelhas. Da mesma forma, nas redes profundas, as camadas iniciais identificam padrões simples, como linhas e formas, enquanto as camadas mais profundas combinam esses padrões para formar algo mais complexo, como a imagem de um cão.

Essas redes são extremamente úteis em situações onde os dados são muito ricos ou difíceis de interpretar, como fotos, sons ou grandes textos. Por exemplo, em sistemas que

analisam exames médicos, as redes neurais profundas conseguem identificar sinais de doenças, como tumores, que podem ser difíceis de perceber até mesmo para especialistas.

Para que uma rede neural profunda funcione, é necessário “treiná-la”. Isso envolve mostrar muitos exemplos para o modelo e permitir que ele ajuste seus parâmetros, chamados de pesos, para que as respostas sejam cada vez mais precisas. Um processo chamado retropropagação é usado para corrigir os erros cometidos pela rede durante o treinamento, ajustando os pesos das conexões entre os neurônios.

Um dos maiores desafios das redes neurais profundas é que elas precisam de muitos dados e recursos computacionais para funcionar bem. Quanto mais camadas uma rede possui, mais dados são necessários para ensinar a máquina de forma eficiente. Além disso, se o modelo for muito complexo e os dados forem insuficientes, ele pode “memorizar” os exemplos do treinamento sem aprender algo útil para novos casos. Esse problema é conhecido como *overfitting*.

Para evitar o *overfitting*, diversas técnicas são aplicadas. Uma delas é aumentar os dados disponíveis, o que pode ser feito gerando variações dos exemplos, como rotacionar imagens ou alterar ligeiramente suas cores. Outra abordagem é chamada de *dropout*, onde alguns neurônios são temporariamente desativados durante o treinamento, forçando a rede a generalizar melhor.

Apesar de seu enorme potencial, as redes neurais profundas também têm limitações. Uma delas é que as decisões tomadas por esses modelos nem sempre são fáceis de explicar, já que os padrões aprendidos estão “escondidos” nas camadas internas. Isso cria um problema chamado de “caixa-preta”, onde os resultados são confiáveis, mas os passos para chegar até eles podem ser difíceis de entender.

Em resumo, redes neurais profundas são ferramentas revolucionárias que permitem às máquinas realizar tarefas complexas, como reconhecer imagens, traduzir idiomas e até diagnosticar doenças. Elas funcionam aprendendo padrões em grandes volumes de dados, mas exigem cuidado no treinamento e na interpretação de seus resultados. Mesmo para quem não é da área, é importante entender seu funcionamento básico, pois essas tecnologias estão cada vez mais presentes em nossa sociedade.

2.6.4 Visão Computacional

As Redes Neurais Convolucionais (Convolutional Neural Networks — CNNs) são uma classe especializada de algoritmos de aprendizado profundo amplamente aplicadas em tarefas de classificação de imagens, reconhecimento de objetos e outras áreas de visão computacional.

O termo "convolucional", empregado em redes neurais convolucionais (CNNs), refere-se à operação matemática de convolução. Isso é o núcleo funcional da extração de caracterísitcas das redes neurais convolucionais (CNNs). Essa operação é amplamente utilizada para processar imagens, e tem como objetivo extrair características relevantes, como bordas, texturas ou padrões.

É inspirado no funcionamento do córtex visual primário dos mamíferos, responsável pelo processamento inicial de estímulos visuais. Neurônios especializados no córtex visual atuam de forma hierárquica, respondendo a padrões simples, como bordas e texturas, e, em seguida, combinando essas informações para interpretar formas mais complexas. Esse modelo de processamento biológico influenciou diretamente o design das *CNN*s.

Nas redes convolucionais, a operação de convolução simula esse mecanismo biológico por meio da aplicação de filtros (ou kernels) sobre diferentes regiões da imagem. Cada filtro, ao sobrepor-se a pequenas áreas do dado visual é capaz de identificar características locais como bordas ou padrões texturais. Essa abordagem reflete a maneira como o córtex visual humano processa informações visuais, em que regiões específicas do campo visual são interpretadas de forma segmentada e integrada.

Além disso, as CNNs utilizam uma hierarquia de camadas convolucionais, onde as primeiras camadas extraem características mais simples, enquanto as camadas posteriores combinam essas informações em representações mais abstratas e complexas, chamadas de camadas profundas. Processam informações mais amplas e complexas de alta ordem.

Esse processo assemelha-se à organização funcional do córtex visual, onde informações simples se integram progressivamente em níveis mais altos de processamento. Essa abordagem não apenas reforça o vínculo entre sistemas biológicos e computacionais, mas também demonstra como princípios da neurociência podem inspirar avanços na inteligência artificial

As *CNN*s se destacam por sua capacidade de processar dados com estrutura de grade, como imagens, de forma escalável e eficiente (GOODFELLOW; BENGIO; COURVILLE, 2016). O

funcionamento das CNNs baseia-se em uma arquitetura composta por camadas interligadas, onde cada nó (ou neurônio) realiza cálculos com base em pesos e limites pré-definidos Figura 7. Quando a saída de um nó excede um valor de limiar especificado, o nó é ativado e os dados são transmitidos para a camada seguinte da rede. Caso contrário, o nó permanece inativo, e nenhum dado é propagado. Essa característica permite que a rede selecione automaticamente os padrões mais relevantes, reduzindo o ruído desnecessário durante o aprendizado. As *CNN*s são estruturadas em blocos fundamentais:

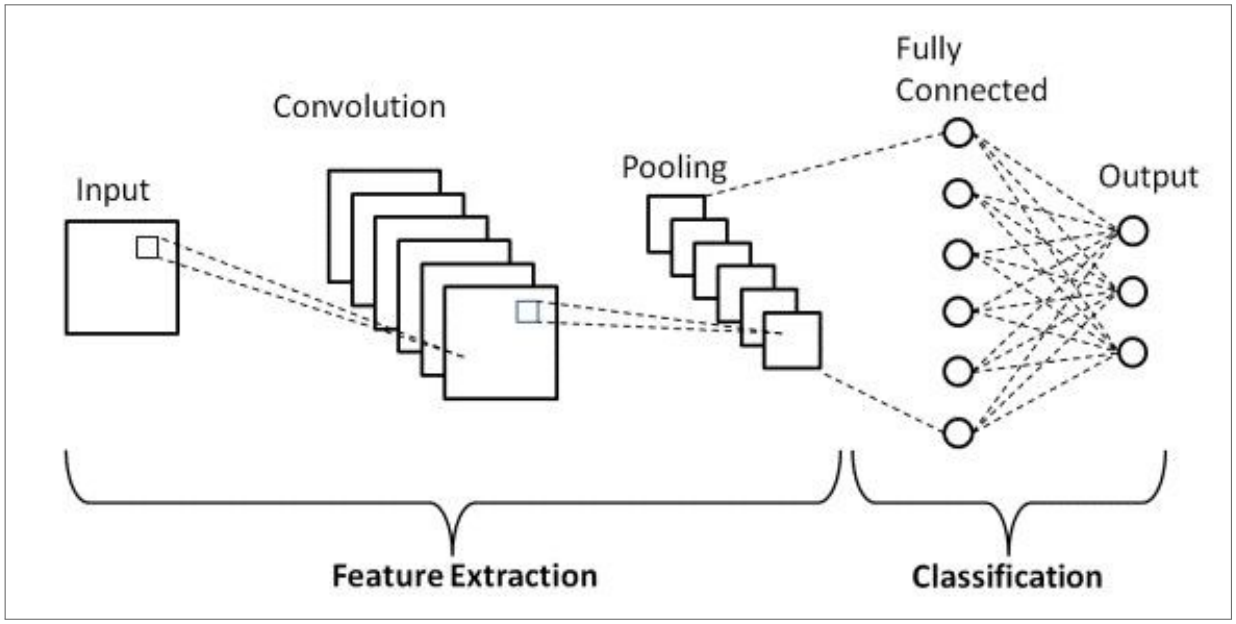
- *Camada Convolucional*: É a camada inicial de uma CNN e responsável pela extração de características locais da entrada. Essa camada utiliza filtros (ou kernels) para realizar operações de convolução sobre a imagem, capturando elementos básicos, como bordas, texturas e cores. A operação convolucional reduz a complexidade do modelo ao focar em regiões específicas da entrada, mantendo as informações espaciais relevantes.

- *Camada de Pooling*: Essa camada é projetada para reduzir a dimensionalidade dos dados e consolidar características importantes, diminuindo o número de parâmetros do modelo e mitigando o risco de *overfitting*.Os tipos mais comuns de *pooling* incluem max *pooling*, que seleciona o valor máximo em uma região, e average *pooling*, que calcula a média dos valores em uma região.

- *Camada Totalmente Conectada (Fully Connected — FC)*: É a camada final da CNN, onde todos os neurônios estão conectados entre si. Essa camada combina as características aprendidas nas etapas anteriores para realizar a classificação ou predição. O objetivo é integrar as informações extraídas em um vetor de saída correspondente às categorias ou valores previstos.

À medida que os dados percorrem essas camadas, a complexidade da CNN aumenta progressivamente. As camadas iniciais identificam características simples, como bordas ou padrões de cores. Em camadas intermediárias e finais, os padrões básicos são combinados para identificar formas maiores e detalhes mais específicos do objeto. Por fim, na camada totalmente conectada, a CNN reconhece o objeto completo, atribuindo-o a uma classe ou valor numérico específico. Esse design hierárquico permite que as CNNs sejam extremamente eficazes em tarefas como reconhecimento facial, análise de imagens

Figura 7 – Arquitetura de uma CNN básica



Fonte: extraída do artigo SENA; ROCHA (2021)

médicas e identificação de objetos em vídeos, consolidando seu papel essencial na visão computacional (KRIZHEVSKY; SUTSKEVER; HINTON, 2012).

2.7 COMITÊS MULTIMODAIS

Os comitês (*Ensemble*) multimodais representam uma classe de modelos de aprendizado profundo capazes de integrar e processar diferentes tipos de dados, ou “modalidades”, simultaneamente. Essas modalidades podem incluir imagens e dados tabulares. A capacidade de lidar com entradas de diferentes formatos torna essas redes particularmente úteis em problemas complexos, onde a informação relevante é frequentemente heterogênea (BALTRUŠAITIS; AHUJA; MORENCY, 2018).

O objetivo das redes multimodais é explorar complementaridades entre diferentes tipos de dados para melhorar a precisão e a robustez das predições. O termo para a interação de duas ou mais redes é "concatenação". Por exemplo, no diagnóstico médico, imagens de exames (como imagens histopatológicas) podem ser concatenadas com dados genômicos para fornecer uma análise mais abrangente. No caso dos *Ensemble* essa concatenação é feita por votos de modelos especialistas. O leitor pode imaginar um comitê de médicos especialistas com diferentes experiências e/ou diferentes focos em suas abordagens. Daí ao final do comitê cada especialista vota no diagnóstico em cada caso e declara a sua confiança no seu voto, como uma probabilidade. Assim um especialista em histopatologia pode votar que um dado caso é diagnóstico A e declarar 70% de confiança, outro especialista também em histopatologia pode votar no mesmo caso pelo diagnóstico B porém declarar apenas 60% de confiança; outro especialista, esse em genética, ao examinar um painel de genes pode votar no diagnóstico A e declarar 70% de confiança e finalmente outro especialista em genética pode votar no diagnóstico B e declarar 95% de confiança. No exemplo simplificado pode-se notar que há duas maneiras de computar os votos, levando em conta a confiança de cada especialista, chamado votação suave (*soft voting*, que nesse caso o resultado seria diagnóstico B ou sem levar em conta a confiança. chamado de voto duro (*hard voting*.

Outra opção digna de nota é a rede multimodal. Ela difere de *Ensemble* por seu método mais complexo de concatenação que permite diversos vetores de informação serem levados em conta antes da camada final. A arquitetura de uma rede multimodal é geralmente composta por três componentes principais:

- Encoders *específicos de modalidade*: cada tipo de dado é processado inicialmente por um modelo especializado, como redes neurais convolucionais (CNNs) para ima-

gens, redes neurais recorrentes (RNNs) ou modelos baseados em *transformers* para texto, e redes *feedforward* para dados tabulares. Esses *encoders* extraem características específicas de cada modalidade.

- *Fusão multimodal*: após a extração das características de cada modalidade, os vetores resultantes são combinados em um espaço comum. A fusão pode ser realizada de várias formas, como concatenação, adição ou mecanismos mais sofisticados, como atenção multimodal, que atribui pesos diferentes às modalidades com base em sua relevância para a tarefa.

- *Camadas de decisão*: A etapa final utiliza as características combinadas para realizar a tarefa desejada, como classificação ou regressão. Essa etapa pode ser composta por camadas totalmente conectadas (*fully connected layers*) ou outras arquiteturas específicas, dependendo da natureza do problema.

O treinamento de redes multimodais apresenta desafios únicos. Um dos mais significativos é o alinhamento entre modalidades, já que diferentes tipos de dados podem ter escalas, dimensões e distribuições distintas. Além disso, a ausência de informações completas em todas as modalidades para alguns exemplos (dados faltantes) exige abordagens robustas, como a utilização de técnicas de imputação ou redes específicas para lidar com entradas incompletas. Assim, o Ensemble Multimodal é uma abordagem mais simples de implementar e com mais explicabilidade final dos modelos já que os votos dos modelos especilista são conhecidos e podem ser mais facilmente avaliados.

As vantagens principais dos Ensembles e redes multimodais é a capacidade de superar limitações de dados individuais. Por exemplo, em tarefas onde informações textuais podem ser ambíguas, as imagens associadas podem fornecer detalhes contextuais essenciais. Da mesma forma, modalidades redundantes podem atuar como uma verificação cruzada, aumentando a confiabilidade do sistema.

2.8 CLASSIFICAÇÃO MOLECULAR DO ADENOCARCINOMA GÁSTRICO

A classificação molecular do câncer gástrico representa um avanço significativo na compreensão da heterogeneidade dessas neoplasias, permitindo uma estratificação mais precisa dos pacientes e, com isso, abrir caminhos para futuros desenvolvimentos de mais

precisos prognósticos e terapias. Em outras palavras, trata-se de um sistema que vai além das características histomorfológicas, incorpora perfis genômicos e epigenômicos para identificar subtipos distintos, o que é particularmente útil em pesquisas de genética e biologia molecular. Identificada no estudo do The Cancer Genome Atlas (TCGA) de 2014 (Cancer Genome Atlas Research Network, 2014).

A classificação molecular divide o adenocarcinoma gástrico com base nas características subjacentes dos processos moleculares em atuação na gênese do tumor. São quatro os subtipos principais identificados: positivo para Epstein-Barr virus (EBV), instável em microssatélites (MSI), genomicamente estável (GS) e com instabilidade cromossômica (CIN) (Cancer Genome Atlas Research Network, 2014).

Essa classificação surgiu da análise computacional não-supervisionada de dados multiômicos, incluindo sequenciamento de exoma, análise de cópias somáticas, metilação de DNA, expressão de mRNA, miRNA e proteínas, aplicada a 295 amostras de adenocarcinoma gástrico primário no estudo TCGA-STAD (*stomach adenocarcinoma*). Em termos práticos, ela facilita a identificação de alvos terapêuticos específicos, otimizando o desenvolvimento de modelos computacionais em bioinformática para predição de resposta a tratamentos.

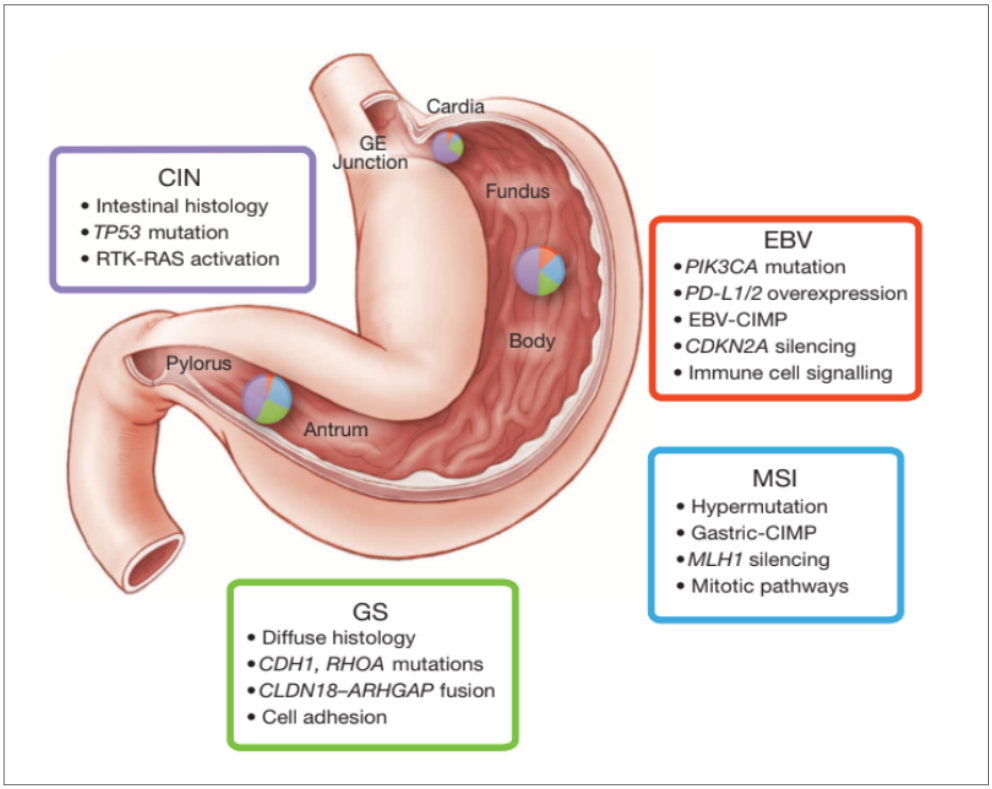
2.8.1 Subtipo EBV

Os tumores positivos para EBV, representando cerca de 9% dos casos, forsm caracterizados por alta carga viral do Epstein-Barr virus, mutações recorrentes em PIK3CA (80% dos casos), hipermetilação extrema de DNA (EBV-CIMP) e amplificações em JAK2, PD-L1 e PD-L2. Em outras palavras, esses tumores exibem um perfil imunogênico pronunciado, com silenciamento epigenético de genes como CDKN2A, mas sem hipermetilação de MLH1, o que os diferencia dos subtipos MSI e sugere potencial para imunoterapias, como inibidores de checkpoint imune (Cancer Genome Atlas Research Network, 2014). Clinicamente, eles são mais comuns em homens e localizados no fundo ou corpo gástrico.

2.8.2 Subtipo MSI (Instabilidade Microssatélite)

Compreendendo 22% dos tumores, o subtipo MSI foi distinguido por altas taxas de mutação (hipermutados), silenciamento de MLH1 via hipermetilação e mutações em genes

Figura 8 – Principais características dos subtipos de câncer gástrico



Fonte: extraída do artigo Cancer Genome Atlas Research Network (2014)

como KRAS, ARID1A e PIK3CA. Em outras palavras, essa instabilidade resulta em um acúmulo de erros no DNA, frequentemente associado a melhor prognóstico e resposta a imunoterapias devido ao alto carga mutacional, que pode ser quantificada em ferramentas bioinformáticas para estimar a neoantigenicidade (FUKAYAMA; RUGGE; WASHINGTON, 2019). Esses tumores são diagnosticados em idades mais avançadas e em mulheres.

2.8.3 Subtipo GS (Genomicamente Estável)

Representando 20% dos casos, os tumores GS foram associados ao tipo histológico difuso (73%), descritos como asociados a mutações em RHOA ou fusões envolvendo proteínas ativadoras de GTPases da família RHO, além de baixa aneuploidia. Em outras palavras, essa estabilidade genômica reflete um número limitado de mutações, porém afetando vias fundamentais de agregação celular (como a da caderina). São diagnosticados em paciente com menos idade.

2.8.4 Subtipo CIN (Instabilidade Cromossômica)

O subtipo mais comum (50%), CIN foi caracterizado por aberrações cromossômicas, aneuploidia marcada, amplificações focais em receptores tirosina quinases (como EGFR, HER2) e mutações em TP53. Em outras palavras, essa instabilidade leva a ganhos e perdas cromossômicas extensas, associadas a tumores intestinais localizados na junção gastroesofágica, com implicações para terapias direcionadas como inibidores de HER2 (FUKAYAMA; RUGGE; WASHINGTON, 2019).

Tabela 6 – Subtipos Moleculares do Câncer Gástrico (TCGA 2014).

Subtipo	Frequência (%)	Características Moleculares Principais	Localização Preferencial	Idade Média
EBV	9	Hipermetilação, mutações PIK3CA, amplificações PD-L1/L2.	Fundo/Corpo	65 anos
MSI	22	Hipermutação, silenciamento MLH1, mutações KRAS.	Corpo/Antrum	72 anos
GS	20	Mutuções RHOA, tipo difuso, baixa aneuploidia.	Difuso	59 anos
CIN	50	Aneuploidia, amplificações RTK, mutações TP53.	Junção GE	68 anos

Fonte: The Cancer Genome Atlas Research Network. (2014). Comprehensive molecular characterization of gastric adenocarcinoma. Nature, 513, 202–209).

2.8.5 Reconhecimento na 5ª edição da OMS

A classificação da OMS 2019 (FUKAYAMA; RUGGE; WASHINGTON, 2019) apresenta um esforço inicial para incorporar os achados moleculares do estudo TCGA 2014 (Cancer Genome Atlas Research Network, 2014). A classificação molecular é citada no tópico prognóstico com as seguintes palavras: "Perfil molecular: Os perfis moleculares recentemente identificados não estão apenas envolvidos na carcinogênese gástrica, mas também podem ajudar a identificar biomarcadores clinicamente relevantes e novos alvos terapêuticos potenciais no futuro". Esse comentário é seguido da tabela **??** e **??** para ilustrar as associações entre os tipos histopatológicos e os subtipos moleculares.

Tabela 8 – Características dos subtipos moleculares de carcinoma gástrico propostos pelo TCGA.

Característica	**EBV-positivo**	**MSI**	**Genomicamente estável**	**Cromossomicamente instável**
Frequência relativa	9%	22%	20%	50%
Histologia representativa	Carcinoma gástrico com estroma linfoide	Nenhum	Tipo difuso*	Tipo intestinal*
Ilha (Metilação)	CIMP	CIMP	Raro	Raro
MSI-alto (Metilação)	Ausente	Todos	Ausente	Ausente
CDKN2A (Metilação)	Todos	Frequente	Raro	Raro
MLH1 (Metilação)	Ausente	Frequente	Raro	Raro
Aberrações no número de cópias	Raro	Raro	Raro	Frequente
Mutações/alterações genômicas	Raro	Frequente	Raro	Frequente
TP53	Raro	Presente	Raro	Frequente
CDH1	Ausente	Raro	Presente	Raro
PIK3CA	Frequente	Presente	Raro	Raro
RHOA	Raro	Raro	Presente	Raro
Fusão CLDN18-ARHGAP	Ausente	Raro	Presente	Raro
ARID1A	Frequente	Presente	Raro	Raro
Amplificação de RTK	Raro	Raro	Raro	Frequente
Mutação de RTK	Raro	Frequente	Raro	Raro
Amplificação de CD274 (PD-L1) e PDCD1LG2 (PD-L2)	Frequente	Raro	Raro	Raro

Fonte: The Cancer Genome Atlas Research Network. (2014). Comprehensive molecular characterization of gastric adenocarcinoma. Nature, 513, 202–209).

No entanto, embora essa iniciativa deva ser reconhecida como um passo positivo em direção à integração de dados moleculares na rotina médica, a maneira como foi implementada peca por simplismo, pois as correlações propostas são frequentemente genéricas e não capturam a complexidade da heterogeneidade tumoral observada no TCGA.

Por exemplo, o subtipo pouco coeso (incluindo células em anel de sinete) na OMS é alinhado ao subtipo GS, enquanto o carcinoma, contudo, essa associação é criticamente limitada pela falta de marcadores moleculares específicos validados que permitam uma predição robusta no nível individual do paciente. O estudo TCGA de 2014 (Figura 1) ilustra

claramente a associação entre subtipos moleculares e tipos histopatológicos de Lauren, revelando uma heterogeneidade evidente: embora o tipo difuso seja o mais frequente no subtipo GS (73% dos casos), nem todo tumor difuso é GS, e nem todo GS é difuso. Essa sobreposição incompleta se aplica igualmente a outras características mencionadas, como a localização anatômica ou o perfil mutacional, destacando que correlações simplistas podem levar a classificações imprecisas e subestimar a variabilidade tumoral (Cancer Genome Atlas Research Network, 2014). Assim, há uma necessidade urgente de sistemas de predição mais robustos, incorporando algoritmos de aprendizado de máquina para análise integrada de múltiplas regras, a fim de refinar essas associações e melhorar a aplicabilidade clínica.

2.9 CARCINOGÊNESE DOS SUBTIPOS MOLECULARES DO ADENOCARCINOMA GÁSTRICO

A carcinogênese gástrica é um processo multifatorial de transformação de células epiteliais normais em neoplasias invasivas, impulsionado por uma cascata de alterações ambientais, epigenéticas e genéticas que culminam em proliferação descontrolada, invasão tecidual e metástases. Em outras palavras, trata-se de uma progressão gradual onde fatores como infecção por Helicobacter pylori, exposição a carcinógenos dietéticos e predisposições genéticas interagem.

Os subtipos moleculares permitem uma compreensão mais individualizada da carcinogênese. Posto que dois pacientes com diagnóstico de tumores tubulares que pertençam a subtipos diferentes passaram por processos de carcinogênese subjacente muito diferentes.

2.9.1 Carcinogênese no Subtipo EBV-Positivo

A carcinogênese nos tumores EBV-positivos inicia-se com a infecção crônica pelo Epstein-Barr virus, que integra seu genoma nas células epiteliais gástricas, desencadeando hipermetilação extrema de promotores de DNA (EBV-CIMP) e silenciamento de genes supressores como CDKN2A (Cancer Genome Atlas Research Network, 2014). Em outras palavras, esse processo viral promove uma reprogramação epigenética que ativa vias oncogênicas, notadamente PI3K/AKT via mutações recorrentes em PIK3CA (80% dos casos), e amplifica genes imunomoduladores como PD-L1 e PD-L2, sugerindo um papel central da imunoeva-

são.

2.9.2 Carcinogênese no Subtipo MSI (Instável em Microssatélites)

Nos tumores MSI, a carcinogênese decorre de defeitos no sistema de reparo de DNA por desemparelhamento (MMR), frequentemente devido ao silenciamento epigenético de MLH1, levando a uma alta taxa de mutações somáticas (hipermutação) (Cancer Genome Atlas Research Network, 2014). Em outras palavras, essa instabilidade acumula erros genéticos em genes como KRAS, ARID1A e PTEN, gerando um ambiente imunogênico rico em neoantígenos, o que favorece a resposta a imunoterapias.

2.9.3 Carcinogênese no Subtipo GS (Genomicamente Estável)

A carcinogênese no subtipo GS é impulsionada por mutações em genes de adesão celular, como RHOA, ou fusões envolvendo GTPases, resultando em uma morfologia pouco coesa (Cancer Genome Atlas Research Network, 2014). Em outras palavras, essas alterações comprometem a coesão celular, facilitando a invasão tecidual e metástase. Apesar da estabilidade genômica, a heterogeneidade observada no TCGA sugere que outros fatores epigenéticos ou microambientais ainda precisam ser elucidados, demandando estudos integrados para capturar a complexidade desse subtipo.

2.9.4 Carcinogênese no Subtipo CIN (Instabilidade Cromossômica)

Nos tumores CIN, a carcinogênese é marcada por aberrações cromossômica, com ganhos e perdas cromossômicas, frequentemente associadas a mutações em TP53 (FUKAYAMA; RUGGE; WASHINGTON, 2019). Em outras palavras, essa desregulação mitótica acelera a progressão tumoral, tornando esses tumores candidatos a terapias direcionadas como inibidores de HER2. A prevalência desse subtipo (50% dos casos) reforça a necessidade de ferramentas bioinformáticas avançadas para monitorar a evolução clonal, especialmente em contextos de resistência terapêutica.

Nos cânceres hematológicos, como leucemias e linfomas, são frequentemente observadas alterações cromossômicas específicas, como deleções e translocações, que estão associadas a subtipos particulares da doença. Por exemplo, a translocação conhecida

como cromossomo Filadélfia, é característica da leucemia mieloide crônica. Essas alterações servem como marcadores diagnósticos e podem orientar o tratamento.

Em contraste, os tumores sólidos, como o adenocarcinoma gástrico, geralmente exibem uma ampla variedade de alterações cromossômicas (heterogeneidade), incluindo aneuploidias e rearranjos complexos, sem padrões específicos que possam ser diretamente associados a subtipos tumorais. Essa heterogeneidade torna o diagnóstico baseado em alterações cromossômicas específicas menos preciso em tumores sólidos. Além disso, a INC em tumores sólidos pode resultar em uma diversidade de alterações genômicas que contribuem para a progressão tumoral, sem um padrão uniforme. Portanto, enquanto nos cânceres hematológicos as alterações cromossômicas específicas desempenham um papel crucial no diagnóstico e na classificação, nos tumores sólidos a diversidade de alterações genômicas torna essa abordagem menos eficaz (RAJAGOPALAN et al., 2003; THOMPSON; BAKHOUM; COMPTON, 2010).

2.10 PAINÉIS IMUNO-HISTOQUÍMICOS E SONDAS GENÔMICAS NA ESTRATÉGIA DIAGNÓSTICA DOS SUBTIPOS MOLECULARES

Um painel apenas utilizando imuno-histoquímica (IHQ) e *hibridização in situ* foi proposto para chegar aos subtipos moleculares na prática médica. Enquanto o TCGA utilizou 6 plataformas moleculares incluindo sequenciamento do exoma e do transcriptoma, os autores do painel propuseram uma maneira de classificar as amostras utilizando 10 recortes histológicos, sendo um para sonda para EBV de *hibridização in situ* (ISH) e 9 para diferentes anticorpos na IHQ (MLH1, PMS2, MSH2, MSH6, HER2, EGFR, MET, PTEN e P53) (KIM et al., 2016). Em uma tese de doutorado um painel bem mais enxuto com 7 lâminas sendo ISIH para EBV e IHQ (MLH1, MSH2, MSH6, PMS2, E-CADERINA e P53)(RAMOS, 2019).

A principal crítica a esses painéis é que os autores não levaram em conta a heterogeneidade das categorias propostas no TCGA, eles apenas realizaram seu painel, mas não o compararam com amostras nas quais foram realizados os sequenciamentos, partiram do conhecimento da literatura em antígenos presentes nas anormalidades moleculares inferidas em outros tipos de câncer e os utilizaram. É, portanto, de grande interesse verificar a correlação entre esses marcadores e a classificação realizada por multi-ômicas no banco de dados originais.

No painel proposto por Ramos, o marcador P53 é de fato o único direcionado especificamente para o subtipo CIN, o que representa uma simplificação significativa em relação ao painel de Kim. A escolha de P53 como marcador é baseada na sua frequente mutação no subtipo CIN, refletindo a instabilidade genômica característica deste subtipo. No entanto, utilizar apenas o P53 para identificar tumores CIN pode ser insuficiente para capturar toda a complexidade molecular desse subtipo. A instabilidade cromossômica é um fenômeno que envolve uma ampla gama de alterações genéticas e epigenéticas, e focar unicamente no P53 pode deixar de lado outras vias críticas que também contribuem para o fenótipo CIN, como as amplificações de HER2, EGFR e MET observadas em outras abordagens mais abrangentes, como a do TCGA.

A limitação de usar apenas P53 como marcador para CIN está na sub-representação do espectro completo das alterações associadas ao subtipo, o que pode levar a uma subdiagnóstico ou até à falta de identificação de certos casos de CIN. Embora a simplificação do painel seja uma vantagem em termos de custo e acessibilidade, a falta de correlação direta com os achados multiômicos detalhados do TCGA compromete a precisão diagnóstica. Assim, enquanto o painel de Ramos oferece uma abordagem mais enxuta e prática, ele pode não ser suficiente para capturar a heterogeneidade completa do subtipo CIN, sugerindo a necessidade de integrar mais marcadores ou validar a eficácia de P53 em conjunto com outras alterações moleculares em amostras já analisadas pelo TCGA.

Uma maneira relativamente simples de checar se as intuições dos autores dos painéis estão corretas é verificar por métodos de bioinformática se, nas amostras originais do TCGA cujos dados de mutações somáticas foram disponibilizados, a correlação com os antígenos propostos. Seguindo esse raciocínio também verificar se há casos que não apresentam esses antígenos e que outras maneiras haveria de identificá-los por IHQ, ou, por outro lado, se todos os marcadores são de fato propostos. Embora, por exemplo, no subtipo instabilidade microssatélite (MSI), que foi verificada na amostra do TCGA pelo grande aumento de mutações e por hipermetilação, com a mutação nos genes de reparo, verificada no painel por MLH1, MSH2, MSH6 e PMS2, no caso do câncer gástrico essa associação precisa ainda ser provada no caso específico por ser concebível que outros mecanismos estejam atuando. Essa verificação da significância entre os marcadores imuno-histoquímicos e a mutação somática necessária à expressão do referido antígeno é importante para estabelecer a sensibilidade e especificidade dos painéis.

2.11 GENES ABORDADOS NA TESE

Na presente sessão são citadas as associações conhecidas dos genes tratados na presente tese. Tanto os que compõe os painéis imuno-histoquímicos propostos na literatura como os identificados aqui por aprendizado de máquina e apresentados nos Capítulos. A tabela 9 lista os genes citados em ordem alfabética servindo para consulta sobre associações conhecidas com carcinogênese e disponibilidade de Anticorpos Ac para diagnóstico imuno-histoquímico de variantes genéticas no gene em questão.

Tabela 9 – Resumo funcional dos genes abordados e sua relevância em carcinogênese.

Gene	**Tipo Funcional Principal**	**Associação com Carcinogênese**	**Ac**
ARID1A	Supressor Tumoral	Perda de função em câncer gástrico, de ovário e endométrio. Promove instabilidade epigenética e progressão tumoral.	Sim
ATM	Reparo de DNA	Mutações aumentam a suscetibilidade a tumores (mama, gástrico). Sua perda favorece instabilidade genômica e metástase.	Sim
BHLHB9	Supressor Tumoral (Potencial)	Silenciamento associado a neoplasias colorretais, sugerindo que sua perda contribui para a progressão de tumores gastrointestinais.	Raros
BOC	Sinalização Celular	Expressão aberrante associada a meduloblastomas e sarcomas. Pode ativar a via oncogênica Hedgehog.	Raros
CD14	Resposta Imune	Pode promover imunoescapismo ou inflamação crônica que facilita a carcinogênese.	Sim

Continua na próxima página

Tabela 9 – Resumo funcional dos genes abordados e sua relevância em carcinogênese. (continuação)

Gene	Tipo Funcional Principal	Associação com Carcinogênese	Ac
CDH1	Supressor Tumoral / Adesão	Perda de função facilita invasão e metástase. Mutações germinativas causam câncer gástrico difuso hereditário.	Sim
CHD1	Remodelador de Cromatina	Deleções são frequentes no câncer de próstata, associadas a agressividade tumoral e instabilidade genômica.	Sim
DOCK3	Motilidade Celular	Expressão aberrante em gliomas pode aumentar a invasão e motilidade tumoral.	Sim
EGFR	Oncogene	Mutações de ganho de função e amplificação impulsionam o crescimento de tumores de pulmão, glioblastoma e outros.	Sim
FAS	Apoptose	Via frequentemente desativada em tumores sólidos, contribuindo para o escape da morte celular programada.	Sim
GGNBP2	Supressor Tumoral (Potencial)	Perda de expressão observada em tumores testiculares, sugerindo um papel na supressão tumoral.	Raros
GLIS2	Fator de Transcrição	Perda de função pode favorecer a progressão de tumores renais e outros carcinomas.	Raros
HERC2	E3 Ubiquitina Ligase	Mutações germinativas associadas à predisposição a tumores, possivelmente pela modulação da função do p53.	Sim

Continua na próxima página

Tabela 9 – Resumo funcional dos genes abordados e sua relevância em carcinogênese. (continuação)

Gene	Tipo Funcional Principal	Associação com Carcinogênese	Ac
HER2	Oncogene	Amplificação e superexpressão em câncer de mama e gástrico conferem crescimento agressivo e são alvos terapêuticos.	Sim
KDM2B	Oncogene (Epigenética)	Superexpressão em leucemias e carcinomas, reprime genes supressores de tumor e promove proliferação.	Sim
KMT2D	Supressor Tumoral (Epigenética)	Frequentemente mutado em linfomas e carcinomas; a perda de função facilita a transformação maligna.	Sim
MEF2C	Fator de Transcrição	Reordenamentos em leucemias promovem autorenovação. Em tumores sólidos, pode contribuir para angiogênese.	Sim
MET	Oncogene	Mutações ativadoras e amplificações são drivers em carcinomas gástricos, renais e de pulmão, estimulando invasão e metástase.	Sim
MLH1	Reparo de DNA (MMR)	Mutações causam a Síndrome de Lynch. A perda de função leva à instabilidade de microssatélites (MSI).	Sim
MSH2	Reparo de DNA (MMR)	Mutações causam a Síndrome de Lynch, levando à instabilidade de microssatélites e acelerando a evolução tumoral.	Sim
MSH6	Reparo de DNA (MMR)	Mutações associadas à Síndrome de Lynch e a cânceres de próstata e endométrio.	Sim

Continua na próxima página

Tabela 9 – Resumo funcional dos genes abordados e sua relevância em carcinogênese. (continuação)

Gene	Tipo Funcional Principal	Associação com Carcinogênese	Ac
MUC16	Biomarcador / Mucina	Superexpresso em câncer de ovário, onde pode promover disseminação peritoneal e evasão imune. Conhecido como CA125.	Sim
MUC6	Mucina Protetora	Perda de expressão está associada à progressão para metaplasia intestinal, precursora do câncer gástrico.	Sim
PIK3CA	Oncogene	Mutações de ganho de função são drivers em carcinomas de mama, cólon e endométrio, ativando vias de crescimento.	Sim
PMS2	Reparo de DNA (MMR)	Mutações causam uma forma da Síndrome de Lynch. Sua perda contribui para a instabilidade de microssatélites.	Sim
PRCC	Fusão Gênica Oncogênica	Translocações com o gene TFE3 produzem proteínas de fusão oncogênicas no carcinoma de células renais.	Raros
PTEN	Supressor Tumoral	Perda de função é comum em glioblastomas e câncer de próstata, resultando em ativação da via pró-sobrevivência PI3K/AKT.	Sim
PTPN14	Supressor Tumoral (Via Hippo)	Perda de função promove proliferação tumoral ao desinibir o oncogene YAP.	Sim

Continua na próxima página

Tabela 9 – Resumo funcional dos genes abordados e sua relevância em carcinogênese. (continuação)

Gene	Tipo Funcional Principal	Associação com Carcinogênese	Ac
SDR9C7	Biomarcador (Metabolismo)	Expressão aumentada em carcinomas de cabeça e pescoço, podendo atuar como biomarcador de prognóstico.	Raros
SEC31A	Fusão Gênica Oncogênica	Fusões com o gene ALK geram um receptor constitutivamente ativo com potencial oncogênico.	Não
SYNE1	Estabilidade Genômica	Deleções podem comprometer a integridade nuclear, favorecendo a instabilidade do genoma e a progressão de carcinomas.	Sim
TP53	Supressor Tumoral	"Guardião do Genoma". Mutado em >50% dos cânceres, abolindo a supressão tumoral. Mutações germinativas causam a Síndrome de Li-Fraumeni.	Sim
ZBTB41	Fator de Transcrição	Mutações identificadas em câncer gastrointestinal, sugerindo contribuição para a desregulação transcricional.	Raros

MMR: Mismatch Repair (Reparo de Erros de Pareamento). A disponibilidade de anticorpos foi baseada nas informações do texto de origem.

2.12 REDES NEURAIS CONVOLUCIONAIS NO DIAGNÓSTICO HISTOPATOLÓGICO DO CÂNCER GÁSTRICO

Outra abordagem de aprimoramento do diagnóstico do câncer em bioinformática é a Inteligência Artificial IA aplicada a problemas de visão computacional VC como classificação e reconhecimento de objetos em imagens histopatológicas digitalizadas.

A própria digitalização de imagens de lâminas inteiras (*Whole Slide Imaging* - WSI) é

recente. O primeiro sistema de patologia digital aprovado para diagnóstico primário foi o *Philips IntelliSite Pathology Solution*, aprovado pela FDA (*Food and Drug Administration* dos EUA) em 2017 e publicado em 2018. Isso foi um marco histórico, permitindo o uso de WSI para diagnósticos primários em patologia cirúrgica, em vez de microscópios tradicionais (EVANS et al., 2018). Embora a fotografia digital de campos específicos seja bem mais antiga, contemporânea da própria fotografia digital, já que o microscópio óptico é essencialmente lentes de aumento, para produzir WSI confiáveis foi necessário o avanço do poder computacional para costurar milhares de imagens em uma única imagem digital da lâmina inteira. Esse é um avanço disruptivo por proporcionar o aumento do diálogo entre especialistas que podem agora ver a mesma WSI estando em locais diferentes.

Assim, em 2020, quando da escrita do projeto da presente tese e do projeto de inovação que a acompanha, a patologia digital havia sido publicizada há somente dois anos. Durante o período do presente trabalho muito aconteceu no campo. Com o desenvolvimento de modelos de redes neurais treinadas por aprendizado profundo (*deep learning*) capazes de auxiliar no diagnóstico patológico. As redes neurais têm o potencial de impulsionar o desenvolvimento da patologia enquanto área do conhecimento por promover padrões quantitativos de análise e diagnóstico.

Foram publicados trabalhos demonstrando a eficiência de IA's no reconhecimento dos padrões histomorfológicos do câncer gástrico (IIZUKA et al., 2020; HUANG, 2021; JANG; SONG; LEE, 2021; KANAVATI et al., 2021) utilizando aprendizado supervisionado a partir de imagens rotuladas por patologistas. Jang et al. (2021) demonstraram que uma CNN Inception-v3 foi capaz de distinguir adenocarcinomas gástricos diferenciados vs. indiferenciados e mucinosos vs. não-mucinosos, alcançando AUCs de 0,932 e 0,979, respectivamente. De forma semelhante, Kanavati & Tsuneki (2021) avaliaram o desempenho de CNNs na classificação do adenocarcinoma difuso (tipo Lauren), utilizando mais de 2.900 biópsias de múltiplos hospitais japoneses. Os modelos atingiram AUCs próximos de 0,95–0,99 em diferentes coortes, mostrando que a IA pode capturar os padrões histológicos reconhecidos por patologistas.

Esse avanço foi possibilitado pelo aumento do poder computacional devido à computação em nuvem. Consistiu no desenvolvimento das técnicas de aprendizado profundo, em especial das redes neurais convolucionais (*Convulotional Neuro Networks* — CNN). Do ponto de vista computacional, as imagens são matrizes matemáticas que podem ser reconhecidas por modelos profundos treinados por métodos de aprendizado supervisionados.

Para uma introdução à visão computacional vide 2.6.4

CNNs podem identificar câncer em imagens histológicas ao serem treinadas em conjunto de dados (*datasets*) rotulados por especialistas utilizando aprendizado supervisionado (IIZUKA et al., 2020; HUANG, 2021). Esses *datasets* são conjuntos de imagens selecionadas com tamanho padronizado acompanhadas de rótulos, ou seja, de informações indicando, por exemplo, se é uma imagem de câncer ou não, ou, por exemplo, que tipo de câncer, etc. Quando são utilizados *datasets* rotulados por patologistas para o treinamento supervisionado de CNN apresenta acurácia comparável ao de humanos no reconhecimento dos padrões específicos associados ao câncer para os quais foi treinado. As CNN's demonstram resultados com grande acurácia em uma quantidade crescente de problemas de visão computacional, permitindo análises quantitativas onde antes só era possível análises qualitativas. Além de aplicações práticas, as CNN tem ainda grande potencial como ferramenta de investigação científica.

2.13 SUPERVISÃO MOLECULAR DE REDES NEURAIS CONVOLUCIONAIS NO CÂNCER GÁSTRICO

Do ponto de vista da visão computacional, um grande desafio a ser superado é a limitação à rotulação humana para o desenvolvimento. Ou seja, o modelo desenvolvido estará limitado ao que os observadores humanos já identificaram e informaram na rotulação do *dataset* de treinamento e essa rotulação é feita por anotação humana. A principal estratégia na ciência da computação é a utilização de métodos não-supervisionados. Um bom exemplo desse tipo de abordagem em imagens histopatológicas é o trabalho de (LEE et al., 2022) que usou CNN para extrair atributos depois usou métodos de clustering não supervisionado e finalmente usou florestas aleatórias (*random forests*) para associar os agrupamentos (*clusters*) encontrados com informações clinicamente significativas. Assim, em diversas aplicações da visão computacional pesquisadores buscam estratégias de treinamento não-supervisionado para superar a limitação dos rótulos humanos e potencializar as CNN como ferramental de investigação científica.

No campo específico do reconhecimento de imagens histológicas foi proposta como solução disruptiva a ideia de supervisão molecular no treinamento de redes neurais (MONJO et al., 2022). Com o advento das ômicas e da visão computacional surgiu uma oportunidade de usar (*deep learning*) para descoberta de fenótipos a partir de dados moleculares

conhecidos. Durante a história do desenvolvimento da genética a ordem dos fatores foi a partir de um fenótipo conhecido pesquisar quais acontecimentos moleculares estavam associados ao fenótipo em questão. Esse foi o caminho desde os trabalhos de Mendel até as recentes investigações multi-ômicas para compreensão do câncer. O que impulsiona essa busca em uma direção diferente é o fato conhecido que o sequenciamento de nova geração trouxe abundância de dados com necessidade do desenvolvimento de técnicas de interpretação. Para uma introdução ao genoma e ao sequenciamento vide 2.5.2.

A supervisão molecular é, assim, a estratégia de pesquisa de oferecer dados moleculares como rótulos moleculares em *datasets* de imagens histopatológicas, sem rótulos humanos, e a partir de treinamento supervisionado clássico descobrir atributos profundos (*deep features*, que devido a sua relação direta com dados genéticos podem ser considerado fenótipos profundos (*deep phenotype*. (YURKOVICH et al., 2020) Fenótipo profundo refere à caracterização detalhada e complexa de fenótipos (características identificáveis) de uma condição, obtida a partir de análises avançadas. Nesse contexto, utiliza-se aprendizado profundo (deep learning) para extrair padrões complexos de dados, especialmente em imagens histopatológicas.

Na supervisão molecular, portanto, embora seja o treinamento supervisionado, não há o viés da supervisão humana. Muitas vezes os rótulos moleculares são resultantes de aprendizado não-supervisionado, como é o caso na presente tese. (Cancer Genome Atlas Research Network, 2014) Monjo e colaboradores (MONJO et al., 2022) demonstraram um modelo capaz de identificar os diferentes tipos celulares em amostras de câncer de mama utilizando dados de transcriptômica espacial para treinar a CNN. O Slide-seq (RODRIGUES; SILVA; FONSECA, 2019) é uma técnica que marca com *barcodes* de localização conhecida no eixo x e y e depois sequência, permitindo um resultado de transcriptômica espacial. O que é muito útil para compreender as diferenças entre as linhagens de um mesmo tecido já que atinge um transcriptoma de single cell. Trata-se de uma técnica nova, de alta tecnologia e alto custo mas que aponta para o potencial futuro dessa abordagem de treinamento de redes neurais.

No tema específico da aplicação da supervisão molecular na classificação de imagens histopatológicas nos subtipos moleculares do CG dois trabalhos se destacam (WANG et al., 2022; FLINNER et al., 2022). Wang et al. (2022) introduziram um modelo para predição dos quatro subtipos do TCGA (CIN, MSI, EBV, GS). Já Flinner et al. (2022) aplicaram deep learning nos quatro subtipos do TCGA e compararam a testes moleculares independentes

e imuno-histoquímica. Ambos com resultados signitifcativos.

Wang et al. (WANG et al., 2022) propuseram o *Deep Ensemble for Molecular Subtyping*, uma abordagem de ensemble baseada em *deep learning* que integra múltiplos classificadores de uma mesma rede (EfficientNEt-v2) para prever os subtipos moleculares diretamente de imagens histopatológicas coradas com hematoxilina-eosina (H&E) do TCGA. O modelo foi treinado em um conjunto de dados derivado de 295 amostras de adenocarcinoma gástrico do TCGA, divididas em *tiles* (pequenos patches de imagem) para análise granular. No nível de tile, alcançou áreas sob a curva ROC (AUROC) de 0.785 para o subtipo CIN (instabilidade cromossômica), 0.668 para MSI (instabilidade de microssatélites), 0.762 para EBV (positivo para vírus Epstein-Barr) e 0.811 para GS (genomicamente estável). No nível de paciente, onde as predições de tiles são agregadas para uma classificação final, os valores de AUROC foram ainda mais elevados: 0.897 para CIN, 0.764 para MSI, 0.890 para EBV e 0.898 para GS. Esses resultados demonstram uma performance robusta, especialmente para os subtipos CIN, EBV e GS, com o ensemble superando modelos individuais e reduzindo o *overfitting* e melhorando a generalização.

Por sua vez, Flinner et al. (FLINNER et al., 2022) desenvolveram uma rede neural convolucional (CNN) de ensemble com técnica de bagging para prever os quatro subtipos moleculares do TCGA diretamente de slides H&E, comparando o desempenho com testes imuno-histoquímicos (IHC) padrão (para MLH1, PMS2, HER2 e EBER-ISH) e análises moleculares independentes (como sequenciamento para MSI e detecção de EBV). O estudo utilizou uma coorte de 438 amostras de GC, com validação externa em conjuntos independentes. Os resultados mostraram que o modelo de deep learning superou o IHC na acurácia geral de predição dos subtipos, com uma acurácia média de 85-90% para classificação binária e AUROC acima de 0.85 para subtipos como EBV e CIN em validação cruzada. Especificamente, para EBV, o modelo alcançou sensibilidade de 92% e precisão de 88%, identificando características como linfócitos infiltrantes e padrões glandulares; para MSI, AUROC de 0.82, destacando heterogeneidade intratumoral em 15-20% dos casos; para CIN, AUROC de 0.89, correlacionado com amplificações em HER2; e para GS, AUROC de 0.80, associado a tipos difusos. Em comparação ao IHC, que teve acurácia de cerca de 75-80% e falhas em casos heterogêneos, o deep learning identificou casos com predições mistas (*intra-tumoral heterogeneity*) em 10-15% das amostras, sugerindo que o método pode detectar variações não capturadas por marcadores tradicionais. Os autores enfatizam que o deep learning é superior em cenários de triagem, reduzindo a ne-

cessidade de testes moleculares caros e permitindo uma abordagem mais personalizada, embora recomendem validação adicional em coortes maiores para robustez clínica.

3 CAPÍTULO 1: G.SUBTVISION – SUBTIPAGEM MOLECULAR DO CÂNCER GÁSTRICO COM MÉTODOS DE ENSEMBLE DE REDES NEURAIS CONVOLUCIONAIS (CNNS)

RESUMO

A classificação molecular do adenocarcinoma gástrico em 4 subtipos: Instabilidade cromossômica (CIN), instabilidade microssatélite (MSI), vírus Epstein-Barr (EBV) e genômica estável (GS) depende de métodos onerosos e de acesso limitado. A classificação de imagens histopatológicas por redes neurais convolucionais (CNNs) é uma alternativa promissora. Este estudo propõe o G.Subtvision (gastro subtypes computational vision), ensemble multiarquitetura de CNNs (MobileNetV2, ShuffleNet e GoogLeNet) treinadas com rótulos moleculares para predição multiclasse do subtipo do adenocarcinoma gástrico. Metodologia: A partir de 263 casos, com 476 lâminas (TCGA-STAD). O treinamento ocorreu em diversas distribuições de treino e validação k-fold, k=10. Avaliação no nível de tile e paciente superou a reprodução controle de modelo previamente publicado em precisão por 4 pontos percentuais na média e 14 pontos no subtipo MSI. O G.Subtvison avança incrementalmente a subtipagem molecular do câncer gástrico por CNN. Código e material suplementar disponíveis.

3.1 INTRODUÇÃO

O consórcio The Cancer Genome Atlas (TCGA) estabeleceu, em 2014, uma classificação molecular em quatro grupos principais: tumores associados ao vírus Epstein-Barr (EBV), instabilidade de microssatélites (MSI), instabilidade cromossômica (CIN) e genômica estável (GS). Esta subtipagem tem impacto direto na estratificação prognóstica e na indicação terapêutica, uma vez que tumores EBV e MSI, por exemplo, demonstram melhor resposta à imunoterapia, enquanto tumores CIN e GS apresentam comportamento clínico distinto e resistência a determinados esquemas quimioterápicos (Sohn et al., 2017). Apesar de seu valor clínico, a subtipagem molecular depende de técnicas laboratoriais caras e pouco acessíveis, como hibridização in situ (ISH), imunohistoquímica (IHC) ampliada e sequenciamento genômico em larga escala. Estas abordagens não apenas elevam custos, como demandam infraestrutura tecnológica nem sempre disponível em países em desenvolvimento, além de prolongarem o tempo para decisão terapêutica. Este cenário motiva a busca por métodos complementares capazes de aproximar a classificação molecular da prática clínica diária (FLINNER et al., 2022; WANG et al., 2022).

Nos últimos anos, o avanço da patologia digital e do aprendizado profundo (deep learning) trouxe a possibilidade de extrair assinaturas moleculares latentes diretamente de imagens histopatológicas coradas em H&E (Hematoxilina e Eosina). Redes neurais convolucionais (CNNs) têm sido utilizadas para predizer alterações moleculares e biomarcadores de forma supervisionada, explorando a relação entre padrões morfológicos e perfis genômicos. Trabalhos pioneiros, como os de Coudray et al. (2018) em câncer de pulmão e Kather et al. (2019) em câncer colorretal, abriram caminho para este campo, posteriormente expandido para o adenocarcinoma gástrico. Diversos estudos exploraram a supervisão molecular em câncer gástrico.

Flinner et al. (2022) mostraram que modelos de deep learning baseados em H&E podem predizer subtipos do TCGA, mas com desempenho limitado em GS e MSI, destacando a dificuldade em classes menos representadas. Jeong et al. (2022) desenvolveram um classificador para EBV, alcançando AUC-ROC elevado em nível de tiles, embora com precisão moderada em nível de pacientes, sugerindo utilidade para triagem clínica. Zheng et al. (2022) propuseram o modelo de aprendizado profundo EBVNet e mostraram que a fusão humano–máquina supera tanto modelos isolados quanto patologistas, ressaltando a importância da integração entre Inteligência Artificial (IA) e prática médica.

Em outra perspectiva, Zhou et al. (2023) aplicaram CNNs para predizer resposta à quimioterapia neoadjuvante, ampliando o escopo da patologia digital para biomarcadores terapêuticos. Finalmente, a revisão sistemática de Cifici et al. (2022) consolidou evidências de que CNNs podem predizer mutações (TP53, KRAS, BRAF), instabilidade de microssatélites e EBV diretamente de H&E, mas destacou a falta de validação externa robusta como limitação central do campo.

Um marco específico para o câncer gástrico foi o estudo de Wang et al. (2022), que introduziu o modelo DEMoS (Deep learning-based Ensemble approach for Molecular Subtyping). Utilizando o modelo EfficientNet em abordagem de ensemble, os autores obtiveram resultados consistentes, mas ainda insuficientes em classes desbalanceadas, sobretudo EBV e GS. Essa limitação abriu espaço para novas investigações focadas em melhorar a robustez dos modelos para essas categorias. No presente estudo, optamos por combinar três arquiteturas de redes neurais convolucionais com características complementares: MobileNetV2, ShuffleNet e GoogLeNet. A MobileNetV2, proposta por Sandler et al. (2018), introduziu o conceito de inverted residuals e linear bottlenecks, permitindo modelos mais leves e eficientes sem perda expressiva de acurácia, o que a torna particularmente adequada para grandes volumes de tiles histopatológicos, reduzindo o custo computacional. A ShuffleNet, desenvolvida por Zhang et al. (2018), utiliza a técnica de channel shuffle para otimizar a comunicação entre grupos convolucionais, alcançando alta eficiência em dispositivos de baixo custo computacional, sendo útil para cenários em que a escalabilidade do processamento de milhares de tiles é um desafio. Já a GoogLeNet (Inception v1), introduzida por Szegedy et al., 2015, marcou a transição para arquiteturas mais profundas e modulares, com o uso de inception modules que permitem captar padrões em múltiplas escalas dentro de uma mesma camada.

Essa diversidade arquitetural garante que cada rede explore aspectos distintos da morfologia tumoral, aumentando a chance de identificar padrões histopatológicos associados a subtipos moleculares. Assim, a combinação dessas arquiteturas em um ensemble multiarquitetura tem como objetivo potencializar a robustez do modelo, explorando simultaneamente eficiência computacional (MobileNetV2 e ShuffleNet) e capacidade de extração de características complexas (GoogLeNet).

O presente trabalho insere-se nesse contexto, propondo ensembles multiarquitetura com o objetivo de aprimorar a predição dos subtipos moleculares do adenocarcinoma gástrico a partir de imagens histopatológicas do TCGA-STAD. O diferencial metodológico con-

siste em combinar arquiteturas distintas usando métodos de votação diferentes, isto é, em hard e soft voting, visando capturar padrões complementares e reduzir o viés de modelos individuais. Além disso, enfatizamos a precisão como métrica prioritária, Em trajetórias diagnósticas existem dois papéis distintos para testes: (i) testes de triagem/triage (ou de rastreio), nos quais privilegia-se alta sensibilidade (recall) para reduzir falsos-negativos e “não perder casos”; e (ii) testes confirmatórios para diagnóstico diferencial, nos quais privilegia-se alta especificidade e alto valor preditivo positivo (PPV/precisão) para reduzir falsos-positivos e “confirmar” com segurança antes de uma decisão terapêutica. Em contextos de subtipagem molecular do câncer gástrico, o uso pretendido é o diagnóstico diferencial; portanto, precisão/PPV deve ser enfatizada à frente de recall, pois decisões errôneas por falso-positivo podem induzir terapias inadequadas e risco direto ao paciente. Essa priorização está alinhada às recomendações clássicas de avaliação de testes diagnósticos (uso de PPV/NPV na prática clínica real, dependência da prevalência e da probabilidade pré-teste) e às diretrizes para estudos de acurácia diagnóstica (BOSSUYT et al., 2015)

O presente estudo propõe três contribuições principais: Avanço metodológico: ensembles multiarquitetura mostraram ganhos em precisão e F1-score, em especial para EBV e GS, subtipos minoritários e de difícil classificação. Comparação direta com a literatura: demonstramos ganhos substanciais em relação a Wang et al. (2022), especialmente em recall de EBV (+32–34 pontos em nível de tiles), além de aproximações e contrastes com Flinner, Jeong, Zheng e Zhou. Pensamento clínico: ao priorizar precisão, destaca-se aqui a importância dos modelos computacionais serem ferramentas de apoio à tomada de decisão confiável, capazes de minimizar danos iatrogênicos de falsos positivos.

3.2 METODOLOGIA

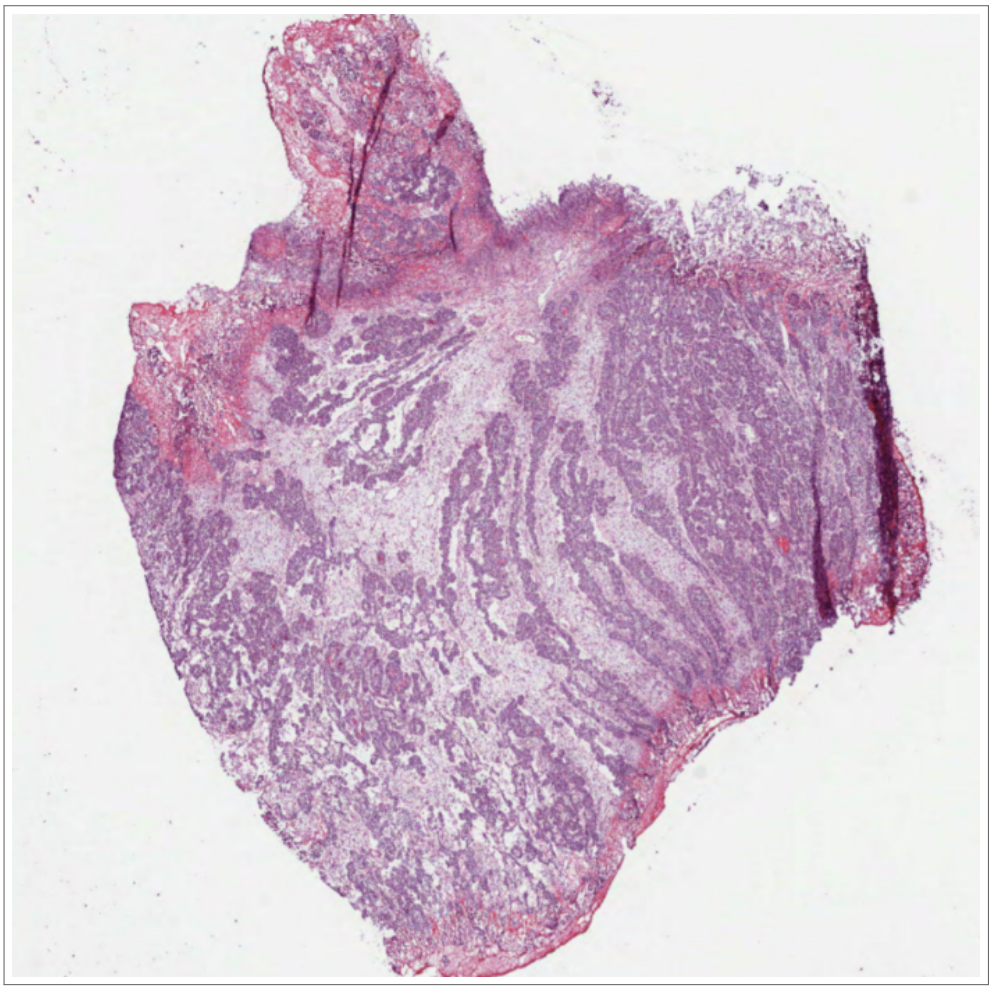
3.2.1 *Dataset*: Conjunto de Dados

O estudo foi realizado utilizando o projeto Stomach Adenocarcinoma (STAD) da base pública do The Cancer Genomic Atlas (TCGA) acessível pelo site (National Cancer Institute, 2025).

3.2.1.1 Imagens de lâminas inteiras

O TCGA disponibiliza imagens de lâminas inteiras (do inglês, Whole Slide Image, WSI) coradas em hematoxilina e eosina (HE) em formato SVS de alta qualidade produzidas por patologia digital a 40x, vide Figura 1. Foram utilizadas 476 lâminas inteiras (referentes a 263 casos do STAD) distribuídas associadas aos rótulos dos subtipos da seguinte maneira: CIN (232 lâminas), seguida por MSI (114 lâminas), GS (73 lâminas) e EBV (57 lâminas).

Figura 1 – Imagem de uma WSI inteira.



Fonte: Extraída de The Cancer Genome Atlas (TCGA), Cancer Genome Atlas Research Network (2014)

3.2.2 Pré-processamento das imagens

O pré-processamento das imagens compreendeu três etapas principais que podem ser observada da Figura 2: (I) segmentação em *tiles* de 224×224 px; (II) detecção e exclusão de imagens borradas; e (III) normalização de cor.

3.2.2.1 Segmentação de Imagem

O Corte das WSI (*Tiling*) foi realizado nas 476 imagens de lâminas inteiras, classificadas em conjuntos segundo as 4 classes distintas em conformidade com o subtipo molecular (CIN, MSI, EBV, GS). Foram incluídos nos metadados da WSI os rótulos correspondentes aos casos. Foi então utilizada a função deepzoom da biblioteca open slide na versão 1.3.1 (OpenSlide, 2023) com fator de ampliação de 10x. As *WSI* foram cortadas em *tiles* no formato de 224x224 pixels. Foram removidos aqueles contendo mais de 50% de fundo. Foram removidas também as imagens borradas utilizando o filtro Laplaciano conforme mostra esquema 2. Ao final desse processo, foram obtidos na ordem de 1.500.000 *tiles*, com distribuição entre os subtipos moleculares CIN (maior proporção, acima de 400.000 tiles), MSI (cerca de 200.000), GS (aproximadamente 150.000) e EBV (em torno de 150.000).

3.2.2.2 Normalização de cor

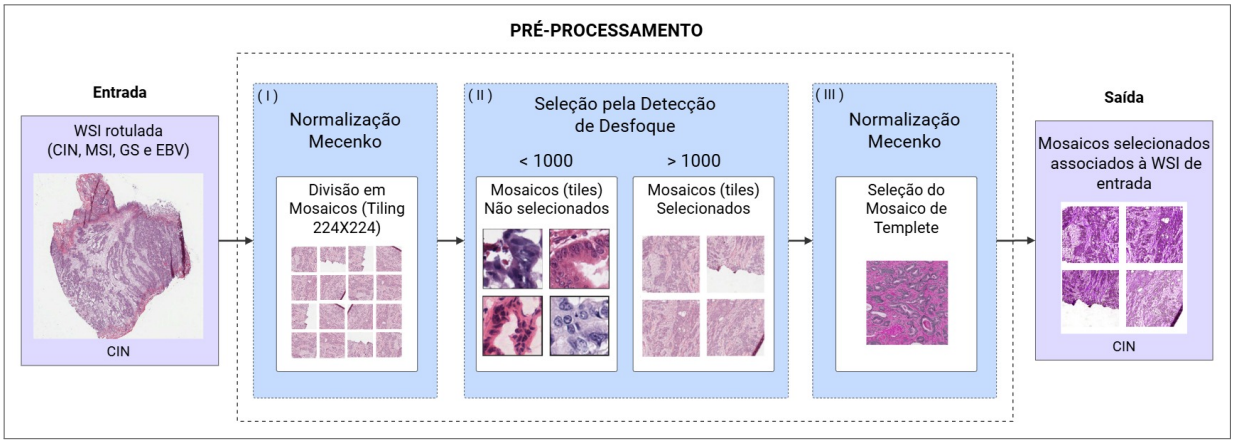
Para abordar a variabilidade na coloração das imagens, implementamos a normalização de cor baseada no método Macenko, conforme eq 3.1. aplicado com uso de um modelo de referência para ajustar ao espaço cromático e padronização de luminosidade (MACENKO et al., 2009).

$$I_{\text{norm}} = I_0 \cdot \exp\left(-\text{decomv}(OD_{\text{fonte}}, S_{\text{fonte}}) \cdot S_{\text{alvo}}\right) \tag{3.1}$$

onde:

- I_{norm} é a intensidade do pixel na imagem normalizada (em RGB)
- I_0 é a intensidade de luz transmitida (geralmente 255 para imagens de 8 bits)
- $OD_{\text{fonte}} = -\log_{10}(I_{\text{fonte}}/I_0)$ é a Densidade Óptica da imagem original
- S_{fonte} é a matriz de vetores extraída da imagem original
- decomv é o processo de deconvolução
- S_{alvo} é a matriz de vetores da imagem de referência

Figura 2 – Fluxograma do corte e processamento das imagens do TCGA.



Fonte: O autor

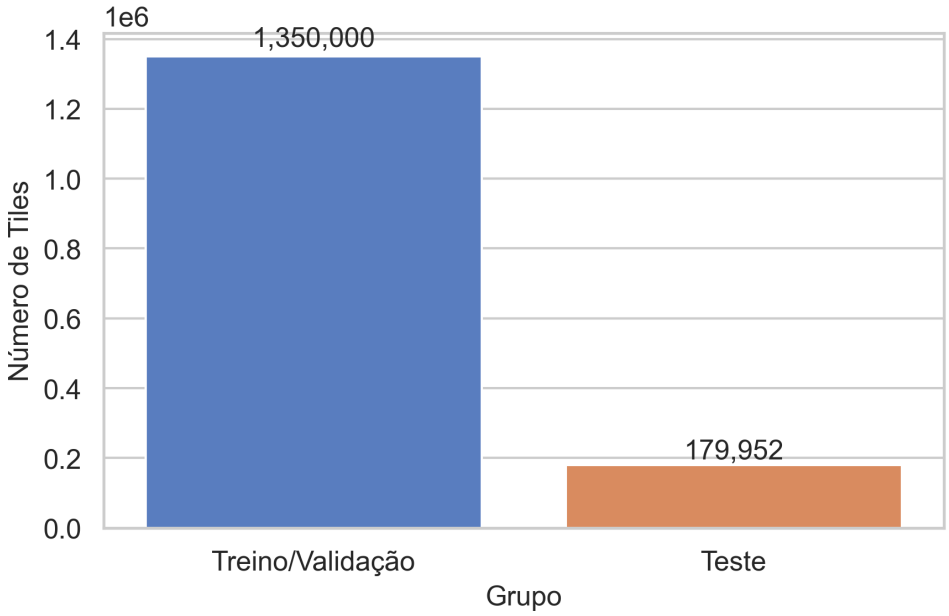
3.3 AGRUPAMENTOS: TREINO, VALIDAÇÃO E TESTE

Nesta etapa, foram criados aleatoriamente dois grupos de *tiles* pré-processados: Um grupo treino/validação com 380 lâminas (1.350.000) *tiles* e um grupo teste *hold out* com 96 lâminas (179.952 tiles), conforme a distribuição da Figura 3.

3.3.1 Grupo Treinamento/Validação

Os grupos de treinamento e validação para cada modelo (explicados no tópico treinamento abaixo) foram separados utilizando o método *K-fold*. Uma utilização parcial do *cross validation* com o objetivo apenas de gerar múltiplas separações aleatórias de conjuntos de treinamento e validação na intenção de reduzir viés em grupos de validação de classes minoritárias. Na proporção 90/10. Consequentemente, treinando 10 modelos. Foi utilizado K=10 e, portanto, foram treinados 10 modelos para cada arquitetura (explicadas no tópico treinamento abaixo) , variando os grupos de teste e validação.

Figura 3 – Proporção da distribuição dos tiles entre os grupos

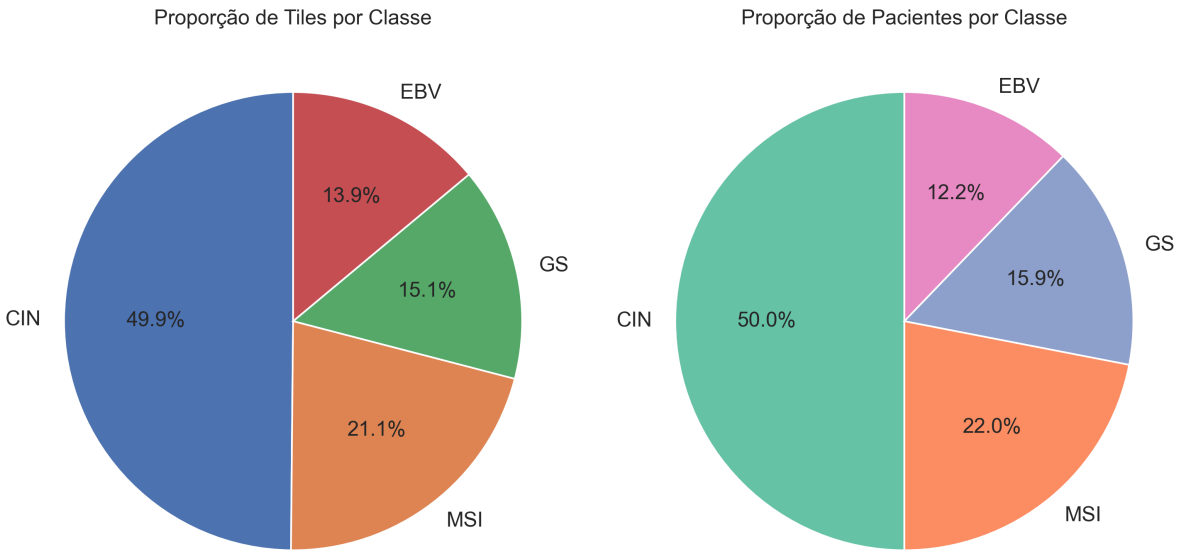


Fonte: O autor (2025).

3.3.2 Grupo Teste

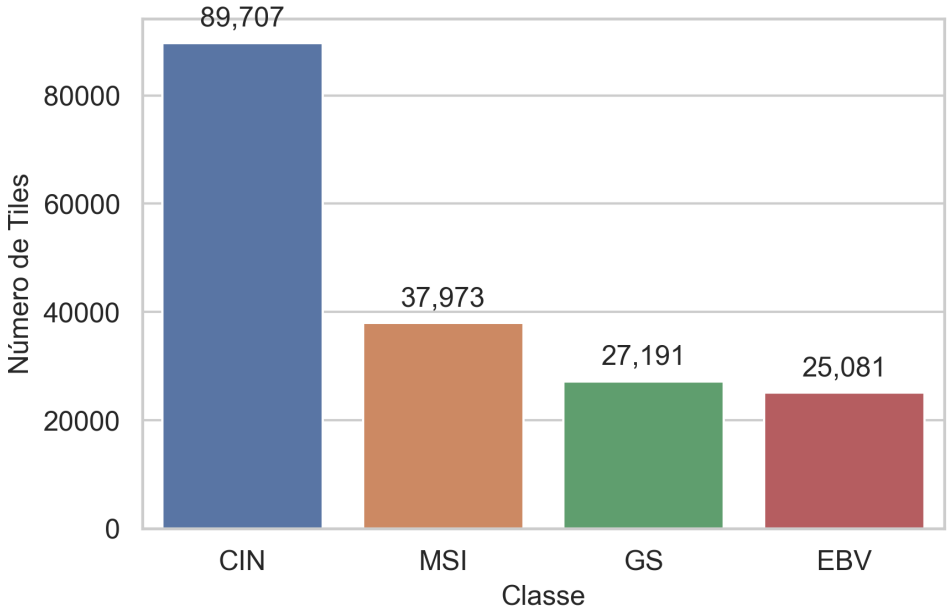
96 lâminas foram separadas para uso no teste final (*hold out*). Medidas foram tomadas para evitar contaminação das imagens de lâminas inteiras do conjunto teste com os outros conjuntos. Após análise da distribuição de lâminas no grupo teste, foram acrescentados aleatoriamente mais casos das classes minoritárias. Totalizando 96 lâminas de 82 casos (179.952 tiles). Distribuídos CIN: 89.707 tiles, provenientes de 47 lâminas inteiras (WSIs), correspondentes a 41 pacientes. MSI: 37.973 tiles, de 23 WSIs, representando 18 pacientes. GS: 27.191 tiles, de 15 WSIs, relativos a 13 pacientes. EBV: 25.081 tiles, oriundos de 11 WSIs, correspondentes a 10 pacientes. Essa distribuição fica melhor evidenciada pelas Figuras [4 - 5 - 6]

Figura 4 – Distribuição do Grupo Teste (Hold-out)



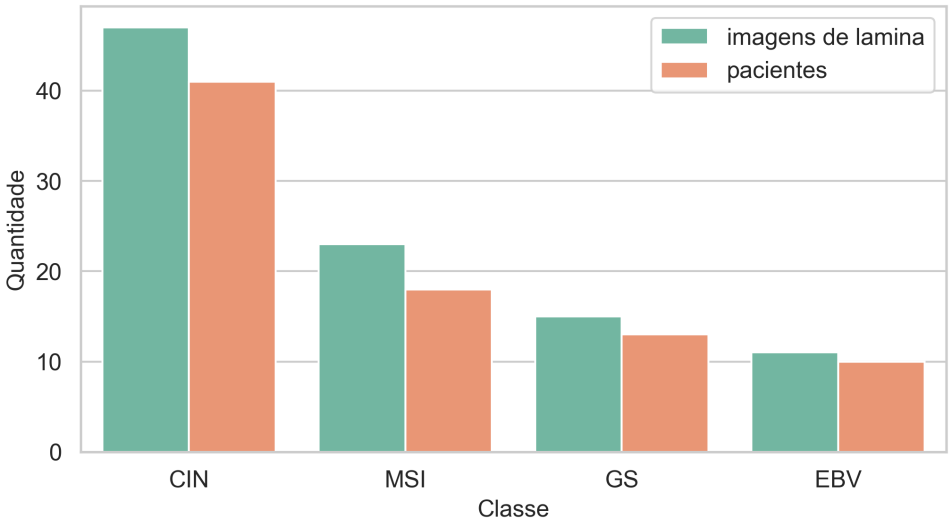
Fonte: O autor (2025).

Figura 5 – Distribuição de tiles por classe



Fonte: O autor (2025).

Figura 6 – Distribuição de imagem de lâmina por paciente por classe



Fonte: O autor (2025).

3.4 TREINAMENTO

O treinamento foi desenvolvido utilizando a versão Python 3.8.20, em um ambiente virtual criado para isolar as dependências e garantir a reprodutibilidade dos resultados. Foram utilizadas as bibliotecas scikit-learn 1.2.2, Pandas 1.5.3, PyTorch 2.4.1+cu118 com documentação detalhada e suas dependências. Foram treinadas quatro redes neurais convolucionais *EfficientNet*, *MobileNetV2*, *GoogleNet* e *ShuffleNet*. Todas as redes foram inicializadas com pesos pré-treinados no *ImageNet* e tiveram suas camadas finais adaptadas para a classificação dos quatro subtipos moleculares do adenocarcinoma gástrico (EBV, MSI, GS e CIN), utilizando os *tiles* do conjunto de treinamento/validação. A otimização foi realizada com o algoritmo *Adam*, empregando taxa de aprendizado inicial de 0,001 e *weight decay* de 1×10^{-4}. A taxa de aprendizado foi ajustada de forma adaptativa pelo agendador *ReduceLROnPlateau*, configurado para reduzir a *learning rate* em um fator de 0,5 sempre que o valor da função de custo no conjunto de validação permanecesse inalterado por três épocas consecutivas, considerando um limiar de 1×10^{-8}. O treinamento foi conduzido por até 50 épocas, com mecanismo de *Early Stopping* e paciência de cinco épocas, interrompendo automaticamente o processo caso não fosse observada melhoria na função de custo no conjunto de validação nesse intervalo. Para cada arquitetura, foram treinados 10 modelos independentes, correspondendo aos 10 *folds* definidos na etapa

anterior. Durante o treinamento, a função de custo adotada foi uma composição de duas funções: a função padrão de entropia cruzada (*Cross Entropy Loss*, $\mathcal{L}CE$) e a *Macro Soft F1 Loss* ($\mathcal{L}F1$). Especificamente, a *loss* final foi definida como a média aritmética dessas duas funções:

$$\mathcal{L} = \frac{1}{2}\big(\mathcal{L}CE + \mathcal{L}F1\big)$$

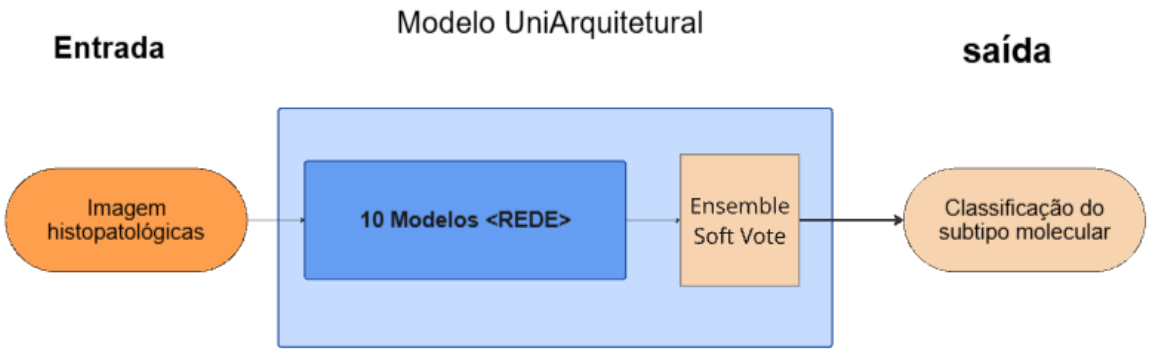
Essa formulação teve como objetivo combinar a estabilidade da *Cross Entropy Loss* na otimização com a capacidade da *Macro Soft F1 Loss* de promover melhor equilíbrio entre precisão (*precision*) e sensibilidade (*recall*) nas diferentes classes, especialmente em cenários de desbalanceamento.Durante o treinamento, além da função de custo padrão (*Cross Entropy Loss*), foi incorporada a função *Macro Soft F1 Loss*, com o objetivo de aprimorar o equilíbrio entre precisão (*precision*) e sensibilidade (*recall*) nas diferentes classes. Os valores intermediários, como a função de custo no conjunto de treino e no conjunto de validação, bem como a taxa de aprendizado, foram monitorados e registrados no *Tensor Board* para acompanhamento e análise posterior. Ao término de cada *fold*, o modelo de melhor desempenho foi armazenado, e os gráficos de evolução das funções de custo foram exportados para análise visual.

3.5 METODOLOGIA DE ENSEMBLES

Ensembles uniarquitetura (Single Architecture SA)

Foi formado um comitê de modelos (Ensemble) que contabiliza os votos levando em conta duas abordagens: (a) o vetor de probabilidade da confiança (soft voting) para chegar a uma predição ou (b) o vetor das classes preditas, para os 10 modelos de uma única arquitetura, conforme ilustra a Figura 7. Na abordagem soft voting, a soma das probabilidades de confiança de cada modelo é realizada para definir a classe predita.

Figura 7 – Ensemble Uniarquitetura (SA)

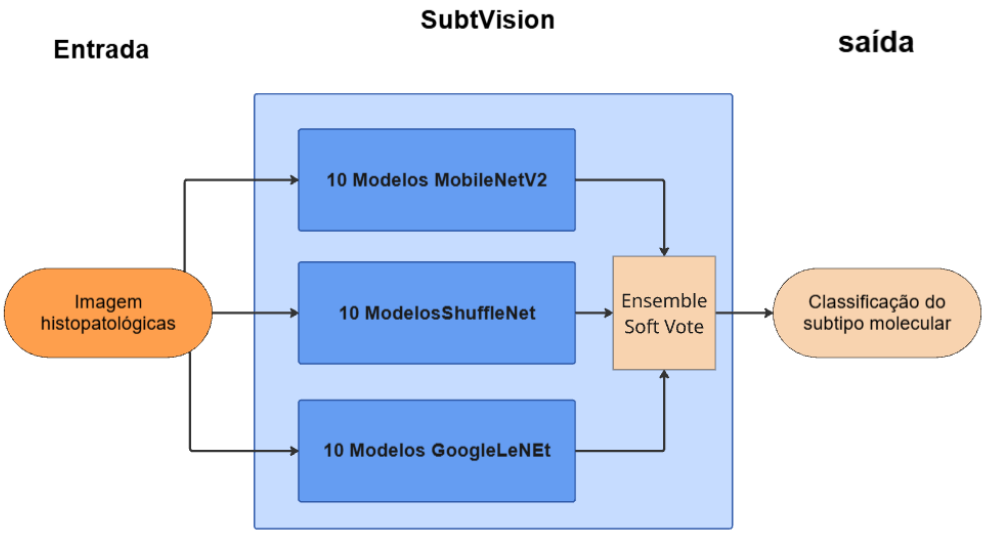


Fonte: O autor (2025).

3.5.1 Ensembles multiarquiteturas (MA)

Os Ensembles MA foram construídos com três arquiteturas de rede (MobileNet V2, ShuffleNet e GoogleNet) Figura 8. Os 30 modelos (10 modelos de cada uma das 3 arquiteturas selecionadas) foram consolidados em um comitê de modelos (Ensemble). Os métodos de ensemble contabilizaram os votos majoritários (*hard voting*) ou os votos ponderados pela confiança (*soft voting*).

Figura 8 – SubtVision



Fonte: O autor (2025).

3.6 MÉTRICAS UTILIZADAS

As métricas foram computadas usando a biblioteca scikit-learn, versão 1.2.2, incluindo médias macro (não ponderadas) e weighted (ponderadas por classe). Para acompanhamento mais detalhado dos resultados, foram utilizados relatórios por fold e ensemble, com curvas ROC visualizadas via TensorBoard. As métricas foram calculadas nos níveis de tiles e consolidadas para o nível dos pacientes, incluindo: Precisão (eq 3.2) – Proporção de predições positivas corretas, ela expressa a confiança no diagnóstico positivo, já que os falsos positivos vão reduzir essa métrica. A precisão expressa a mesma intenção da especificidade, porém o faz ao representar a proporção de verdadeiros positivos no total de positivos indicados pelo modelo. Sensibilidade ou Recall (eq 3.3) – Proporção de positivos reais corretamente identificados, expressa, portanto, a proporção de verdadeiros positivos sobre o total de casos positivos, já que o total de casos positivos é a soma dos verdadeiros positivos com os falsos negativos. F1-Score (eq 3.4) – Média harmônica de precisão e recall, é uma métrica que combina precisão e recall em uma única medida, oferecendo um balanço entre a capacidade de identificar corretamente os positivos (sensibilidade - recall) e a confiabilidade dessas predições (precisão). AUC-ROC: Área sob a curva ROC (one-vs-rest por classe) é uma ferramenta gráfica utilizada para avaliar o desempenho de um modelo de classificação binária, representando o trade-off entre a taxa de verdadeiros positivos (Recall) (TPR) e a taxa de falsos positivos (FPR) à medida que a confiança do modelo na predição aumenta. A área abaixo da curva ROC (neste texto chamada de AUC-ROC) corresponde à medida numérica obtida ao calcular a área sob a curva ROC. Intuitivamente, ela representa a probabilidade de o modelo atribuir um valor de score mais alto para uma instância positiva do que para uma negativa escolhida aleatoriamente. Quanto maior a AUC-ROC (próxima de 1), melhor a capacidade de separação entre as classes; valores próximos de 0,5 indicam um modelo aleatório, e valores abaixo disso sugerem um modelo que classifica pior do que o acaso.

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{3.2}$$

$$\text{Recall} = \frac{VP}{VP + FN} \tag{3.3}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}} \tag{3.4}$$

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(\text{FPR}) d(\text{FPR}) \tag{3.5}$$

onde:

- VP = Verdadeiros Positivos (*True Positives*)
- FP = Falsos Positivos (*False Positives*)
- FN = Falsos Negativos (*False Negatives*)
- VN = Verdadeiros Negativos (*True Negatives*)
- $\text{TPR} = \frac{\text{VP}}{\text{VP}+\text{FN}}$ (Taxa de Verdadeiros Positivos ou *Recall*)
- $\text{FPR} = \frac{\text{FP}}{\text{FP}+\text{VN}}$ (Taxa de Falsos Positivos)

3.6.1 Teste de Wilcoxon Signed-Rank

O teste de Wilcoxon Signed-Rank foi aplicado para comparar os resultados dos ensembles com os modelos individuais ao longo dos folds de validação cruzada. Por ser um método não paramétrico. Para uma explicação mais detalhada sobre os fundamentos e a aplicação do teste, consulte a documentação da Statsoft sobre o tema (acessando a seção "Wilcoxon matched pairs test").

3.7 RESULTADOS E DISCUSSÃO

Este capítulo apresenta resultados e discussões dos experimentos realizados. No primeiro momento, serão apresentados os resultados a nível de tile, em seguida a nível de paciente, discutindo-se como as métricas de precisão, revocação, F1-Score e AUC-ROC se comportaram nos modelos individuais e o G.SubtVision, usando a abordagem de votos soft voting.

3.7.1 Nível dos TILEs

PRECISION no nível do TILE

No nível dos tiles, os ensembles multiarquitetura (MA) com hard voting e soft voting (G.SubtVision), alcançaram médias macro de precisão de 0,53 e 0,56, respectivamente, tendo o segundo superado as arquiteturas individuais MobileNetV2 (0,55), ShuffleNet (0,53) e GoogLeNet (0,48), como apresentados na Tabela 1. Por classe, os valores para CIN variaram de 0,61 a 0,64 nos modelos individuais e 0,62-0,63 nos ensembles; para EBV, de 0,40 a 0,62 nos individuais e 0,50-0,62 nos ensembles; para MSI, de 0,51 a 0,65 nos individuais e 0,61-0,65 nos ensembles.

Tabela 1 – Comparação de PRECISION entre modelos (no nível de tiles).

Classe	Wang et al. (2022)	Reprodução EfficientNet	Mobile NetV2	Shuffle Net	Google LeNet	Ensemble MA Hard	G.Subt Vision
CIN	0,45	0,61	0,62	0,61	0,64	0,63	0,62
EBV	0,14	0,56	0,61	0,62	0,40	0,50	0,62
GS	0,76	0,42	0,31	0,26	0,36	0,36	0,34
MSI	0,56	0,51	0,65	0,62	0,53	0,61	0,65
Macro AVG	0,48	0,52	0,55	0,53	0,48	0,53	0,56
Weighted AVG	0,55	0,55	0,58	0,56	0,54	0,57	0,59

Fonte: O autor (2025).

Comparando com Wang et al. (2022), que reportou média macro de 0,48, observou-se melhoria de 8 pontos percentuais com o G.SubtVision, mas ao comparar com a reprodução realizada aqui com EfficientNet, a precisão melhorou apenas 4 pontos percentuais (WANG et al., 2022).

O único subtipo que apresentou redução de precisão foi GS, de 0,42 na reprodução de Wang com EficientNet a 0,36 nos individuais e 0,34-0,36 nos ensembles; Em relação a Flinner et al. (2022), que também analisaram os quatro subtipos em TCGA e UKC, os valores de precisão relatados foram próximos aos observados aqui (CIN = 0,53; EBV = 0,52; MSI = 0,43; GS = 0,55), confirmando a consistência metodológica e a dificuldade comum em MSI e GS. O maior aumento de precisão foi encontrado na categoria EBV com melhora de 48 pontos percentuais em relação a Wang et al. 2022 e 10 pontos percentuais em relação a Flinner et al (2022); A precisão para MSI aumentou entre 9 pontos percentuais

em relação à Wang e 22 em relação a Flinner. Na categoria CIN a precisão aumentou entre 17 pontos em relação à Wang e 9 pontos em relação à Flinner. Ao superar limitações comuns de abordagens individuais, esta estratégia demonstra robustez metodológica e relevância clínica, uma vez que a redução de falsos positivos e amplia a confiabilidade diagnóstica dos modelos e CNN.

3.7.1.1 Recall no nível do TILE

No nível dos tiles, os ensembles MA com hard voting e G.SubtVision com soft voting, obtiveram médias macro de recall de 0,46 e 0,47, respectivamente,ligeiramente superiores às arquiteturas individuais (0,44-0,46), como apresentados na Tabela 2. Por classe, CIN apresentou recall de 0,74-0,84 nos individuais e 0,80 nos ensembles; EBV, de 0,30-0,44 nos individuais e 0,39 nos ensembles; GS, de 0,24-0,44 nos individuais e 0,39-0,40 nos ensembles; e MSI, de 0,23-0,33 nos individuais e 0,27-0,28 nos ensembles. A média ponderada foi de 0,57 nos ensembles. O recall mede a capacidade de detectar instâncias verdadeiras, crítico para classes minoritárias em contextos de triagem. Os modelos do presente estudo apresentaram sensibilidade elevada para CIN (até 0,84 nos modelos individuais e 0,80 nos ensembles MA). O que representou um aumento de 32 pontos percentuais em relação aos resultados do artigo publicado por Wang em 2022, cuja metodologia foi reproduzida. A sensibilidade na categoria MSI do modelo reprodução da metodologia, ao contrário, da categoria CIN, foi pior 45 pontos que o artigo previamente publicado. Esse modelo reprodução mesmo assim foi melhor que os resultados do ensemble MA em 6 ou 7 pontos percentuais. Para as categorias minoritárias, os modelos aqui apresentados obtiveram resultados significativamente melhores que os anteriormente publicados. Para o subtipo GS, o modelo G.SubtVision demonstrou resultados 8 pontos percentuais mais sensíveis que os resultados publicados e 16 pontos mais sensíveis que a reprodução da metodologia no presente agrupamento aleatório entre o grupo de treinamento e o teste. No grupo EBV foi identificado o maior avanço de sensibilidade com aumento de 23 a 34 pontos percentuais em relação à metodologia de Wang 2022.

Tabela 2 – Comparação de RECALL entre modelos (no nível de tiles).

Classe	Wang et al. (2022)	Reprodução EfficientNet	Mobile NetV2	Shuffle Net	Google LeNet	Ensemble MA Hard	G.Subt Vision
CIN	0,52	0,84	0,78	0,81	0,74	0,80	0,80
EBV	0,07	0,35	0,30	0,40	0,44	0,39	0,39
GS	0,32	0,24	0,44	0,25	0,43	0,39	0,40
MSI	0,78	0,33	0,31	0,32	0,23	0,27	0,28
Macro AVG	0,42	0,44	0,46	0,44	0,46	0,46	0,47
Weighted AVG	0,54	0,57	0,56	0,56	0,54	0,57	0,57

Fonte: O autor (2025).

3.7.1.2 F1 score no nível do TILE

No nível dos tiles, os ensembles MA e G.SubtVision registraram médias macro de F1-score de 0,47 e 0,48, respectivamente, superando ligeiramente as individuais (0,45-0,47), como é possível observar na Tabela 3. Por classe, CIN variou de 0,68-0,71 nos individuais e 0,70 nos ensembles; EBV, de 0,40-0,48 nos individuais e 0,44-0,48 nos ensembles; GS, de 0,25-0,40 nos individuais e 0,36-0,38 nos ensembles; e MSI, de 0,32-0,42 nos individuais e 0,38-0,39 nos ensembles. A média ponderada foi de 0,55-0,56 nos ensembles. Comparando com Wang et al. (2022), que apresentou média macro de 0,42, os ensembles melhoraram 5-6 pontos, com ganhos notáveis em EBV (31-39 pontos em relação aos 0,09). Para GS, houve redução em relação aos 0,45 de Wang, mas melhoria de 5-9 pontos sobre a reprodução com EfficientNet (0,31). Na literatura, F1-scores em ensembles para histopatologia gástrica, como em Li et al. (2022), variam de 0,44-0,49, enfatizando a robustez dos ensembles em datasets desbalanceados, alinhando-se aos nossos achados onde o F1 reflete equilíbrio melhorado em classes majoritárias.

F1 score expressa o equilíbrio entre sensibilidade e precisão, sendo aqui utilizado para análise geral do equilíbrio do desempenho dos modelos. No nível dos tiles, os ensembles alcançaram média macro F1 de 0,48, superando em 6 pontos os 0,42 de Wang, superando ainda em 2 pontos os resultados da reprodução com Efficient Net. Apresentou melhorias notáveis em EBV com 31 a 39 pontos percentuais (0,48 vs. 0,09). No entanto, GS permaneceu desafiador (0,36). O ensemble multiarquiterura melhora a robustez geral.

Tabela 3 – Comparação de F1-score entre modelos (no nível de tiles).

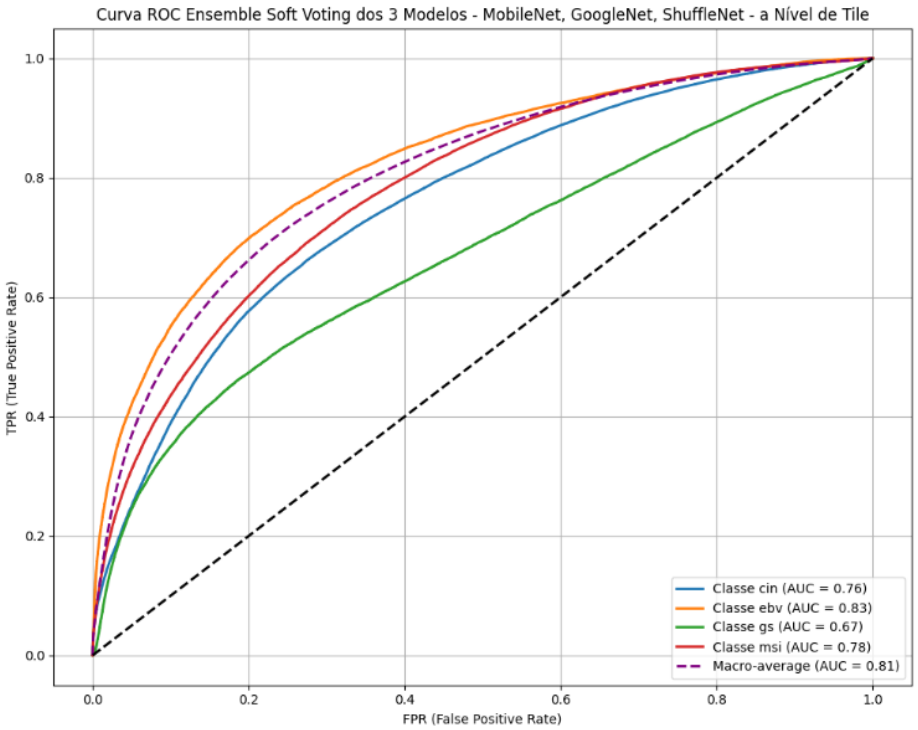
Classe	Wang et al. (2022)	Reprodução EfficientNet	Mobile NetV2	Shuffle Net	Google LeNet	Ensemble MA Hard	G.Subt Vision
CIN	0,48	0,71	0,69	0,70	0,68	0,70	0,70
EBV	0,09	0,43	0,40	0,48	0,42	0,44	0,48
GS	0,45	0,31	0,36	0,25	0,40	0,38	0,36
MSI	0,66	0,40	0,42	0,42	0,32	0,38	0,39
Macro AVG	0,42	0,46	0,47	0,46	0,45	0,47	0,48
Weighted AVG	0,51	0,54	0,54	0,54	0,53	0,55	0,56

Fonte: O autor (2025).

3.7.1.3 AUC-ROC no nível do TILE

No nível dos tiles, o G.SubtVision obteve AUC-ROC de 0,76 para CIN, 0,83 para EBV, 0,67 para GS e 0,78 para MSI, com média macro implícita de 0,81, como é possível observar na Tabela 4. Comparando com Wang et al. (2022), que reportou AUC-ROC de 0,762 (CIN), 0,668 (EBV), 0,785 (GS) e 0,811 (MSI), observou-se melhoria em EBV (aumento de 0,162) e redução em GS (0,115) e MSI (0,031), possivelmente devido a vieses na confiança das predições. A área sob a curva ROC foi construída, como é bastante frequente na análise de resultados de redes neurais, a partir da confiança dos modelos em suas predições. Aqui se enfatiza, do ponto de vista de estatística médica, essa confiança é “auto-declarada” pelo modelo. Ao analisar os resultados de Wang para a categoria EBV, percebe-se como essa métrica pode ser enviesada. O F1 score para o EBV em Wang (encontrado em seu material suplementar) foi de apenas 0.09 (precisão foi 0.14 e recall 0.07); no entanto, a área sob a curva ROC para EBV foi 0.67. Uma provável razão para esse aparente viés é que a confiança do modelo em suas predições teve magnitudes desalinhadas com o acerto do ground truth pela predição do modelo.

Figura 9 – AUC-ROC Ensemble Soft Voting 3 arquiteturas (nível do tile)



Fonte: O autor (2025).

Tabela 4 – Comparação da Curva ROC (AUC-ROC) entre modelos (no nível de tiles).

Curva ROC	Wang et al. (2022)	Reprod. EfficientNet	Mobile NetV2	Shuffle Net	Google LeNet	G.Subt Vision
CIN	0,762	0,75	0,74	0,73	0,74	0,76
EBV	0,668	0,83	0,81	0,81	0,76	0,83
GS	0,785	0,69	0,67	0,62	0,68	0,67
MSI	0,811	0,75	0,79	0,72	0,74	0,78
Macro AVG		0,81	0,80	0,79	0,78	0,81

Fonte: O autor (2025).

3.7.2 Resultados no nível dos pacientes

3.7.2.1 PRECISION dos modelos no nível do paciente

Na Tabela 5 pode-se observar que no nível dos pacientes, os ensembles MA e o G.SubtVision alcançaram médias macro de precisão de 0,80, com valores por classe: CIN

0,57-0,58; EBV 1,00; GS 0,62; MSI 1,00. A média ponderada foi de 0,73. Comparando com Wang et al. (2022), que obteve média macro de 0,77, os ensembles melhoraram 3 pontos, com perfeição em EBV e MSI (1,00 vs. 1,00 e 0,65 de Wang). Para GS, redução de 21 pontos (0,62 vs. 0,83). A reprodução com EfficientNet obteve o melhor resultado apresentando um aumento de 17 pontos (1,00 vs 0,83 de Wang). No subtipo CIN pode-se observar que os resultados entre os ensembles e os dados publicados por Wang foram semelhantes. Porém ao comparar o nível dos tiles com o nível do paciente observa-se que os resultados de Wang apresentaram melhora de 13 pontos percentuais (de 0.45 a 0.58), por outro lado a reprodução reduziu 3 pontos (de 0.61 para 0.55). Os Resultados dos ensembles reduziram 5 pontos (de 0.63 ou 0.62 no nível do tile para 0.58 ou 0.57).

Para EBV no nível do paciente, Wang publicou precisão de 1.00. Partindo no nível do Tile de uma precisão de 0,14. A reprodução do artigo partindo de 0.56 de precisão atingiu também 1.00 de pecisão ao nível do paciente. Ou seja, no nível do tile, a razão (Verdadeiros Positivos / [Verdadeiros Positivos + Falsos Negativos]) atingiu 0.14, mas a agregação por maior frequência (nível do paciente) eliminou o falso negativo. Os modelos de ensemble apresentou precisão de 1,00, ou seja, não houve falsos positivos. O mesmo ocorreu para o MSI, no qual os ensemble MA atingiram 1,00, ou seja, zeraram os falsos positivos. Para o subtipo GS, o resultado foi desafiador 0,62, abaixo de 21 pontos do resultado de Wang, e 38 pontos abaixo da reprodução.

Tabela 5 – PRECISION: Wang et al. (2022) vs. ensemble MA com soft voting em nível de pacientes

Classe	Wang et al. (2022)	Reprodução EfficientNet	Mobile NetV2	Shuffle Net	Google LeNet	Ensemble MA Hard (PACIENTE)	G.Subt Vision (PACIENTE)
CIN	0,58	0,55	0,57	0,61	0,58	0,58	0,57
EBV	1,00	1,00	1,00	0,62	0,57	1,00	1,00
GS	0,83	1,00	0,57	0,26	0,71	0,62	0,62
MSI	0,65	0,60	1,00	0,62	0,50	1,00	1,00
Macro AVG	0,77	0,79	0,78	0,53	0,59	0,80	0,80
Weighted AVG	0,71	0,69	0,71	0,56	0,58	0,73	0,73

Fonte: O autor (2025).

Um achado inesperado em relação aos resultados publicados por Wang foi a irregularidade do aumento ou diminuição das métricas no nível do paciente comparado ao nível dos Tiles. Em outras palavras, Wang apresentou após a agregação do nível dos tiles para

o nível do paciente aumento em todas as métricas. Aqui isso não foi verificado, houve aumentos e reduções das métricas ao agregar o nível do tile para o nível do Paciente, usando o mesmo método. Os resultados e discussão do nível do paciente serão apresentados seguindo a mesma sequência: primeiro precision seguida de *recall*, f-1 Scores e Área sob a curva Roc.

3.7.2.2 RECALL no nível do paciente

No nível dos pacientes, o ensemble MA e o G.SubtVision obtiveram médias macro de *recall* de 0,46 e 0,45, respectivamente. Por classe: CIN 0,95; EBV 0,30; GS 0,38; MSI 0,17-0,22. Média ponderada de 0,61-0,62. A Tabela 6 mostra que os resultados de CIN foram equilibrados tendendo mais à sensibilidade 0,95 que à precisão descrita acima. Na categoria EBV o contrário acontece, o ensemble é preciso (1.00), mas muito menos sensível (0.30) o que significa que quando indica EBV essa predição é confiável, mas que no entanto quando o resultado são outras categorias pode estar sendo falso negativo para EBV. O mesmo acontece para MSI, o G.SubtVision é preciso para MSI (1.00) no entanto a sensibilidade foi bem baixa para MSI (0,17) sendo a categoria mais desafiadora para o modelo G.SubtVision aqui proposto.

Comparando com Wang et al. (2022), que reportou média macro de 0,49, houve leve redução de 3 a 4 pontos, com ganhos em CIN (41 pontos) mas perdas em MSI (72-77 pontos). Ao comparar os resultados do autor com a reprodução de sua metodologia (EfficientNet) observa-se que a reprodução foi melhor que os resultados previamente publicados na categoria CIN (0.98 vs 0.54) e bem inferior na categoria MSI (0.60 vs 0.94).

3.7.2.3 F1 Paciente

No nível dos pacientes, os ensembles MA e G.SubtVision alcançaram médias macro de F1-score de 0,51 e 0,48. Por classe: CIN 0,72; EBV 0,46; GS 0,48; MSI 0,29-0,36. Média ponderada de 0,55-0,57, como mostrado na Tabela 7. Comparando com Wang et al. (2022), que obteve 0,52 na média macro, houve variação mínima (-4 a -1 ponto), com ganhos em CIN (16 pontos) mas reduções em MSI (41-48 pontos). Superior à EfficientNet (0,42) em 6-9 pontos na média macro. Na literatura, F1 paciente em subtipagem, como em Wang et al. (2023) extensão, atinge 0,45-0,55, reforçando a agregação como desafio.

Tabela 6 – RECALL: Wang et al. (2022) vs. MA com soft voting em nível de pacientes

Classe	Wang et al. (2022) (PCT_IMG)	Reprodução EfficientNet	Mobile NetV2	Shuffle Net	Google LeNet	Ensemble MA Hard (PACIENTE)	G.Subt Vision (PACIENTE)
CIN	0,54	0,98	0,95	0,81	0,93	0,95	0,95
EBV	0,14	0,30	0,30	0,40	0,40	0,30	0,30
GS	0,36	0,15	0,31	0,25	0,38	0,38	0,38
MSI	0,94	0,60	0,17	0,32	0,06	0,22	0,17
Macro AVG	0,49	0,40	0,43	0,44	0,44	0,46	0,45
Weighted AVG	0,66	0,59	0,60	0,56	0,59	0,62	0,61

Fonte: O autor (2025).

Tabela 7 – F1-score: Wang et al. (2022) vs. ensemble MA com soft voting em nível de pacientes

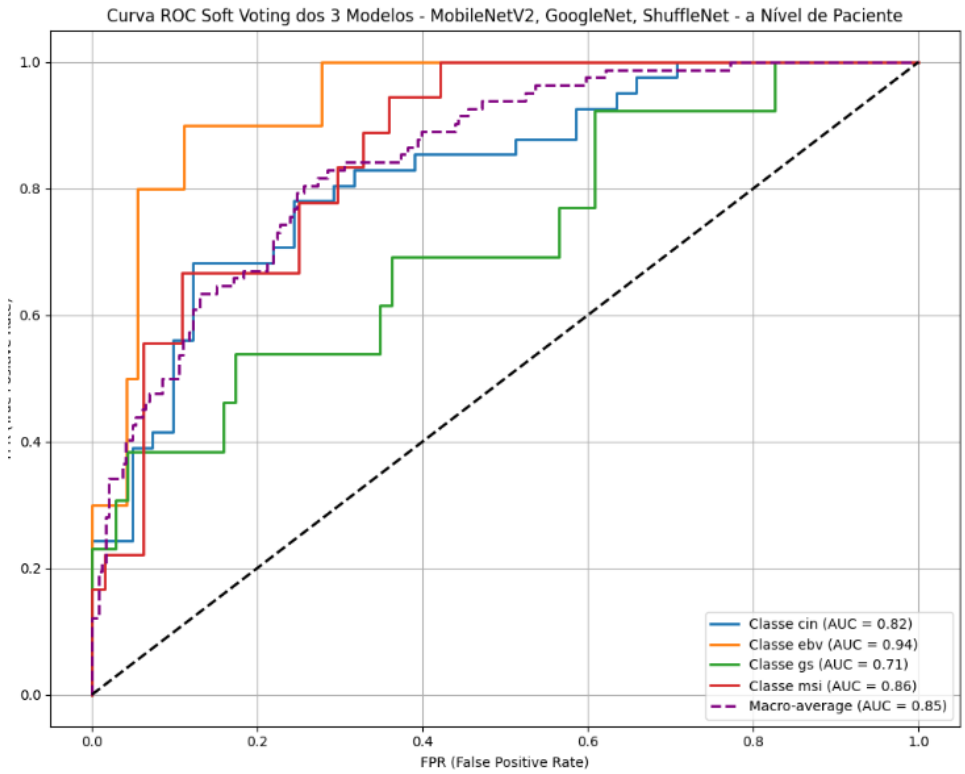
Classe	Wang et al. (2022) (PCT_IMG)	Reprodução EfficientNet	Mobile NetV2	Shuffle Net	Google LeNet	Ensemble MA Hard (PACIENTE)	G.Subt Vision (PACIENTE)
CIN	0,56	0,71	0,71	0,70	0,71	0,72	0,72
EBV	0,25	0,46	0,46	0,48	0,47	0,46	0,46
GS	0,50	0,27	0,40	0,25	0,50	0,48	0,48
MSI	0,77	0,26	0,29	0,42	0,10	0,36	0,29
Macro AVG	0,52	0,42	0,46	0,46	0,45	0,51	0,48
Weighted AVG	0,62	0,59	0,54	0,54	0,51	0,57	0,55

Fonte: O autor (2025).

3.7.2.4 AUC-ROC paciente

É apresentado No nível dos pacientes, o ensemble MA soft voting registrou AUC-ROC de 0,82 (CIN), 0,94 (EBV), 0,71 (GS) e 0,86 (MSI), com média macro de 0,85, como é possível observar na Tabela 8. Comparando com Wang et al. (2022), que reportou 0,890 (CIN), 0,764 (EBV), 0,897 (GS) e 0,898 (MSI), observou-se melhoria em EBV (0,176) e redução em GS (0,187). Superior à EfficientNet em média (0,73 vs. 0,85). Na literatura, AUC-ROC paciente em ensembles gástricos, como em Huang et al. (2022), varia de 0,80-0,90, com ensembles elevando valores em EBV.

Figura 10 – AUC-ROC G.SubtVision (nível do paciente)



Fonte: O autor (2025).

Tabela 8 – AUC-ROC: Wang et al. (2022) vs. ensemble MA com soft voting em nível de pacientes

Classe	Wang et al. (2022)	Reprodução EfficientNet	Mobile NetV2	Shuffle Net	Google LeNet	Ensemble MA Soft (PACIENTE)
CIN	0,890	0,80	0,80	0,81	0,79	0,82
EBV	0,764	0,94	0,95	0,89	0,87	0,94
GS	0,897	0,71	0,72	0,70	0,74	0,71
MSI	0,898	0,82	0,86	0,81	0,84	0,86
Macro AVG		0,83	0,85	0,83	0,83	0,85

Fonte: O autor (2025).

Em comparação com Wang et al. (2022), observamos ganhos de +32–34 pontos em recall para EBV no nível de tiles e melhora substancial em PPV no nível de pacientes, atingindo precisão perfeita (1,00) para EBV e MSI. Esses resultados contrastam com Jeong et al. (2022), que reportaram recall elevado mas precisão mais baixa, e complementam Zheng et al. (2022), que demonstraram aumento de robustez ao integrar CNNs e patologistas. Achados semelhantes foram descritos por Flinner et al. (2022) e sintetizados na

revisão sistemática de Cifici et al. (2022), que destaca a necessidade de validação externa. Em linha com Zhou et al. (2023), nossos resultados reforçam o potencial translacional da Inteligência Artificial na patologia digital. Assim, os resultados não apenas confirmam avanços recentes na literatura, mas evidenciam o potencial translacional dos ensembles como ferramenta de apoio à prática em patologia digital, aproximando a classificação molecular baseada em imagens da acurácia obtida por métodos genômicos mais custosos.

3.8 CONCLUSÃO

O presente estudo demonstrou que o G.SubtVision, um modelo de ensemble Soft com MobileNetV2, ShuffleNet e GoogLeNet, melhorou significativamente a predição dos subtipos moleculares do adenocarcinoma gástrico (CIN, MSI, EBV e GS) a partir de imagens histopatológicas. Para uma compreensão mais profunda do desempenho do modelo, incluindo uma discussão sobre suas limitações, a contribuição individual de cada arquitetura no ensemble e perspectivas futuras, convidamos o leitor a consultar o Material Suplementar. Assim, este trabalho não apenas confirma avanços recentes na literatura, mas também aprimora o potencial translacional das CNNs como ferramenta acessível de apoio ao diagnóstico.

4 CAPÍTULO 2: REDES NEURAIS CONVOLUCIONAIS CLASSIFICAM SUBTIPO MOLECULAR DO CÂNCER GÁSTRICO EM DATASET TUBULAR-CONTROLADO

RESUMO

Uma abordagem promissora em desenvolvimento para a classificação dos subtipos moleculares do câncer gástrico é o treinamento de redes neurais convolucionais (CNN) em imagens histopatológicas supervisionadas por rótulos moleculares. Essa supervisão molecular pode estar identificando novos atributos por aprendizado profundo (deep learning). A evidência atual, no entanto, ainda não demonstra de forma conclusiva a descoberta de atributos inéditos. Metodologia: O presente estudo utilizou dados TCGA-STAD (the cancer genomic atlas - stomach adenocarcinoma) e organizou um novo conjunto de dados (dataset), apenas com tipos histológicos tubulares (WHO-2019), denominado dataset tubular-controlado (22 casos de tipo tubular categorizados como CIN ou não-CIN) e outro conjunto denominado dataset geral (263 casos dos 4 subtipos - CIN, MSI, GS e EBV). MobilleNet- V2 foi treinada em ambos os datasets e os resultados foram contrastados. Adicionalmente foram treinadas apenas no dataset tubular-controlado outras 5 redes: VGG19, DenseNet, ResNet50-v2, Inception-v3 e NASNet-Mobile). Diversas redes obtiveram resultados significativos. A NASNet-Mobile apresentou o melhor desempenho global (AUROC >0,72). O desempenho da MobileNetV2 no dataset tubular controlado para o subtipo CIN teve precisão, recall, F1-score a AUC-ROC respectivamente de 0.62/ 0.73/ 0.66/ 0.64 enquanto no dataset geral a mesma rede obteve 0.63/0.69/0.66/0.69. Concluiu-se, ao contrastar esses resultados, que a predição do subtipo molecular instabilidade cromossômica CIN em adenocarcinomas gástricos por CNN persiste no dataset tubular-controlado, reforçando o papel das CNN em identificar fenótipos profundos.

4.1 INTRODUÇÃO

O câncer gástrico permanece como uma das principais causas de mortalidade por câncer em todo o mundo, apresentando elevada heterogeneidade clínica, histológica e molecular. Em 2014, o consórcio The Cancer Genome Atlas (TCGA) propôs uma classificação molecular que subdivide os adenocarcinomas gástricos em quatro grupos principais: associado ao vírus Epstein-Barr (EBV), instável por microssatélites (MSI), instabilidade cromossômica (CIN) e genômica estável (GS). Essa estratificação revela subgrupos com características genéticas e fenotípicas distintas, com impacto direto no prognóstico e nas opções terapêuticas. Entre esses subtipos, a instabilidade cromossômica (CIN) destaca-se como o subgrupo mais prevalente, associado a alterações cromossômicas extensas, aneuploidia e padrões clínicos específicos (Cancer Genome Atlas Research Network, 2014).

Tradicionalmente, a prática diagnóstica em patologia baseia-se na supervisão de especialistas, utilizando critérios morfológicos vistos à microscopia óptica, como o sistema de Lauren (intestinal, difuso, misto) uma classificação clássica de 1965 Correção sugerida: que diferencia o câncer gástrico em tipos Intestinal, Difuso e Misto (LAURÉN, 1965) ou a classificação da OMS-2019 (que classifica conforme a morfologia os câncer gástricos em: Papilífero, Tubular (bem diferenciado, moderadamente diferenciado e mal diferenciado), Pouco Coeso (Anel de sinete ou não-anel de sinete), Mucinoso, Misto, Adenoescamoso, Carcinoma de células escamosas, Carcinoma indiferenciado, Carcinoma de estroma linfóide, Adenocarcinoma hepatóide, com diferenciação enteroblástica, tipo glándula fúndica e micropapilar. Essas classificações refletem décadas de conhecimento acumulado e aumento de sua complexidade buscando categorias mais histomorfológicas cada vez mais específicas.

Os avanços dos métodos de sequenciamento e da compreensão dos processos carcinogênicos com o desenvolvimento de tratamentos específicos têm potencializado o avanço da classificação de cânceres de diversas topografias, como por exemplo o de mama e da próstata, com tipos imunofenotípicos bem estabelecidos. No câncer gástrico, a abordagem molecular ainda não está bem estabelecida. A classificação molecular foi proposta pelo TCGA em 2014, mas ainda há limitações de acessibilidade aos métodos multiômicos nos quais ela foi inicialmente identificada com tentativas ainda imaturas de estabelecer painéis de imuno-histoquímica (KIM et al., 2016; FUKAYAMA; RUGGE; WASHINGTON, 2019).

Recentemente, modelos de aprendizado profundo têm demonstrado a capacidade de

predizer subtipos moleculares diretamente a partir de imagens histopatológicas coradas em Hematoxilina e Eosina (H&E). Treinando CNN´s por supervisão de rótulos moleculares. Estudos pioneiros como Kather et al. (2019) demonstraram a capacidade do deep learning em predizer MSI diretamente de imagens histológicas. Nesse contexto, trabalhos como os de Wang et al. (2022), Flinner et al. (2022) mostraram que redes neurais convolucionais (CNNs) podem alcançar desempenhos robustos na classificação de subtipos do TCGA. (KATHER et al., 2020; WANG et al., 2022; FLINNER et al., 2022)

O uso de aprendizado profundo em patologia digital pode ser agrupado em duas grandes vertentes: a supervisão de especialistas, em que os rótulos são definidos por patologistas a partir de critérios morfológicos convencionais, e a supervisão molecular, em que os rótulos derivam de dados genômicos ou biomarcadores independentes da morfologia. Esta distinção é fundamental para compreender tanto os avanços recentes quanto as lacunas que ainda persistem.

Na primeira vertente, os modelos buscam replicar ou ampliar classificações já estabelecidas por patologistas. Jang et al. (2021) demonstraram que uma CNN Inception-v3 foi capaz de distinguir adenocarcinomas gástricos diferenciados vs. indiferenciados e mucinosos vs. não-mucinosos, alcançando AUCs-ROC de 0,932 e 0,979, respectivamente, em nível de patch. O estudo reforça que a inteligência artificial pode reduzir a subjetividade e acelerar tarefas que já fazem parte da rotina diagnóstica. De forma semelhante, Kanavati & Tsuneki (2021) avaliaram o desempenho de CNNs na classificação do adenocarcinoma difuso (tipo Lauren), utilizando mais de 2.900 biópsias de múltiplos hospitais japoneses. Os modelos atingiram AUCs-ROC próximos de 0,95–0,99 em diferentes coortes, mostrando que a IA pode capturar padrões histológicos que patologistas já reconhecem, mas com maior rapidez e reprodutibilidade. Em comum, esses trabalhos utilizam abordagem de treinamento supervisionado dependente dos rótulos atribuídos por especialistas tomando-os como verdade de base (ground truth) (JANG; SONG; LEE, 2021)

Na segunda vertente, emergem os estudos que classificam imagens de H&E em subtipos moleculares. Wang et al. (2022) introduziram o método para predição dos quatro subtipos do TCGA (CIN, MSI, EBV, GS). Já Flinner et al. (2022) aplicaram deep learning nos quatro subtipos do TCGA e compararam a testes moleculares independentes e imunohistoquímica. Essas abordagens têm em comum o treinamento supervisionado tomando dados moleculares como ground truth. Esses dados moleculares devem ser alcançados por meios não operador-dependente, através de sequenciamento ou sondas, auxiliadas

métodos de bioinformática.

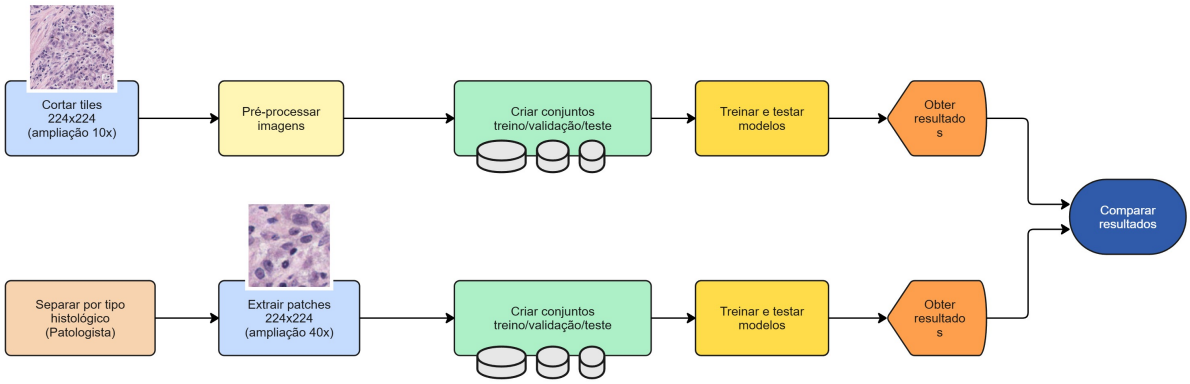
A ideia central deste estudo é fazer uso das redes neurais computacionais como instrumento de descoberta de fenótipos previamente desconhecidos associados ao genótipo do rótulo molecular. Por serem descobertos por métodos de aprendizado profundo, podem ser de maneira geral chamados de atributos profundos (deep features) e, quando especificamente referentes a genótipos de rótulos moleculares, são por definição fenótipos profundos (deep phenotypes). Apesar dos avanços, a literatura de supervisão molecular ainda sofre de uma limitação central: a ausência de controles estritos de morfologia. Não está ainda comprovado se as CNNs estão de fato aprendendo atributos profundos subjacentes decorrentes de genótipos (fenótipos profundos) ou apenas reproduzindo associações com tipologias já conhecidas (como tubular vs. papilífero). Até o momento nenhum estudo isolou um único subtipo morfológico e testou se o desempenho se mantém. (FLINNER et al., 2022; WANG et al., 2022)

No presente trabalho se avalia se esses modelos de fato aprendem atributos não previamente descritos de genótipos, ou apenas captam associações com tipologias morfológicas já estabelecidas. Caso a precisão dos modelos possa ser explicada por associações com tipos histopatológicos, o potencial da supervisão de rótulos moleculares seria apenas a automatização de classificações já disponíveis. Por outro lado, caso haja identificação de padrões fenótipos previamente desconhecidos se apoia o potencial das CNN como ferramenta de investigação científica no tema. O presente trabalho busca preencher essa lacuna, avaliando se a predição por CNNs persiste em um conjunto tubular-controlado, no qual a tipologia histológica é mantida constante segundo a classificação OMS-2019.

4.2 METODOLOGIA

Essa seção descreve a metodologia adotada nos experimentos realizados. Inicialmente, foi feita a organização e descrição dos conjuntos de dados utilizados, desde as imagens de lâminas inteiras até a construção dos datasets específicos (tubular-controlado e geral), bem como a divisão em grupos de treinamento, validação e teste. Em seguida, são detalhadas a arquitetura de redes e o processo de treinamento, assim como as métricas de avaliação aplicadas. Por fim, são discutidos os aspectos relacionados à ética, reprodutibilidade e disponibilidade dos dados e modelos. A Figura 1 apresenta um fluxograma resumindo o pipeline metodológico.

Figura 1 – Fluxograma das abordagens do dataset tubular controlado e geral



Fonte: O autor (2025).

4.2.1 *Dataset*: Conjunto de Dados

O estudo foi realizado utilizando o projeto STAD (Stomach Adenocarcinoma) da base pública do TCGA (The Cancer Genomic Atlas) O conjunto de dados de imagens histopatológicas e rótulos dos subtipos moleculares foi extraído do projeto STAD (Stomach Adenocarcinoma), disponível no banco de dados público TCGA (The Cancer Genomic Atlas) (The Cancer Genome Atlas Research Network, 2014).

4.2.1.1 Imagens de lâminas inteiras

O TCGA disponibiliza imagens de lâminas inteiras (WSI) coradas em hematoxilina e eosina (H&E) em formato SVS de alta qualidade produzidas por patologia digital a 40x associadas aos subtipos moleculares CIN, EBV, MSI e GS.

4.2.1.2 Tipos Histopatológicos

A classificação histopatológica disponível na base do TCGA é a classificação de Lauren, Para os objetivos do presente estudo foi utilizada a classificação da OMS de 2019.

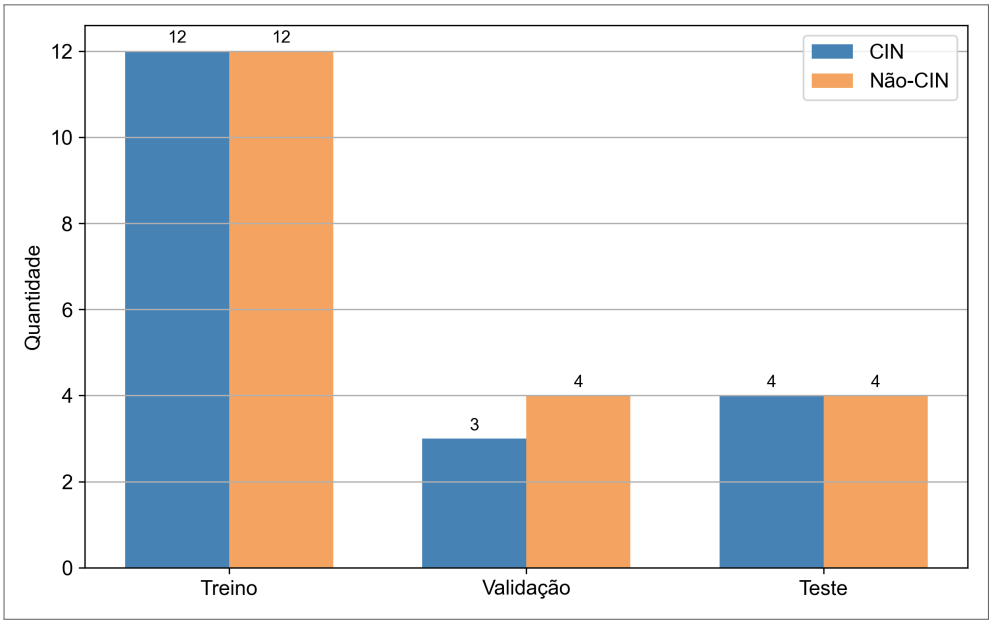
4.2.1.3 Construção do dataset tubular-controlado

95 casos do TCGA foram examinados por médicos especialistas em patologia com mais de 10 anos de atuação diagnóstica que classificaram os casos segundo a classificação da OMS 2019 e descreveram detalhes das características particulares da morfologia. Foram selecionados apenas casos classificados como Tubulares. Foram incluídos 23 casos representados por 37 imagens de lâmina inteiras. As imagens foram agrupadas por 3 subtipos de tubulares: bem (2), moderadamente (32) e mal diferenciados (5). Devido ao relativamente pequeno número de casos e da diminuta presença das classes minoritárias escolheu-se tratar o problema de maneira binária entre as classes majoritárias. O dataset foi por isso categorizado para CIN ou não-CIN (MSI) O pré-processamento das imagens compreendeu duas etapas principais: (I) segmentação das whole slide images (WSIs) em patches de 224×224 px representando apenas áreas cancerígenas tendo sido orientado por médicos patologistas experientes; e (II) normalização de cor aplicando o método Macenko (MACENKO et al., 2009).

4.2.1.4 Grupos treinamento, validação e teste dos experimentos com o dataset tubular-controlado

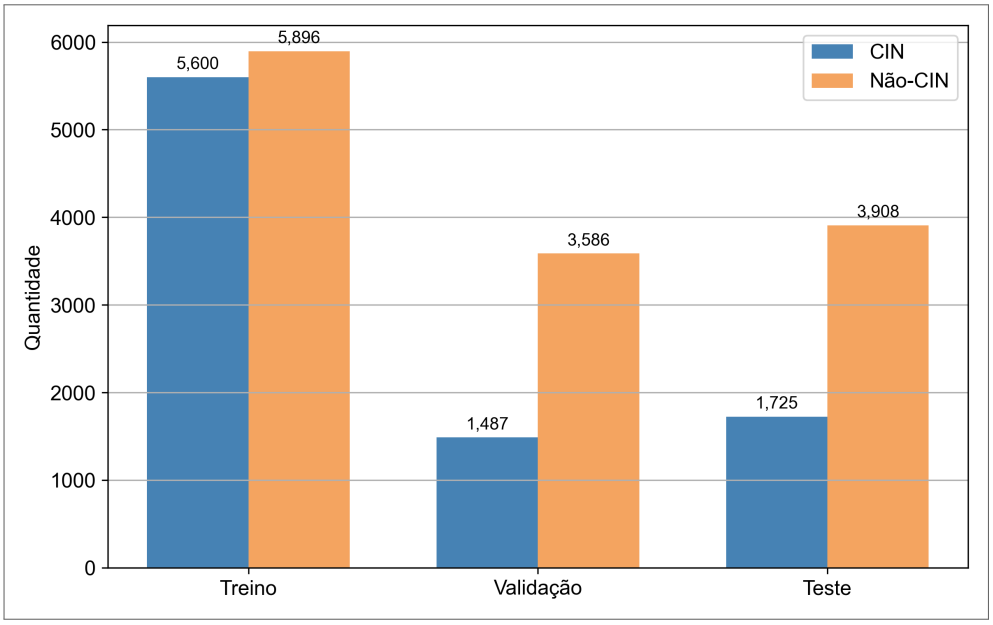
Os grupos treinamento, validação e teste foram construídos de maneira a ter uma distribuição semelhante entre tubulares bem, moderadamente e mal diferenciados em nossos grupos de treinamento, validação e teste. A separação foi feita por casos, não permitindo contaminação de patches entre os grupos. Para a classe CIN, as imagens foram distribuídas em 12 para treino, 3 para validação e 4 para teste. Para a classe não-CIN as imagens foram distribuídas em 12 para treino, 4 para validação e 4 para teste Figura 2. Ao se extrair o patches de cada imagem obteve-se na classe CIN uma distribuição de 5600 patches para treino, 1487 para validação e 1725 para teste. Enquanto que na classe não-CIN obteve-se uma distribuição de 5896 patches para treino, 3586 patches para validação e 3908 patches para teste Figura 3.

Figura 2 – Distribuição de Imagens por Classe e Conjunto



Fonte: O autor

Figura 3 – Distribuição de Patches por Classe e Conjunto



Fonte: O autor

4.2.1.5 Construção do dataset geral

Foram utilizadas 476 imagens de lâminas inteiras (WSI) coradas com Hematoxilina e Eosina (H&E), com resolução de 40x. Como o número de imagens é relativamente bem maior essas imagens foram categorizadas nos quatro subtipos moleculares: CIN (232), MSI (114), GS (73) e EBV (57). Durante o processo de criação do dataset geral para o

modelo de análise foram aplicadas três etapas principais de pré-processamento, a saber: Segmentação em Tiles - as 476 imagens WSI foram divididas em pequenas imagens quadradas de 224x224 pixels (tiles). O corte foi realizado com um fator de ampliação de 10x. Durante este processo, tiles com menos de 50% de presença de tecidos foram removidos. Ao final, cerca de 1.500.000 tiles foram gerados, distribuídos entre os subtipos moleculares. Por fim, aplicou-se uma normalização de cor, um template de referência foi utilizado para ajustar o espaço de cor e a luminosidade dos tiles, utilizando o método de Macenko (MACENKO et al., 2009)

4.2.1.6 Grupos treinamento, validação e teste dos experimentos com dataset geral

Nesta etapa, aplicou-se divisão aleatória de dados para treinar, validar e testar os modelos. Primeiro, divisão inicial dos dados para os grupos Treino (241 casos) (950.434 tiles) correspondendo a 72% do dataset, Validação (153 casos) (105.603 tiles), 8% do dataset e grupo Teste Hold-Out com 82 casos (179.952 tiles), 20% do dataset. Medidas foram tomadas para não permitir contaminação entre os tiles de um mesmo paciente entre os grupos.

4.2.2 Arquitetura e treinamento

Foram treinadas no dataset tubular-controlado 6 arquiteturas de redes neurais convolucionais (CNNs): MobileNet-V2, VGG19, DenseNet, ResNet50-v2, 2, Inception-v3 e NASNet-Mobile. No dataset geral foi treinada MobileNet-V2. Todas as redes tiveram pesos inicializados a partir do ImageNet (DENG et al., 2009) Os modelos foram treinados com dois valores distintos de learning rate (1e-4 e 1e-3), a fim de avaliar a estabilidade e sensibilidade ao parâmetro. Utilizou-se o otimizador Adam (KINGMA; BA, 2015), com weight decay e estratégia de redução do learning rate on plateau. O treinamento foi realizado em mini-batches de 32 imagens, com early stopping monitorando a perda de validação, interrompendo o processo quando não havia melhora após 15 épocas consecutivas. (PRECHELT, 1997)

4.2.3 Métricas utilizadas

As métricas empregadas foram: Precisão (valor preditivo positivo, PPV): métrica prioritária, refletindo a proporção de predições positivas corretas. A ênfase no PPV se justifica pelo contexto clínico de diagnóstico diferencial, no qual falsos positivos podem levar a condutas terapêuticas inadequadas e risco direto ao paciente. Essa escolha está alinhada às recomendações das diretrizes STARD 2015 para estudos de acurácia diagnóstica. Recall (sensibilidade): proporção de verdadeiros positivos corretamente identificados, importante para mensurar a capacidade de detecção de casos. F1-score: média harmônica entre precisão e recall, avaliando o equilíbrio entre ambas. Área sob a curva ROC (AUROC): métrica global de discriminação avaliando confiança do modelo nas predições.

4.2.4 Ética, reprodutibilidade e disponibilidade

Os dados utilizados neste estudo são provenientes do The Cancer Genome Atlas – Stomach Adenocarcinoma (TCGA-STAD), um repositório público e de acesso aberto, disponível no Genomic Data Commons. Por se tratar de dados previamente coletados, anonimizados e disponibilizados em domínio público, não se faz necessário a submissão ao comitê de ética local, em conformidade com as diretrizes internacionais para o uso secundário de dados públicos.

4.3 RESULTADOS E DISCUSSÃO

Resultados no Dataset Tubular-Controlado

No conjunto tubular-controlado, composto exclusivamente por adenocarcinomas gástricos tubulares reclassificados de acordo com a OMS-2019 por médicos especialistas em patologia com mais de 10 anos de atuação., As redes neurais convolucionais (CNNs) demonstraram capacidade de predizer o subtipo molecular CIN de forma consistente e estatisticamente superior ao acaso.A Tabela 1 apresenta os resultados para as seis arquiteturas avaliadas (MobileNetV2, VGG19, DenseNet, ResNet50-v2, Inception-v3 e NASNet-Mobile), treinadas com dois valores de taxa de aprendizado (learning rate, lr: 1e-4 e 1e-3). As métricas incluem acurácia (acc), F1-score, precisão (PPV), recall e área sob a curva ROC

(AUROC). Observa-se que o NASNet-Mobile obteve o melhor desempenho global com lr = 0,001, alcançando F1 = 0,71, precisão = 0,72, recall = 0,71 e AUROC = 0,73. Essa arquitetura destacou-se pela robustez, com valores médios de AUROC > 0,70, confirmando sua eficiência em cenários restritos. As demais redes apresentaram F1-scores variando de 0,59 a 0,70 para lr = 0,0001, e de 0,61 a 0,71 para lr = 0,001, com AUROC consistentemente acima de 0,60 em todos os casos. A macro-F1 média no dataset tubular-controlado foi de 0,47–0,49, indicando que, mesmo sob controle morfológico estrito, os modelos mantêm discriminação molecular para CIN. Notavelmente, a taxa de aprendizado mais alta (lr = 0,001) tende a melhorar o recall em várias arquiteturas, como na MobileNetV2 (recall = 1,00), embora às custas de uma leve redução na precisão em alguns casos.

Tabela 1 – Resultados CNN´s no dataset tubular controlado.

Arquitetura	lr	Acurácia (%)	F1-Score	Precisão	Recall	AUROC
MobileNetV2	0,0001	0,59	0,66	0,62	0,73	0,64
	0,001	**0,56**	**0,71**	**0,56**	**1,00**	**0,56**
VGG19	**0,0001**	**0,67**	**0,70**	**0,69**	**0,73**	**0,69**
	0,001	0,61	0,66	0,64	0,71	0,64
DenseNet	**0,0001**	**0,65**	**0,68**	**0,68**	**0,69**	**0,66**
	0,001	0,54	0,64	0,56	0,77	0,61
ResNet50-v2	**0,0001**	**0,65**	**0,69**	**0,67**	**0,71**	**0,67**
	0,001	0,60	0,61	0,67	0,57	0,63
Inception-v3	0,0001	0,60	0,59	0,68	0,53	0,62
	0,001	**0,64**	**0,67**	**0,69**	**0,66**	**0,68**
NASNet-Mobile	0,0001	0,65	0,67	0,70	0,66	0,67
	0,001	**0,69**	**0,71**	**0,72**	**0,71**	**0,73**

Resultados no dataset tubular-controlado. As métricas são calculadas em nível de tile, priorizando a precisão (PPV) conforme o contexto clínico de diagnóstico diferencial. Valores em negrito indicam o melhor desempenho por arquitetura para os valores de learning rate (1e-4 e 1e-3) avaliados.

4.3.1 Resultados no Dataset Geral

Os resultados da MobileNetV2 no dataset geral são visualizados na Tabela 2. Essa detalha as métricas por subtipo molecular (CIN, EBV, GS, MSI), com precisão, recall, F1-

score. No subtipo CIN (classe majoritária), a precisão foi de 0,63, recall de 0,69 e F1 de 0,66, com suporte de 89.707 tiles. Para classes minoritárias, os valores foram mais modestos: EBV (precisão 0,49, recall 0,33, F1 0,39), GS (precisão 0,29, recall 0,44, F1 0,35) e MSI (precisão 0,50, recall 0,32, F1 0,39).

Tabela 2 – Resultados da MobileNetV2 no dataset geral.

Classe	Precisão	Recall	F1-Score	AUC-ROC
CIN	0,63	0,69	0,66	0,69
EBV	0,49	0,33	0,39	0,77
GS	0,29	0,44	0,35	0,67
MSI	0,50	0,32	0,39	0,69
Macro AVG	0,48	0,44	0,45	
Weighted AVG	0,53	0,52	0,52	
Micro AVG				0,76

Métricas calculadas em nível de tile, com ênfase na precisão para o subtipo CIN (0,63). O suporte reflete a distribuição desbalanceada, com CIN como classe dominante.

Em resumo, o NASNet-Mobile apresentou o melhor desempenho no dataset controlado, destacando-se na identificação de CIN em adenocarcinomas tubulares (F1 médio > 0,70) mas várias redes tiveram um desempenho acima do aleatório. Isso demonstra que as CNNs capturam padrões histopatológicos não previamente conhecidos, rejeitando a hipótese nula e confirmando a utilidade da supervisão molecular para revelar fenótipos profundos.

Os resultados da MobileNet-V2 no dataset tubular-controlado e no dataset geral para o subtipo CIN foram muito aproximados. Há uma evidente limitação metodológica ao comparar datasets diferentes, sendo um deles muito maior e outro categorizado para um problema binário e, portanto, não é possível afirmar que os resultados foram iguais do ponto de vista estatístico, pois não é o mesmo grupo de teste. No entanto, não seria necessário provar que ambos são idênticos para demonstrar que a CNN está identificando um atributo profundo que vai além do tipo tubular, caso o desempenho da CNN fosse acima do aleatório ambora inferior no dataset tubular-controlado em relação ao grupo controle do dataset geral já estaria demonstração que há a participação parcial de um atributo profundo. O contraste de resultados, ainda mais, foi surpreendente por sua proximidade: O desempenho da MobileNet-V2 no dataset tubular controlado para o subtipo CIN teve precisão, sensibilidade (Recall), F1-score e AUC-ROC respectivamente de 0.62/ 0.73/ 0.66/

0.64 enquanto no dataset geral a mesma rede obteve 0.63/ 0.69/ 0.66/ 0.69. A diferença de precisão foi de 1 ponto percentual, e o F1-Score foi idêntico. Indicando uma diferença muito pequena de falsos positivos.

Os autores compreendem a limitação metodológica de comparar os resultados de datasets diferentes, essa porém é o núcleo da abordagem que precisará ser metodologicamente aprimorada. Defendem, no entanto, o potencial explicativo da abordagem que permite testar se de fato a CNN está classificando atributos profundos (deep features) ao organizar um dataset controlado para os atributos conhecidos.

É uma contribuição médica ao campo, já que os especialistas podem desafiar o poder de predição da CNN ao organizar um dataset específico para testar uma hipótese. Um próximo passo no desenvolvimento da abordagem é ajustar o treinamento para poder rodar o modelo treinado no dataset tubular-controlado no grupo teste do dataset geral, assim podendo comparar estatisticamente os resultados dos modelos já que o grupo teste seria então o mesmo.

4.4 CONCLUSÃO

Este estudo aponta para que redes neurais convolucionais (CNNs) são capazes de predizer a instabilidade cromossômica (CIN) em adenocarcinomas gástricos no dataset tubular-controlado (composto exclusivamente por tumores tubulares). A manutenção de desempenho acima do acaso e aproximada (por contraste) com o grupo controle do dataset geral sugere que esses modelos identificam padrões histomorfológicos subjacentes ao subtipo molecular e não associações com a classificação histopatológica WHO 2019.

5 CAPÍTULO 3: G.SUBTFOREST – CLASSIFICADOR DE SUBTIPOS MOLECULARES DO CA GÁSTRICO COM TCGA VIA RANDOM FOREST EM PAINÉIS OTIMIZADOS

RESUMO

A aplicação da classificação molecular do adenocarcinoma gástrico permanece um desafio. Este estudo apresenta os G.SubtForest (Gastro Subtyping Trough Random Forest) classificadores com base em painéis de mutação para subtipos moleculares. Dois paineis são aqui propostos com 18 e 9 genes respectivamente. Metodologia: A partir de 18.600 variantes de nucleotídeo (SNV) somáticas não-sinônimas da base TCGA-STAD (The Cancer Genomic Atlas - Stomach Adenocarcinoma) foram organizados 10 grupos treinamento e validação utilizando K-fold (k=10) foram então treinados modelos de *Random Forest* e utilizado *SHapley Additive exPlanations* (SHAP) para identificar os genes de maior influência colaborativa nas predições. Os resultados dos 10 modelos foram consolidados em dois painéis otimizados: um com 18 genes, adequado ao sequenciamento de nova geração, e outro com 9 genes, apropriado para imuno-histoquímica. Novos modelos G.SubtForest 18 e G.SubtForest 9 foram treinados para classificação de casos a partir da informação da mutação em cada um dos paineis. Os G.subtForest mostraram desempenho consistente (AUC-ROC avg 0,91 e 0,89, respectivamente). Os resultados evidenciam ganhos relevantes na estratificação de pacientes e oferecem solução reprodutível e escalável para uso translacional. Código e material suplementar disponíveis.

5.1 INTRODUÇÃO

O adenocarcinoma gástrico representa uma das principais causas de mortalidade por câncer no mundo, com heterogeneidade histológica e molecular que complica o diagnóstico e tratamento (Cancer Genome Atlas Research Network, 2014). A classificação molecular, proposta pelo *The Cancer Genome Atlas* no projeto *Stomach Adenocarcinome* (TCGA-STAD), descobriu quatro *clusters* de dados multiômicos denominados subtipos moleculares: positivo para vírus Epstein-Barr (EBV), instabilidade de microssatélites (MSI), genomicamente estável (GS) e instabilidade cromossômica (CIN) (Cancer Genome Atlas Research Network, 2014).

Apesar dos avanços científicos, o avanço do diagnóstico de rotina enfrenta lacunas significativas que não capturam a heterogeneidade molecular, contribuindo para altas taxas de recidiva(KIM et al., 2016).

A classificação da Organização Mundial da Saúde (WHO) para o câncer gástrico é baseada em apenas em aspectos histomorfológicos. Depois de descrever toda a classificação dos tipos histopatológicos cita a classificação molecular apenas no tópico sobre prognóstico, descrevendo apenas o que é frequente ou não em cada subtipo molecular. Embora a classificação da (WHO) represente um marco na padronização do diagnóstico histopatológico do câncer gástrico, seu escopo permanece centrado em critérios morfológicos e não incorpora, de forma sistemática, informações moleculares ou genômicas.

Essa lacuna limita a capacidade de correlacionar padrões histológicos com processos carcinogênicos mais precisos, restringindo o potencial de estratificação prognóstica e preditiva. A crescente disponibilidade de dados multiômicos e o avanço da bioinformática, indicam caminhos para futuras revisões das diretrizes, capazes de integrar morfologia, marcadores imuno-histoquímicos e assinaturas genéticas. Tal abordagem ampliará a utilidade clínica das classificações, permitindo diagnósticos mais alinhados com a definição de terapias-alvo e maior alinhamento com a medicina de precisão(FUKAYAMA; RUGGE; WASHINGTON, 2019) O presente estudo busca contribuir nessa direção.

Trabalhos propuseram painéis imuno-histoquímicos (IHC) para subtipos moleculares do adenocarcinoma gástrico, como o de Kim et al., que propôs MLH1, PMS2, MSH2, MSH6, HER2, EGFR, MET, PTEN e P53 e ISH para EBV em 438 pacientes. encontrando apenas 14 EBV,(3,3%); 21 MSI (4,8%); (associando-o a deficiência nas proteínas de reparo de mismatch mmr - MLH1, PMS2, MSH2, MSH6). 218 (49,8%) sobreexpressão de RTKs

(HER2,EGFR,MET) e em 258 (59,1%) p53 overexpressed/null foi identificada uma alta prevalência de mutações ou inativação proteica, associando ambos os fenótipos a CIN. No entanto, há grandes limitações metodológicas nas associações entre os fenótipos imunohistoquímicos e os subtipos moleculares feitos pelos autores. Eles o fazem apenas por inferência indutiva, argumentando razoabilidade apenas por associações probabilísticas com o que foi publicado em artigo (The Cancer Genome Atlas Research Network, 2014) sem verificar nos dados originais suas suposições ou utilizar os mesmos critérios diagnósticos do artigo original (sequenciamento de exoma e metiloma).

Já Flinner et al (FLINNER et al., 2022) apresentaram um modelo de classificação de subtipos moleculares com IA em imagens histopatológicas do TCGA e compararam os resultados desse modelo aos dos marcadores imunohistoquímicos propostos por Kin et al em um grupo com controle externo por análise de variação de número de cópias (copy number variation) encontrando que o modelo de IA em imagens foi melhor que o painel previamente proposto para subtipo CIN. Por outro lado, Wang et al. (2022) desenvolveram um modelo de rede neural convolucional usando TCGA como fonte de imagens histopatológicas e rótulos derivados de sequenciamento multiômico, oferecendo uma base molecular robusta para os subtipos EBV, MSI, GS e CIN.

Paineis genéticos são de suma importância para a classificação de tumores, como já ocorre em outros tumores como mama e próstata. Esses paineis devem somar informações significativas aos achados histomorfológicos. Dependendo do tamanho do painel, métodos variam: IHC para painéis pequenos e sequenciamento de próxima geração (NGS) para painéis maiores, mas problemas como cobertura inadequada em blocos de parafina fixados em formalina (FFPE) de painéis grandes e inadequação de acesso e aumento do custo saúde persistem, limitando a translação para a rotina de painéis grandes (KIN et al., 2016; FUKAYAMA; RUGGE; WASHINGTON, 2019).

O aprendizado de máquina tem grande potencial na descoberta de mutações chave para diagnóstico diferencial de câncer, com algoritmos como Random Forest (RF) destacando-se por sua robustez em dados de alta dimensionalidade, redução de viés em classes desbalanceadas (ex.: EBV minoritário) e seleção de features via importância Gini, como demonstrado em estudos recentes para subtipos TCGA (XU et al., 2023). (JANG et al., 2023).

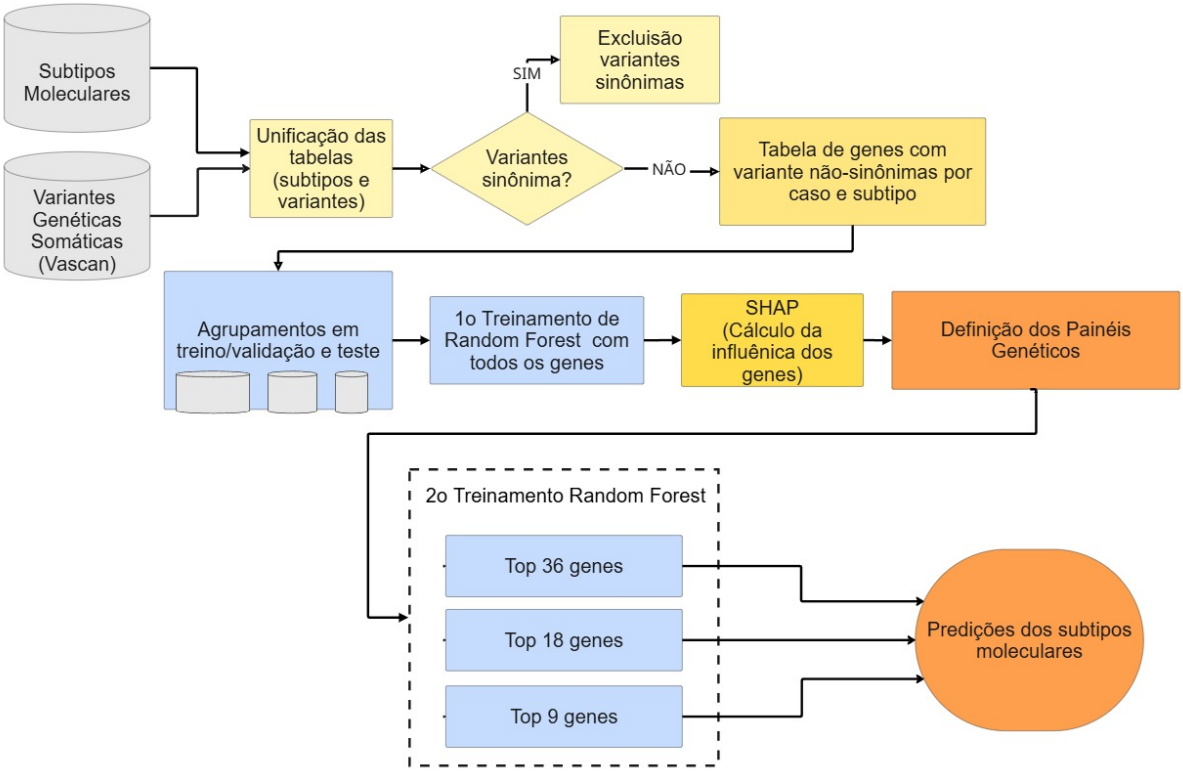
O objetivo deste estudo é desenvolver um sistema preditivo para subtipos moleculares do câncer gástrico otimizando Paineis Genéticos ao fazer uso de Random Forest em dados do TCGA-STAD, A importância deste estudo reside no avanço do campo, fornecendo uma

ferramenta bioinformática reprodutível (com código suplementar) para estratificação de pacientes, facilitando terapias direcionadas e reduzindo mortalidade, alinhado a abordagens multiômicas integradas (LIU et al., 2024).

5.2 METODOLOGIA

A metodologia desta etapa foi delineada para organizar o processo de aquisição, preparação e análise dos dados do SNV-TCGA STAD. Inicialmente, os dados foram coletados e estruturados, seguidos pela definição dos agrupamentos em treino, validação e teste. Na sequência, descreve-se o processo de treinamento dos modelos e a consolidação dos painéis obtidos. Por fim, são apresentadas as métricas utilizadas para avaliação dos resultados. A Figura 1 ilustra, em formato de fluxograma, o fluxo metodológico, desde a aquisição dos dados até a avaliação final.

Figura 1 – Fluxograma do pipeline da criação dos paineis genéticos.



Fonte: O autor (2025).

5.2.1 Aquisição de dados SNV-TCGA STAD

A detecção de mutações somáticas por variação de nucleótidos únicos (SNV) foi inicialmente realizada pelo pipeline do TCGA utilizando o software VarScan2 (v2.4.4), que compara as amostras tumorais ao DNA germinativo emparelhado para identificar mutações somáticas. Os arquivos de saída (geralmente .maf, .vcf ou .tsv) foram posteriormente anotados com ferramentas como SnpEff ou ANNOVAR, que classificam o tipo funcional de cada variante (por exemplo: sinônima, missense, nonsense, splicing, etc.) e atribuem a cada uma um gene correspondente. Para esta pesquisa, foram utilizadas apenas as variantes com impacto funcional não-sinônimo, com exclusão sistemática das mutações classificadas como "synonymous_variant", a fim de focar em alterações com consequências de alterações de tradução proteica. Após o processamento, as mutações foram agrupadas por gene e por subtipo molecular (EBV, MSI, GS, CIN), com o objetivo de determinar os genes mais frequentemente mutados em cada categoria.

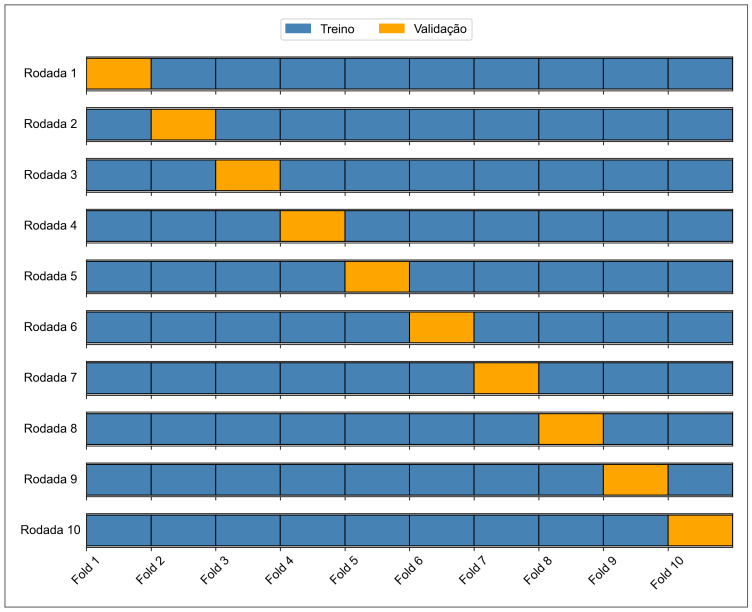
5.2.2 Agrupamentos: Treino, Validação e Teste.

Nesta etapa, foram criados aleatoriamente dois grupos de casos: Um grupo treino e validação com (290) casos e um grupo teste *hold out* com 81 casos oriundos de um dataset com montante de 443 casos.

5.2.2.1 Grupo Treinamento/Validação

Os grupos de treinamento e validação para cada modelo (explicados no tópico treinamento abaixo) foram separados utilizando o método *K fold*. Uma utilização parcial do *k-fold cross validation* com o objetivo, apenas, de gerar múltiplas separações aleatórias de conjuntos de treinamento e validação. Consequentemente, treinando 10 modelos (K=10), como pode ser obsevada pela Figura 2. A intenção do uso do método foi reduzir o viés em grupos de validação de classes minoritárias. As separações treino/validação foram 90/10.

Figura 2 – Exemplo de Validação Cruzada K-Fold (K=10)

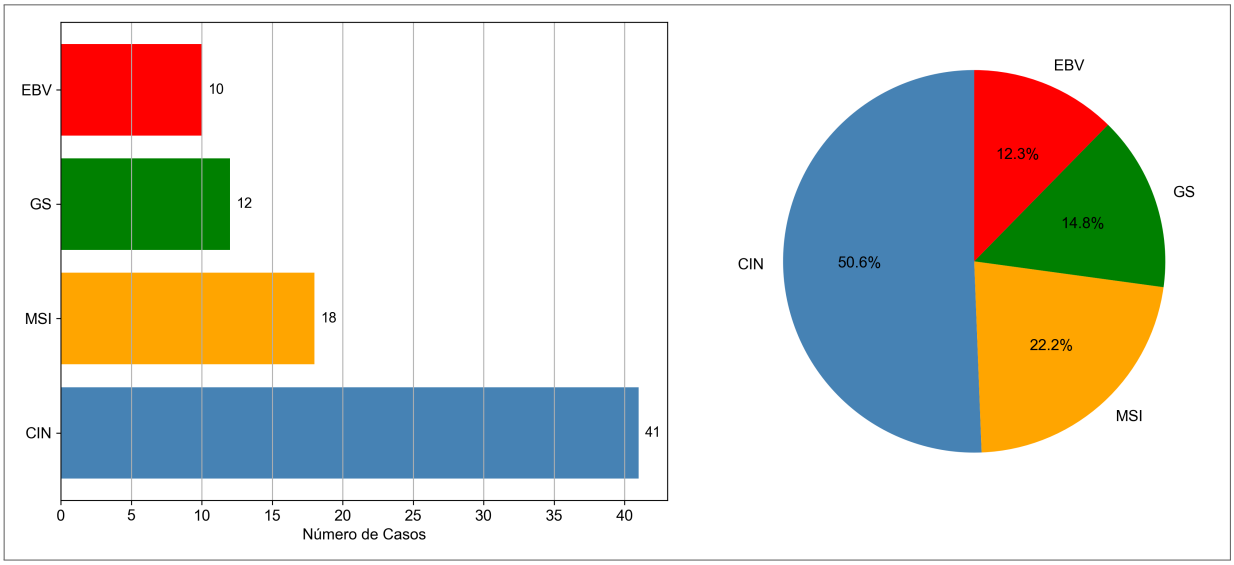


Fonte: O autor

5.2.2.2 Grupo Teste

81 casos foram separados para uso no teste final (*hold out*) tanto com imagens histopatológicas como com as informações sobre SNV dos 10 genes do painel de genes influentes. Medidas foram tomadas para evitar contaminação das imagens dos casos do conjunto teste com os outros conjuntos. Após análise da distribuição de casos no grupo teste, foram acrescentados aleatoriamente mais casos das classes minoritárias. Totalizando 81 casos. Distribuídos 41 CIN, 18 MS, 12 GS e 10 EBV, vide Figura 3.

Figura 3 – Distribuição de Casos no Grupo Teste (Hold-out)



Fonte: O autor

5.2.3 Treinamento

Foi empregado o algoritmo *Random Forest* (implementação RandomForestClassifier), configurado com `random_state=42`, execução paralela (`n_jobs=-1`) e balanceamento automático das classes (`class_weight='balanced'`). A etapa de ajuste de hiperparâmetros foi conduzida por meio do RandomizedSearchCV, definido com n_iter=10, validação cruzada de três partições (cv=3), processamento paralelo (`n_jobs=-1`) e `random_state=42`. O espaço de busca contemplou os seguintes hiperparâmetros: quantidade de árvores no ensemble (n_estimators $\in$ 100, 200, 300, 400, 500), número máximo de atributos considerados por divisão (max_features $\in$ 'sqrt', 'log2', None), profundidade máxima permitida (max_depth $\in$ None, 10, 20, 30), número mínimo de amostras exigido para realizar uma divisão (min_samples_split $\in$ 2, 5, 10), número mínimo de amostras por nó folha (min_samples_leaf $\in$ 1, 2, 4) e uso ou não do método bootstrap para amostragem (bootstrap $\in$ True, False).

10 modelos de *Random Forest* foram treinados (K-fold, k=10) nos dados tabulados SNV (Varscan) não-sinônimos contendo 18.600 genes para a tarefa de classificação dos 4 subtipos moleculares (denominada lista todos os genes).

Em cada modelo treinado em um dos 10 folds (explicados em agrupamentos) foi empregado o método SHAP (SHapley Additive exPlanations) para avaliação da influência de cada gene nas predições dos modelos. Para cada subtipo, foram selecionados os 10 genes

com maior magnitude de influência para a predição de cada modelo. 2.4 Consolidação das listas (SHAPLEY, 1953; CHEN et al., 2025)

As 40 listas de genes (10 para cada um dos 4 subtipos) foram consolidadas por diferentes metodologias explicadas a seguir.

Novos modelos de *Random Forest* (K-fold=10) foram treinados conforme cada painel consolidado, mantendo somente as informações relevantes aos genes presentes no painel em treinamento.

5.2.4 Consolidação dos painéis

Inicialmente, as 40 listas (maiores magnetudes de SHAP) foram consolidadas por votação simples (Hard Voting), compondo-se assim um painel dos 10 genes mais frequentes nas listas de influentes dos modelos. Denominado painel de 10 genes mais influentes por frequência de aparência nas listas.

Um segundo método utilizado para consolidar as listas da maneira a ter mais explicabilidade biológica foi primeiro excluir os genes das listas cujos SHAP eram negativos. O que significa que a sua influência se dá quando não está presente.

A partir das listas contendo somente SHAP positivo, os painéis foram consolidados por método de pontuação ponderada que considera a posição nos rankings e a frequência de aparição entre subtipos para 4 listas (denominadas listas 10 mais por subtipo).

Essas quatro listas foram então consolidadas em uma única lista de 36 genes ordenados por influência na predição dos quatro subtipos por método de pontuação ponderada (denominada Painel 36 mais influentes).

O Método do Cotovelo foi então utilizado para determinar o número ideal do ponto de vista de custo-efetividade de genes no painel final, identificando o ponto de corte onde a variância explicada se estabiliza, 18 genes (75% de variância e até o 5º de cada subtipo). A dimensão do painel foi também escolhida para ser apropriada à NGS (New Generation Sequencing) com alta profundidade de cobertura em bloco de parafina, maior que 500X.

Foi também selecionado um painel viável reduzido para ser apropriado à imuno-histoquímica, com 9 genes (Os 3 primeiros de cada subtipo, excluindo dois genes que não têm ainda anticorpos listados no genecard.org).(STELZER et al., 2016) Foram também treinados modelos de *Random Forest* para classificação a partir do painel imuno-histoquímico proposto (KIN et al., 2016)

5.2.5 Métricas

As métricas foram computadas com scikit-learn (v1.2.2), incluindo médias macro (não ponderadas) e ponderadas (ponderadas por classe). Relatórios por fold e ensemble, com curvas ROC visualizadas via TensorBoard. As métricas foram calculadas nos níveis de tiles e pacientes, incluindo: Precisão (eq 5.1) – Proporção de predições positivas corretas; Sensibilidade ou Recall (eq 5.2) – Proporção de positivos reais corretamente identificados; F1-Score (eq 5.3) –Média harmônica de precisão e recall; AUC-ROC: Área sob a curva ROC (one-vs-rest por classe)

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{5.1}$$

$$\text{Recall} = \frac{VP}{VP + FN} \tag{5.2}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}} \tag{5.3}$$

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(\text{FPR}) d(\text{FPR}) \tag{5.4}$$

onde:

- VP = Verdadeiros Positivos (*True Positives*)
- FP = Falsos Positivos (*False Positives*)
- FN = Falsos Negativos (*False Negatives*)
- VN = Verdadeiros Negativos (*True Negatives*)
- $\text{TPR} = \frac{\text{VP}}{\text{VP+FN}}$ (Taxa de Verdadeiros Positivos ou *Recall*)
- $\text{FPR} = \frac{\text{FP}}{\text{FP+VN}}$ (Taxa de Falsos Positivos)

A curva ROC é uma ferramenta gráfica utilizada para avaliar o desempenho de um modelo de classificação binária, representando o trade-off entre a taxa de verdadeiros positivos (Recall) (TPR) e a taxa de falsos positivos (FPR) à medida que a confiança do modelo na predição aumenta.

5.3 RESULTADOS E DISCUSSÃO

Primeiro são apresentados os painéis com os genes mais influentes. Na sequência, são apresentados os resultados do poder de predição por Random Forest, discutidas as diferentes métricas de precisão, recall, F1 score e AUC-ROC. Depois é apresentada as comparações das métricas entre entre os painéis imuno-histoquímico aqui proposto e o poder de predição do painel imunoistoquímico proposto por Kim et al 2016 na base do TCGA STAD.

5.3.1 Descrição dos Painéis Genéticos

A análise de frequência e importância resultou na definição dos seguintes painéis genéticos:

- **10 genes mais frequentes:** ARID1A, TP53, RNF213, MUC16, PIK3CA, KMT2D, HERC2, DOCK3, SYNE1, PCDHB13.
- **Painel TOP 36:** TP53, ARID1A, MUC16, ZBTB41, GGNBP2, PIK3CA, SYT17, MEF2C, MUC6, RNF213, SEC31A, BOC, CDH18, NFASC, BHLHB9, FAS, HERC2, SYNE1, ATM, CHD1, GRIP1, PCDHA2, PRCC, GJD4, KMT2D, DOCK3, KDM2B, KIF21A, SDR9C7, CD14, CTNNBL1, DYSF, XKR6, GLIS2, MYO15A, PTPN14.

  Nota: No material suplementar estão disponíveis os pesos ponderados e a influência na predição de cada subtipo dos 36 genes.
- **Painel TOP 18:** TP53, ARID1A, MUC16, ZBTB41, GGNBP2, PIK3CA, SYT17, MEF2C, MUC6, RNF213, SEC31A, BOC, CDH18, NFASC, BHLHB9, FAS, HERC2, SYNE1.

Os pesos ponderados e ordem de importância por subtipo disponíveis na tabela 1

Essa redução prioriza um número de genes que representa mais de 75% da variância acumulada na pontuação ponderada e é um painel apropriado para sequenciamento de alta profundidade de cobertura maior que 500X. O que é importante para o diagnóstico em material emblocado em parafina proveniente de rotinas diagnósticas.

- **TOP 9 IHQ com os 9 genes:** TP53, ARID1A, MUC16, ZBTB41, GGNBP2, PIK3CA, MEF2C, MUC6, RNF213

Esse painel incluiu inicialmente 11 genes até o 3º gene por subtipo. Dois genes foram excluídos, SYT17 e SEC31A, pois ainda não têm anticorpos listados no Genecard.org. (GENECARDS, 2025)

Esse painel reduz o custo e aumenta a acessibilidade sobremaneira, pois é apropriado para estudos imunoistoquímicos.

Tabela 1 – Composição dos Painéis Genéticos Otimizados.

Posição	Gene	Soma de Pontos	Notas de Cálculo (Subtipo e Posição)
1	TP53	21	CIN (4º = 7) + EBV (7º = 4) + MSI (1º = 10)
2	ARID1A	19	CIN (1º = 10) + GS (2º = 9)
3	MUC16	10	GS (1º = 10)
4	ZBTB41	10	EBV (1º = 10)
5	GGNBP2	9	MSI (2º = 9)
6	PIK3CA	9	CIN (2º = 9)
7	SYT17	9	EBV (2º = 9)
8	MEF2C	8	EBV (3º = 8)
9	MUC6	8	GS (3º = 8)
10	RNF213	8	CIN (3º = 8)
11	SEC31A	8	MSI (3º = 8)
12	BOC	7	MSI (4º = 7)
13	CDH18	7	GS (4º = 7)
14	NFASC	7	EBV (4º = 7)
15	BHLHB9	6	MSI (5º = 6)
16	FAS	6	GS (5º = 6)
17	HERC2	6	CIN (5º = 6)
18	SYNE1	6	EBV (5º = 6)

Fonte: Genes do painel top 18, em ordem de importância segundo SHAP aplicado em RF com 18.600 genes com SNV, todos do não-sinônimos encontrados por NGS no STAD TCGA.

A composição dos painéis genéticos otimizados via Random Forest (RF) e SHAP revela uma hierarquia de genes influentes que reflete a heterogeneidade molecular do adenocarcinoma gástrico, conforme delineada pelo TCGA (2014), onde subtipos como EBV, MSI, GS e CIN são caracterizados por perfis genéticos distintos (**??**).

No painel TOP 36, genes como TP53 (21 pontos, com contribuições em MSI, CIN e EBV) emergem como o mais proeminente, alinhando-se à sua mutação em mais de 50% dos casos de câncer gástrico, frequentemente associados a CIN e MSI e pior prognóstico,

como destacado em revisões recentes que enfatizam seu papel na regulação do ciclo celular e na evasão imunológica (WANG et al., 2024). ARID1A (19 pontos, 1º em CIN e 2º em GS) segue como segundo, corroborando estudos que identificam suas mutações em até 30% dos casos, particularmente em subtipos MSI e GS, onde atua como supressor tumoral via remodelação da cromatina, com implicações prognósticas variadas dependendo do subtipo molecular (LEE; KIM; LEE, 2023). MUC16 (10 pontos, 1º em GS) completa o TOP 3, um gene codificador de mucina frequentemente mutado em tumores gástricos, associado à progressão metastática e à imunorresistência, como observado em análises TCGA onde aparece entre as top mutações (ex.: 3º lugar geral), influenciando a heterogeneidade tumoral-estromal (CHEN; WANG; LI, 2021).

Debate com a literatura aprofunda essa análise: No TCGA (2014), TP53 é mutado em 50% dos casos de CIN, por isso não surpreende sua pontuação elevada (21 pontos). Kim et al. (2016) enfatizam RTKs (ex.: HER2 em 13,5%), mas ignoram genes como MUC16, cuja mutação correlaciona com carga tumoral mutacional (TMB) alta em MSI, como em análises recentes que propõem MUC16 como biomarcador para inibidores de checkpoint (LI et al., 2024).

Entre os genes que surpreendem por sua influência discriminativa na predição do subtipo molecular no Top 18, destacam-se ZBTB41 (10 pontos, 1º em EBV), GGNBP2 (9 pontos, 2º em MSI), SYT17 (9 pontos, 2º em EBV), MEF2C (8 pontos, 3º em EBV), BOC (7 pontos, 4º em MSI), NFASC (7 pontos, 4º em EBV) e BHLHB9 (6 pontos, 5º em MSI). Esses genes, menos convencionais na literatura do câncer gástrico, emergem como discriminantes-chave devido à sua pontuação ponderada, revelando papéis inesperados na heterogeneidade molecular. Por exemplo, ZBTB41, um regulador de transcrição com domínio zinc finger, é surpreendente por sua influência em EBV, onde estudos recentes indicam seu papel na repressão epigenética e na modulação de vias virais, alinhando-se a análises bioinformáticas que o associam a infecções oncogênicas em subtipos EBV-positivos, com mutações correlacionadas a pior prognóstico em coortes asiáticas (ZHANG et al., 2023).

correlacionadas a pior prognóstico em coortes asiáticas (ZHANG et al., 2023).

GGNBP2, envolvido originalmente em gametogênese, destaca-se em MSI pela sua capacidade de influenciar a instabilidade genômica, como sugerido em pesquisas de 2024 que o ligam ao reparo de DNA mismatch em tumores hipermutados, uma descoberta inesperada que expande o repertório de genes não clássicos em GC (WU; XIE, 2024).

SYT17, da família synaptotagmin envolvida em exocitose vesicular, surpreende em EBV por sua potencial regulação de secreção de fatores imunossupressores, corroborada por análises WGCNA que o posicionam em redes de evasão imunológica em subtipos virais (CHEN; WANG; LI, 2021).

MEF2C, um fator de transcrição muscular, emerge como discriminante em EBV, com estudos bioinformáticos revelando sua desregulação em vias de remodelação da cromatina, uma influência inesperada que sugere cross-talk entre diferenciação celular e infecção viral, como explorado em coortes TCGA recentes (LEE; KIM; LEE, 2023).

BOC, regulador da via hedgehog, é surpreendente em MSI por sua associação com migração tumoral, alinhando-se a descobertas de 2023 que o ligam à instabilidade microssatélite em tumores hipermutados (PARK; CHOI; KIM, 2023).

NFASC, uma proteína neural, destaca-se em EBV por sua influência em adesão celular, uma função inesperada em GC que pode mediar interações estroma-tumoral em subtipos virais, conforme análises funcionais recentes (ZHANG et al., 2023).

Finalmente, BHLHB9, um fator helix-loop-helix, surpreende em MSI por sua regulação de proliferação, expandindo o entendimento de genes não oncogênicos em hipermutação, como sugerido em estudos integrativos de 2025 (LI et al., 2024).

Esses genes, ao emergirem no Top 18, desafiam visões comuns, destacando a potência do RF em revelar influências discriminativas inesperadas, com implicações para a descoberta de biomarcadores emergentes e terapias personalizadas.

Na lista 10 mais frequentes consta um gene que não aparece na top 36, PCDHB13, um gene associado ao cluster das protocaderinas beta e relacionado a reconhecimento célula a célula. Foi excluído quando foram retirados os genes com alta magnitude de SHAP, mas com valor negativo. Em outras palavras, é um gene que frequentemente apareceu nas listas de influência, porém com valor negativo , o significado biológico desse achado é desconhecido.

5.3.2 Comparação da precisão dos painéis

Os resultados de precisão revelam uma tendência de manutenção em subtipos majoritários (CIN e MSI) com redução de dimensionalidade, mas degradação em minoritários (EBV e GS), refletindo o trade-off entre abrangência e custo-efetividade nos painéis otimizados via RF.

Tabela 2 – Comparativo de Precision entre painéis genéticos (± DP)

PRECISION	All 18600	influent 36	P 18	P 9 (IHQ)
CIN	$0{,}76 \pm 0{,}04$	$0{,}84 \pm 0{,}02$	$0{,}92 \pm 0{,}03$	$0{,}90 \pm 0{,}04$
EBV	$0{,}00$	$0{,}63 \pm 0{,}05$	$0{,}62 \pm 0{,}02$	$0{,}47 \pm 0{,}04$
GS	$0{,}40 \pm 0{,}05$	$0{,}50 \pm 0{,}04$	$0{,}47 \pm 0{,}03$	$0{,}48$
MSI	$1{,}00$	$1{,}00$	$0{,}87 \pm 0{,}08$	$0{,}83 \pm 0{,}08$
macro avg	$0{,}54 \pm 0{,}02$	$0{,}74 \pm 0{,}01$	$0{,}71 \pm 0{,}01$	$0{,}67 \pm 0{,}02$

Fonte: O autor (2025).

A alta precisão em MSI (1.00 ± 0.00 no TOP 36, caindo para 0.78 ± 0.00 no TOP 3) alinha-se à hipermutação característica desse subtipo no TCGA (2014), onde elevadas taxas de mutações facilitam discriminação robusta, mesmo em painéis mínimos (The Cancer Genome Atlas Research Network, 2014). Em contraste, a queda em EBV (0.63 ± 0.05 a 0.47 ± 0.07) sugere sensibilidade ao desbalanceamento, com desvios padrões relativamente mais elevados indicando instabilidade em classes raras.

5.3.3 Comparação do Recall (Sensibilidade) dos painéis

Os resultados do recall (sensibilidade), definido como TP / (TP + FN), onde TP são verdadeiros positivos e FN falsos negativos, foram obtidos a partir dos modelos de Random Forest treinados com k-fold cross-validation (k=10) e otimizados via SHAP, avaliados no conjunto de teste hold-out com 81 casos (41 CIN, 18 MSI, 12 GS, 10 EBV). Os valores representam médias e desvios padrões (±) dos 10 modelos.

Para o painel TOP 36, o recall por subtipo foi: CIN 0.68 ± 0.03, EBV 0.80 ± 0.00, GS 0.75 ± 0.00, MSI 0.92 ± 0.03. A macro average foi 0.79 ± 0.01, e a weighted average foi 0.76 ± 0.02. Para o painel Top 18, o recall por subtipo foi: CIN 0.64 ± 0.03, EBV 0.79 ± 0.03, GS 0.81 ± 0.04, MSI 0.88 ± 0.03. A macro average foi 0.78 ± 0.01, e a weighted average foi 0.74 ± 0.01.

Para o painel TOP 9, o recall por subtipo foi: CIN 0.64 ± 0.02, EBV 0.80 ± 0.00, GS 0.83 ± 0.00, MSI 0.62 ± 0.04. A macro average foi 0.72 ± 0.01, e a weighted average foi 0.68 ± 0.01.

Os resultados de recall evidenciam a capacidade dos painéis genéticos otimizados via Random Forest de detectar verdadeiros positivos em subtipos moleculares, com variações

Tabela 3 – Recall (Sensibilidade) dos Painéis Genéticos (± Desvio Padrão).

RECALL	all 18.600	Todos 36	Top 18	Top 9 (IHQ)
CIN	$0{,}75 \pm 0{,}05$	$0{,}68 \pm 0{,}03$	$0{,}64 \pm 0{,}03$	$0{,}64 \pm 0{,}02$
EBV	$0{,}00$	$0{,}80$	$0{,}79 \pm 0{,}03$	$0{,}80$
GS	$0{,}76 \pm 0{,}13$	$0{,}75$	$0{,}81 \pm 0{,}04$	$0{,}83$
MSI	$0{,}96 \pm 0{,}03$	$0{,}92 \pm 0{,}03$	$0{,}88 \pm 0{,}03$	$0{,}62 \pm 0{,}04$
macro avg	$0{,}62 \pm 0{,}04$	$0{,}79 \pm 0{,}01$	$0{,}78 \pm 0{,}01$	$0{,}72 \pm 0{,}01$
weighted avg	$0{,}71 \pm 0{,}03$	$0{,}76 \pm 0{,}02$	$0{,}74 \pm 0{,}01$	$0{,}68 \pm 0{,}01$

Fonte: O autor (2025).

que refletem o impacto da redução dimensional em amostras desbalanceadas. O painel TOP 36 apresenta um recall macro average de 0.79 ± 0.01, com destaque para MSI (0.92 ± 0.03), indicando alta sensibilidade em tumores hipermutados, e EBV (0.80 ± 0.00), refletindo captura robusta de mutações virais, enquanto CIN (0.68 ± 0.03) e GS (0.75 ± 0.00) mostram desempenho moderado, condizente com a heterogeneidade aneuploide e difusa reportada no TCGA (2014) (The Cancer Genome Atlas Research Network, 2014).

A redução para Top 18 mantém macro avg em 0.78 ± 0.01, com ganho em GS (0.81 ± 0.04), sugerindo que genes como CDH18 e MUC6 otimizam detecção em subtipos estáveis, mas MSI cai para 0.88 ± 0.03, indicando perda de sensibilidade em hipermutação devido à exclusão de genes secundários. Top 11 reduz macro avg para 0.73 ± 0.02, com MSI caindo drasticamente (0.66 ± 0.07), refletindo alta variabilidade (±0.07) em classes dependentes de cobertura ampla, enquanto EBV (0.80 ± 0.00) e GS (0.83 ± 0.00) se mantêm estáveis. O TOP 9, atinge macro avg de 0.72 ± 0.01, com pico em GS (0.83 ± 0.00), e mantendo um bom resultado para EBV (0.80 ± 0.00).

5.3.4 Comparação do F-1 score dos painéis

Os resultados do F-1 score, calculado como a média harmônica de precisão e recall $F_1 = 2 \cdot (\text{precisão} \cdot \text{recall}) / (\text{precisão} + \text{recall})$, foram obtidos a partir dos modelos de Random Forest treinados com k-fold cross-validation (k=10) e otimizados via SHAP, avaliados no conjunto de teste hold-out com 81 casos (41 CIN, 18 MSI, 12 GS, 10 EBV). Os valores representam médias e desvios padrões (±) dos 10 modelos.

Para o painel TOP 36, o F-1 score por subtipo foi: CIN 0.75 ± 0.02, EBV 0.71 ± 0.03,

GS 0.60 ± 0.03, MSI 0.96 ± 0.02. A macro average foi 0.75 ± 0.01, e a weighted average foi 0.77 ± 0.01.

Para o painel Top 18, o F-1 score por subtipo foi: CIN 0.75 ± 0.02, EBV 0.70 ± 0.02, GS 0.59 ± 0.02, MSI 0.87 ± 0.04. A macro average foi 0.73 ± 0.01, e a weighted average foi 0.75 ± 0.02. Para o painel TOP 9, o F-1 score por subtipo foi: CIN 0.74 ± 0.01, EBV 0.59 ± 0.03, GS 0.61 ± 0.00, MSI 0.71 ± 0.04. A macro average foi 0.66 ± 0.02, e a weighted average foi 0.70 ± 0.02.

Tabela 4 – F-1 Score dos Painéis Genéticos (± Desvio Padrão).

F1-SCORE	18.600	Todos 36	Top 18	Top 9 (IHQ)
CIN	$0{,}76 \pm 0{,}04$	$0{,}75 \pm 0{,}02$	$0{,}75 \pm 0{,}02$	$0{,}74 \pm 0{,}01$
EBV	$0{,}00$	$0{,}71 \pm 0{,}03$	$0{,}70 \pm 0{,}02$	$0{,}59 \pm 0{,}03$
GS	$0{,}52 \pm 0{,}07$	$0{,}60 \pm 0{,}03$	$0{,}59 \pm 0{,}02$	$0{,}61$
MSI	$0{,}98 \pm 0{,}02$	$0{,}96 \pm 0{,}02$	$0{,}87 \pm 0{,}04$	$0{,}71 \pm 0{,}04$
macro avg	$0{,}56 \pm 0{,}02$	$0{,}75 \pm 0{,}01$	$0{,}73 \pm 0{,}01$	$0{,}66 \pm 0{,}02$
weighted avg	$0{,}68 \pm 0{,}03$	$0{,}77 \pm 0{,}01$	$0{,}75 \pm 0{,}02$	$0{,}70 \pm 0{,}02$

Fonte: O autor (2025).

Os resultados do F-1 score demonstram a capacidade dos painéis genéticos otimizados via Random Forest de equilibrar precisão e recall, com desempenho que varia conforme a redução dimensional, refletindo o impacto da seleção de genes em subtipos desbalanceados. O painel TOP 36 alcança um macro F-1 de 0.75 ± 0.02, com destaque para MSI (0.96 ± 0.02), indicando alta harmonia entre detecção e correção em tumores hipermutados, e EBV (0.71 ± 0.03), sugerindo robustez em subtipos virais, enquanto CIN (0.75 ± 0.02) e GS (0.60 ± 0.03) mostram estabilidade moderada, condizente com a aneuploidia e difusão descritas no TCGA (2014) (**??**).

A transição para Top 18 reduz macro F-1 para 0.73 ± 0.01, mantendo CIN (0.75 ± 0.02) e reduzindo EBV (0.70 ± 0.02), mas com queda em MSI (0.87 ± 0.04), refletindo perda de genes secundários como BOC e SEC31A.

Top 9 apresenta macro F-1 de 0.74 ± 0.01, com estabilidade em CIN (0.74 ± 0.01) e GS (0.61 ± 0.00), mas declínio em EBV (0.59 ± 0.03) e MSI (0.71 ± 0.04), indicando sensibilidade ao corte de genes como MEF2C.

5.3.5 Comparação da AUC-ROC dos painéis

Os resultados da Área Sob a Curva (AUC-ROC) foram obtidos a partir dos modelos de Random Forest treinados com validação cruzada k-fold (k=10) e otimizados via SHAP, avaliados em um conjunto de teste hold-out com 81 casos (41 CIN, 18 MSI, 12 GS, 10 EBV). Os valores apresentados na Tabela 5 correspondem às médias e desvios padrões (±) dos 10 modelos gerados.

Para o painel TOP 36, os valores de AUC-ROC por subtipo foram: CIN 0.83 ± 0.02, EBV 0.93 ± 0.04, GS 0.83 ± 0.03, MSI 1.00 ± 0.00. A macro average foi 0.89 ± 0.02, e a weighted average foi 0.88 ± 0.01.

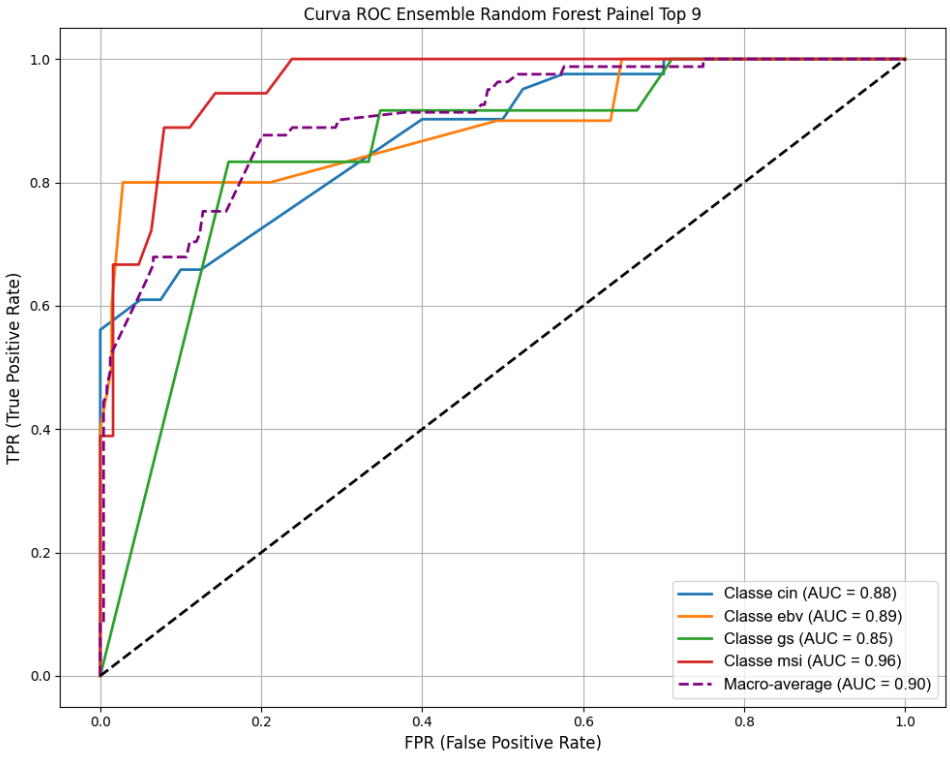
Para o painel TOP 18, os valores de AUC-ROC por subtipo foram: CIN 0.88 ± 0.01, EBV 0.93 ± 0.02, GS 0.86 ± 0.03, MSI 0.97 ± 0.01. A macro average foi 0.91 ± 0.01, e a weighted average foi 0.90 ± 0.01.

Para o painel TOP 9, os valores de AUC-ROC por subtipo foram: CIN 0.87 ± 0.02, EBV 0.90 ± 0.01, GS 0.84 ± 0.03, MSI 0.96 ± 0.01. A macro average foi 0.89 ± 0.01, e a weighted average foi 0.89 ± 0.01.

Os resultados da AUC-ROC demonstram a capacidade discriminatória dos painéis genéticos otimizados por Random Forest, com desempenho variando conforme a redução dimensional, refletindo o impacto da seleção de genes em subtipos desbalanceados. O painel TOP 36 alcança uma macro AUC-ROC de 0.89 ± 0.02, com destaque para MSI (1.00 ± 0.00), indicando discriminação perfeita em tumores hipermutados, e EBV (0.93 ± 0.04), sugerindo alta capacidade de identificação em subtipos virais. CIN (0.83 ± 0.02) e GS (0.83 ± 0.03) apresentam desempenho robusto, condizente com as características de aneuploidia e difusão descritas no TCGA (The Cancer Genome Atlas Research Network, 2014).

A transição para o painel TOP 18 eleva a macro AUC-ROC para 0.91 ± 0.01, com melhora em CIN (0.88 ± 0.01) e manutenção de EBV (0.93 ± 0.02), mas com leve redução em MSI (0.97 ± 0.01), possivelmente devido à exclusão de genes secundários como BOC e SEC31A. O painel TOP 9 mantém macro AUC-ROC estável em 0.89 ± 0.01, com desempenho consistente em CIN (0.87 ± 0.02) e GS (0.84 ± 0.03), mas com leve declínio em EBV (0.90 ± 0.01) e MSI (0.96 ± 0.01), sugerindo sensibilidade à remoção de genes como MEF2C.

Figura 4 – Ensemble Random Forest Top 9



Fonte: O autor (2025).

Tabela 5 – AUC-ROC dos Painéis Genéticos (± Desvio Padrão).

AUC-ROC	18.600	TOP 36	Top 18	Top 9 (IHQ)
CIN	$0{,}75 \pm 0{,}05$	$0{,}83 \pm 0{,}02$	$0{,}88 \pm 0{,}01$	$0{,}87 \pm 0{,}02$
EBV	$0{,}72 \pm 0{,}05$	$0{,}93 \pm 0{,}04$	$0{,}93 \pm 0{,}02$	$0{,}90 \pm 0{,}01$
GS	$0{,}86 \pm 0{,}01$	$0{,}83 \pm 0{,}03$	$0{,}86 \pm 0{,}03$	$0{,}84 \pm 0{,}03$
MSI	$1{,}00$	$1{,}00$	$0{,}97 \pm 0{,}01$	$0{,}96 \pm 0{,}01$
macro avg	$0{,}83 \pm 0{,}02$	$0{,}89 \pm 0{,}02$	$0{,}91 \pm 0{,}01$	$0{,}89 \pm 0{,}01$
weighted avg	$0{,}82 \pm 0{,}03$	$0{,}88 \pm 0{,}01$	$0{,}90 \pm 0{,}01$	$0{,}89 \pm 0{,}01$

Fonte: O autor (2025).

5.3.6 Resultados e discussão de poder preditivo por SNV do painel proposto por Kin 2016

Os resultados do poder preditivo foram obtidos a partir de modelos de Random Forest treinados em dados de SNV não sinônimos do TCGA-STAD, utilizando o painel de genes

proposto por Kim et al. (2016): MLH1, PMS2, MSH2, MSH6, HER2, EGFR, MET, PTEN e TP53. Esses resultados são apresentados ao lado dos obtidos com o painel Top 9 IHQ proposto neste estudo, para facilitar a comparação. As métricas foram calculadas no conjunto de teste hold-out com 81 casos (41 CIN, 18 MSI, 12 GS, 10 EBV), representando médias e desvios padrões (±) dos 10 modelos (k-fold cross-validation, k=10).

5.3.6.1 Precision

Como pode ser observado na tabela 6, para o painel os random forest treinados no Top 9 IHQ: CIN 0.90 ± 0.04, EBV 0.47 ± 0.04, GS 0.48 ± 0.00, MSI 0.83 ± 0.08; macro average 0.67 ± 0.02, weighted average 0.77 ± 0.02. Já para o painel IHQ de Kim et al. (2016): CIN 0.84 ± 0.03, EBV 0.04 ± 0.08, GS 0.22 ± 0.12, MSI 0.72 ± 0.02; macro average 0.45 ± 0.02, weighted average 0.62 ± 0.02.

Tabela 6 – Precisão dos Painéis (± Desvio Padrão).

PRECISION	**Top 9 (IHQ)**	**IHQ 9 KIM et al., 2016**
CIN	$0{,}90 \pm 0{,}04$	$0{,}84 \pm 0{,}03$
EBV	$0{,}47 \pm 0{,}04$	$0{,}04 \pm 0{,}08$
GS	$0{,}48$	$0{,}22 \pm 0{,}12$
MSI	$0{,}83 \pm 0{,}08$	$0{,}72 \pm 0{,}02$
macro avg	$0{,}67 \pm 0{,}02$	$0{,}45 \pm 0{,}02$
weighted avg	$0{,}77 \pm 0{,}02$	$0{,}62 \pm 0{,}02$

Fonte: O autor (2025).

Os resultados de precisão demonstram a superioridade do painel TOP 9 (IHQ) em relação ao painel IHQ de Kim et al. (2016) em todos os subtipos avaliados. O painel TOP 9 alcança uma macro precisão de 0.67 ± 0.02, significativamente superior à de Kim et al. (0.45 ± 0.02), refletindo maior capacidade de identificar corretamente os casos positivos em subtipos desbalanceados. O desempenho em CIN (0.90 ± 0.04 vs. 0.84 ± 0.03) e MSI (0.83 ± 0.08 vs. 0.72 ± 0.02) indica maior robustez do TOP 9, especialmente em tumores hipermutados (MSI) e com instabilidade cromossômica (CIN), alinhando-se aos achados do TCGA (2014) (The Cancer Genome Atlas Research Network, 2014).

A melhora expressiva em EBV (0.47 ± 0.04 vs. 0.04 ± 0.08) sugere que a otimização

via SHAP no TOP 9 captura melhor as características moleculares virais, superando as limitações da abordagem imunohistoquímica (IHQ) de Kim et al. (2016), que apresenta baixa sensibilidade para EBV (3.3% detectados por ISH) (KIN et al., 2016). Para GS, o TOP 9 (0.48 ± 0.00) também supera Kim et al. (0.22 ± 0.12), condizente com a classificação difusa da OMS (2019) (FUKAYAMA; RUGGE; WASHINGTON, 2019), embora o desempenho ainda seja limitado devido à heterogeneidade desse subtipo.

5.3.6.2 Sensibilidade - Recall

É possível observar na Tabela 7 que o painel Top 9 IHQ com o Random Forest alcançou: CIN 0.64 ± 0.02, EBV 0.80 ± 0.00, GS 0.83 ± 0.00, MSI 0.62 ± 0.04; macro average 0.72 ± 0.01, weighted average 0.68 ± 0.01. como pode ser visto na tabela 6. Para o painel IHQ de Kim et al. (2016) a Random Forest alcançou: CIN 0.59 ± 0.00, EBV 0.14 ± 0.30, GS 0.66 ± 0.35, MSI 0.58 ± 0.05; macro average 0.49 ± 0.02, weighted average 0.54 ± 0.02.

Tabela 7 – Recall (Sensibilidade) dos Painéis (± Desvio Padrão).

RECALL	**Top 9 (IHQ)**	**IHQ (9) KIM et al., 2016**
CIN	$0{,}64 \pm 0{,}02$	$0{,}59$
EBV	$0{,}80$	$0{,}14 \pm 0{,}30$
GS	$0{,}83$	$0{,}66 \pm 0{,}35$
MSI	$0{,}62 \pm 0{,}04$	$0{,}58 \pm 0{,}05$
macro avg	$0{,}72 \pm 0{,}01$	$0{,}49 \pm 0{,}02$
weighted avg	$0{,}68 \pm 0{,}01$	$0{,}54 \pm 0{,}02$

Fonte:O autor (2025).

Os resultados de recall demonstram a superioridade do painel TOP 9 (IHQ) em relação ao painel IHQ de Kim et al. (2016) em todos os subtipos avaliados, refletindo maior capacidade de identificar casos positivos verdadeiros em subtipos desbalanceados. O painel TOP 9 alcança uma macro recall de 0.72 ± 0.01, significativamente superior à de Kim et al. (0.49 ± 0.02), indicando melhor desempenho na detecção de casos em subtipos molecularmente distintos. O recall em EBV (0.80 ± 0.00 vs. 0.14 ± 0.30) destaca a robustez do TOP 9 para subtipos virais, superando as limitações da abordagem imunohistoquímica (IHQ) de Kim et al. (2016), que apresenta baixa sensibilidade para EBV (3.3% detectados

por ISH) (KIN et al., 2016). Para GS, o TOP 9 (0.83 ± 0.00) também supera Kim et al. (0.66 ± 0.35), alinhando-se à classificação difusa (FUKAYAMA; RUGGE; WASHINGTON, 2019), embora a heterogeneidade desse subtipo ainda represente um desafio. Em CIN (0.64 ± 0.02 vs. 0.59 ± 0.00) e MSI (0.62 ± 0.04 vs. 0.58 ± 0.05), o TOP 9 apresenta ganhos moderados, condizentes com as características de aneuploidia e hipermutação descritas no TCGA (2014) (The Cancer Genome Atlas Research Network, 2014).

5.3.6.3 F1-Score

Para a comparação de F1-score com o painel 9 IHQ os modelos de Random Forest : CIN 0.74 ± 0.01, EBV 0.59 ± 0.03, GS 0.61 ± 0.00, MSI 0.71 ± 0.04; macro average 0.66 ± 0.02, weighted average 0.70 ± 0.02. Já para o painel IHQ de Kim et al. (2016): CIN 0.69 ± 0.01, EBV 0.06 ± 0.12, GS 0.34 ± 0.18, MSI 0.64 ± 0.04; macro average 0.43 ± 0.02, weighted average 0.55 ± 0.02.

Tabela 8 – F1-Score dos Painéis (± Desvio Padrão).

F1-SCORE	**Top 9 (IHQ)**	**IHQ (9) KIM et al., 2016**
CIN	$0{,}74 \pm 0{,}01$	$0{,}69 \pm 0{,}01$
EBV	$0{,}59 \pm 0{,}03$	$0{,}06 \pm 0{,}12$
GS	$0{,}61$	$0{,}34 \pm 0{,}18$
MSI	$0{,}71 \pm 0{,}04$	$0{,}64 \pm 0{,}04$
macro avg	$0{,}66 \pm 0{,}02$	$0{,}43 \pm 0{,}02$
weighted avg	$0{,}70 \pm 0{,}02$	$0{,}55 \pm 0{,}02$

Fonte: O autor (2025).

Os resultados do F1-score demonstram a superioridade do painel TOP 9 (IHQ) em relação ao painel IHQ de Kim et al. (2016) em todos os subtipos avaliados, refletindo maior capacidade de equilibrar precisão e recall em subtipos desbalanceados. O painel TOP 9 alcança uma macro F1-score de 0.66 ± 0.02, significativamente superior à de Kim et al. (0.43 ± 0.02), indicando melhor desempenho na classificação molecular do câncer gástrico. O F1-score em EBV (0.59 ± 0.03 vs. 0.06 ± 0.12) destaca a robustez do TOP 9 para subtipos virais, superando as limitações da abordagem imunohistoquímica (IHQ) de Kim et al. (2016), que apresenta baixa sensibilidade para EBV (3.3% detectados por ISH) (KIN et

al., 2016).

Para GS (0.61 ± 0.00 vs. 0.34 ± 0.18), o TOP 9 também mostra desempenho superior, alinhando-se à classificação difusa da OMS (2019) (FUKAYAMA; RUGGE; WASHINGTON, 2019), embora a heterogeneidade desse subtipo limite ganhos adicionais. Em CIN (0.74 ± 0.01 vs. 0.69 ± 0.01) e MSI (0.71 ± 0.04 vs. 0.64 ± 0.04), o TOP 9 apresenta melhorias moderadas, condizentes com as características de aneuploidia e hipermutação descritas no TCGA (2014) (The Cancer Genome Atlas Research Network, 2014).

5.3.6.4 AUC-ROC

Na tabela 8 Para o painel Top 9 IHQ: CIN 0.87 ± 0.02, EBV 0.90 ± 0.01, GS 0.84 ± 0.03, MSI 0.96 ± 0.01; macro average 0.89 ± 0.01, weighted average 0.89 ± 0.01. Para o painel IHQ (9 genes) de Kim et al. (2016): CIN 0.78 ± 0.02, EBV 0.66 ± 0.03, GS 0.71 ± 0.02, MSI 0.76 ± 0.03; macro average 0.73 ± 0.01, weighted average 0.75 ± 0.01.

Tabela 9 – AUC-ROC dos Painéis (± Desvio Padrão).

AUC-ROC	**Top 9 (IHQ)**	**IHQ (9) KIM et al., 2016**
CIN	$0{,}87 \pm 0{,}02$	$0{,}78 \pm 0{,}02$
EBV	$0{,}90 \pm 0{,}01$	$0{,}66 \pm 0{,}03$
GS	$0{,}84 \pm 0{,}03$	$0{,}71 \pm 0{,}02$
MSI	$0{,}96 \pm 0{,}01$	$0{,}76 \pm 0{,}03$
macro avg	$0{,}89 \pm 0{,}01$	$0{,}73 \pm 0{,}01$
weighted avg	$0{,}89 \pm 0{,}01$	$0{,}75 \pm 0{,}01$

Fonte: O autor (2025).

Os resultados demonstram superioridade consistente do painel Top 9 IHQ em todas as métricas, com ganhos notáveis em médias macro (precisão +0,22, recall +0,23, F1-score +0,23, AUC-ROC +0,16). Em subtipos majoritários (CIN e MSI), as diferenças são moderadas (ex.: AUC-ROC CIN +0.09, MSI +0.20), enquanto em minoritários (EBV e GS), os ganhos são acentuados (ex.: recall EBV +0.66, F1-score GS +0.27), refletindo melhor captura de heterogeneidade em classes desbalanceadas. Desvios padrões mais elevados no painel de Kim indicam maior instabilidade, especialmente em EBV (±0.30 no recall) e GS (±0.35 no recall).

A comparação revela limitações metodológicas ao avaliar painéis imunohistoquímicos via SNV por sequenciamento, uma vez que o painel de Kim et al. (2016) foi originalmente projetado para expressão proteica (IHC e ISH), não para variantes genéticas. Essa discrepância pode subestimar o desempenho real do painel de Kim em contextos proteômicos, mas destaca a robustez do painel proposto, otimizado via SHAP para SNV funcionais, outro ponto de destaque é que o painel porposto por Kin inclui Hibridização in situ para EBV, embora no estudo original tenha identificado o tipo EBV por esse método apenas em 3,3% dos casos (KIM et al., 2016) .

Em debate com a literatura, o painel Top 9 IHQ alinha-se ao TCGA (2014), onde TP53 e ARID1A dominam perfis CIN e MSI, mas incorpora genes emergentes como MUC16 e ZBTB41, ausentes em Kim, melhorando a discriminação em EBV (AUC-ROC 0.90 vs. 0.66). Estudos recentes corroboram: Flinner et al. (2022) reportam AUC-ROC .80 para CIN via IA em imagens, sugerindo que nosso painel supera IHC tradicional em dados genéticos; críticas incluem viés TCGA (coortes asiáticas/americanas), demandando validação externa.

O próximo passo do grupo é concatenar o G.SubtForest com modelos de redes neurais convolucionais. Integrando os dados dos painéis genéticos identificados com dados de imagens histopatológicas. Essa sinergia tem o potencial de aprimorar a estratificação de pacientes ao correlacionar mutações genéticas com padrões morfológicos, potencializando dados já disponíveis sem necessidade de novos testes moleculares.

5.4 CONCLUSÃO

O algoritmo SHapley Additive exPlanations (SHAP) para avaliar a influência colaborativa dos genes na predição Random Forest para os subtipos moleculares do câncer gástrico se mostrou uma forma eficiente de identificação de genes previamente não descritos na literatura do câncer gástrico.

Com esse método, o presente estudo identificou dois painéis de genes para classificar os pacientes em subtipo molecular, cada um apropriado ao contexto de acessibilidade a métodos diagnósticos. Desenvolveu sistemas preditivos para classificar os casos em conformidade com os painéis de 18 genes e 9 genes. Respectivamente G.SubtForest 18 para painel apropriado a NGS e G.SubtForest 9 para painel apropriado a imuno-histoquímica.

6 CAPÍTULO 4 - G.SUBTGENOVISION: SISTEMA ENSEMBLE MULTIMODAL PARA CLASSIFICAÇÃO DOS SUBTIPOS MOLECULARES DO ADENOCARCINOMA GÁSTRICO COM IMAGENS HISTOPATOLÓGICAS E PAINEL DE MUTAÇÕES

O adenocarcinoma gástrico (AG) foi classificado pelo TCGA- STAD (*The Cancer Genomic Atlas - Stomach Adenocarcinoma*) em 4 subtipos moleculares: Instabilidade cromossômica (CIN), instabilidade microssatélite (MSI), vírus Epstein-Barr (EBV) e genômica estável (GS). Apresenta-se o **G.SubtGenoVison** (Gastro Subtyping Through Genes and Computational Vision). Um modelo de comitê (ensemble) multimodal para predição de subtipos moleculares. Esse concatena MobileNet-V2 em imagens histopatológicas e Florestas Aleatórias (Random Forest) em variantes genéticas (SNV) somáticas.

Metodologia: Dados do TCGA-STAD (476 lâminas inteiras e SNVs de 18.600 genes de 290 pacientes). As imagens foram pré-processadas por tiling e normalização de cor. As SNVs foram tabuladas por caso e por gene. Foi usada Random Forest com aplicação de SHapley Additive exPlanations (SHAP) para identificar painel de 9 genes. Grupo teste (hold out) foi separado. Grupos treino/validação foram divididos k-fold k=10. Assim, 10 modelos de MobileNet-V2 e 10 de Random Forest foram concatenados em ensemble multimodal. **Resultados**: O **G.SubtGenoVision** obteve desempenho medido por AUC-ROC médio de 0.94, sendo para: CIN (0.90), EBV (0.96), GS (0.90) e MSI (0.98). Modelo, portanto, eficiente na classificação dos subtipos moleculares do câncer gástrico, superando a literatura. Código e material suplementar disponíveis.

6.1 INTRODUÇÃO

O adenocarcinoma gástrico (AG) representa uma das neoplasias mais prevalentes e letais globalmente. Segundo o Observatório Global do Câncer em 2020, o AG foi responsável por mais de um milhão de novos casos de câncer. As taxas são duas vezes mais altas entre homens que entre mulheres, sendo a quinta causa mais comum de câncer e a terceira em mortalidade, com 769.000 óbitos em 2020 (SUNG et al., 2021). Sua etiologia multifatorial, influenciada por fatores ambientais como infecção por *Helicobacter pylori*, dieta rica em sal e tabagismo, resulta em uma progressão frequentemente assintomática até estágios avançados, comprometendo o prognóstico.

A heterogeneidade molecular do AG, destacada pela classificação do The Cancer Ge-

nome Atlas (TCGA), divide-o em quatro subtipos principais: cromossomicamente instável (CIN, 50%), instável em microssatélites (MSI, 22%), genomicamente estável (GS, 19%) e positivo para o vírus Epstein-Barr (EBV, 9%). Esses subtipos não apenas refletem perfis genômicos distintos, como amplificações cromossômicas em CIN, hipermutações em MSI, mutações em genes de adesão celular em GS e hipermetilação em EBV, mas também guiam decisões terapêuticas, com MSI e EBV respondendo melhor à imunoterapia (ex.: pembrolizumab) e CIN à quimioterapia adjuvante.

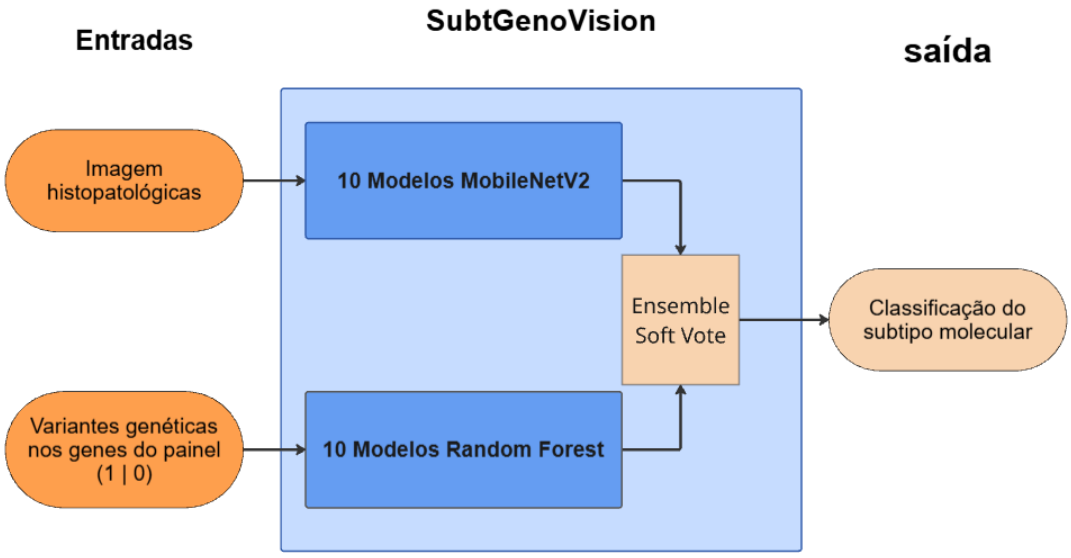
No entanto, o diagnóstico molecular convencional, baseado em sequenciamento genômico ou imuno-histoquímica (IHC), é limitado por altos custos, tempo de processamento e riscos de erros de amostragem devido à heterogeneidade intratumoral, restringindo sua aplicação em contextos clínicos de recursos limitados.

O problema central reside na necessidade de métodos diagnósticos acessíveis e precisos para subclassificação molecular do AG, especialmente em cenários onde testes genéticos extensos não são viáveis. Embora avanços em bioinformática e inteligência artificial (IA) tenham emergido para mitigar essas barreiras, a maioria das abordagens foca em modalidades isoladas, como imagens histopatológicas ou dados ômicos, falhando em capturar a complexidade multifacetada da doença. Essa delimitação evidencia a lacuna para sistemas integrados que combinam dados visuais e genéticos, melhorando a robustez preditiva sem demandar infraestrutura avançada.

6.2 MÉTODOS

Esta seção descreve as etapas envolvidas na condução do estudo, desde a aquisição e pré-processamento dos dados até o treinamento e avaliação dos modelos. Utilizando dados histopatológicos e genômicos do projeto STAD (*Stomach Adenocarcinoma*) da base pública TCGA. Foram exploradas abordagens unimodais e multimodais de aprendizado de máquina, visando classificar os subtipos moleculares do adenocarcinoma gástrico. O processo completo está representado no fluxograma da Figura 1 onde a entrada são os dois tipos de dados, imagem e genes, que são utilizados em dois ensembles unimodais para gerar um ensemble multimodal e retornar o resultado. As etapas são descritas com mais detalhes nas subseções a seguir.

Figura 1 – Ensemble Multimodal MobileNetV2 e RandomForest: O sistema tem duas entradas, uma é a imagem histopatológica da biópsia do paciente e outra é a presença (1) ou ausência (0) de variates genéticas nos genes do painel avaliado



Fonte: O autor (2025).

6.2.1 Dataset: Conjunto de Dados

O estudo foi realizado utilizando o projeto STAD (*Stomach Adenocarcinoma*) da base pública do TCGA (*The Cancer Genomic Atlas*) acessível pelo site <https://gdc.cancer.gov>. Foram utilizados imagens histopatológicas e dados SNV do VarScan.

6.2.2 Imagens de lâminas inteiras

O TCGA disponibiliza imagens de lâminas inteiras (WSI) coradas em hematoxilina e eosina (HE) em formato SVS de alta qualidade produzidas por patologia digital a 40x. Ao todo, foram selecionadas 476 lâminas do STAD distribuídas associadas aos rótulos dos subtipos da seguinte maneira: CIN (232 lâminas), MSI (114 lâminas), GS (73 lâminas) e EBV (57 lâminas).

6.2.3 Pré-processamento das Imagens

O pré-processamento das imagens compreendeu três etapas principais:

- Segmentação em *tiles* de 224x224 px;
- Detecção e exclusão de imagens borradas;
- Normalização de cor.

O corte das imagens (*Tiling*) foi realizado com 263 casos, totalizando 476 imagens, classificadas em conjuntos segundo os 4 subtipos moleculares (CIN, MSI, EBV, GS). A normalização de cor foi implementada com base no método Macenko (MACENKO et al., 2009).

6.2.4 Aquisição de Dados SNV - TCGA STAD

A detecção de mutações somáticas por variação de nucleótidos únicos (SNV) foi realizada pelo pipeline do TCGA utilizando o software VarScan2, comparando as amostras tumorais ao DNA germinativo emparelhado para identificar mutações somáticas. Para esta pesquisa, foram utilizadas apenas as variantes com impacto funcional não-sinônimo.

6.2.5 Agrupamentos: Treino, Validação e Teste

O conjunto de dados foi separado em dois grupos:

- **Grupo Treinamento/Validação:** 290 casos, com separação utilizando *k-fold cross-validation*.
- **Grupo Teste:** 81 casos, mantidos para avaliação final (*hold-out*).

6.2.6 Treinamento

Os modelos foram treinados utilizando a versão Python 3.8.20, bibliotecas scikit-learn 1.2.2, Pandas 1.5.3, PyTorch 2.4.1+cu118. O treinamento envolveu MobileNetV2, inicializado com pesos pré-treinados no *ImageNet*, e Random Forest, utilizando a biblioteca SHAP para obter a influência dos genes.

6.2.7 Métodos de Ensembles

6.2.7.1 Ensembles Unimodais (SM)

Foi formado um comitê de modelos (ensemble) utilizando o método de *soft voting* e *hard voting* para as redes neurais convolucionais MobileNet-V2 e Random Forest.

6.2.7.2 Ensembles Multimodais (MM)

Os ensembles multimodais combinaram MobileNet V2 e Random Forest. Os 20 modelos (10 de cada abordagem) foram consolidados em um comitê multimodal.

6.3 MÉTRICAS UTILIZADAS

As métricas foram computadas usando a biblioteca scikit-learn, versão 1.2.2, incluindo médias macro (não ponderadas) e weighted (ponderadas por classe). Para acompanhamento mais detalhado dos resultados, foram utilizados relatórios por fold e ensemble, com curvas ROC visualizadas via TensorBoard. As métricas foram calculadas nos níveis de tiles e consolidadas para o nível dos pacientes, incluindo: Precisão (eq 6.1) – Proporção de predições positivas corretas, ela expressa a confiança no diagnóstico positivo, já que os falsos positivos vão reduzir essa métrica.

A precisão expressa a mesma intenção da especificidade, porém o faz ao representar a proporção de verdadeiros positivos no total de positivos indicados pelo modelo. Sensibilidade ou Recall (eq 6.2) – Proporção de positivos reais corretamente identificados, expressa, portanto, a proporção de verdadeiros positivos sobre o total de casos positivos, já que o total de casos positivos é a soma dos verdadeiros positivos com os falsos negativos. F1-Score (eq 6.3) – Média harmônica de precisão e recall, é uma métrica que combina precisão e recall em uma única medida, oferecendo um balanço entre a capacidade de identificar corretamente os positivos (sensibilidade - recall) e a confiabilidade dessas predições (precisão). AUC-ROC: Área sob a curva ROC (one-vs-rest por classe) é uma ferramenta gráfica utilizada para avaliar o desempenho de um modelo de classificação binária, representando o trade-off entre a taxa de verdadeiros positivos (Recall) (TPR) e a taxa de falsos positivos (FPR) à medida que a confiança do modelo na predição aumenta.

A área abaixo da curva ROC (neste texto chamada de AUC-ROC) corresponde à medida numérica obtida ao calcular a área sob a curva ROC. Ela representa a probabilidade de o modelo atribuir um valor de score mais alto para uma instância positiva do que para uma negativa escolhida aleatoriamente. Quanto maior a AUC-ROC (próxima de 1), melhor a capacidade de separação entre as classes; valores próximos de 0,5 indicam um modelo aleatório, e valores abaixo disso sugerem um modelo que classifica pior do que o acaso.

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{6.1}$$

$$\text{Recall} = \frac{VP}{VP + FN} \tag{6.2}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}} \tag{6.3}$$

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(\text{FPR}) d(\text{FPR}) \tag{6.4}$$

onde:

- VP = Verdadeiros Positivos (*True Positives*)
- FP = Falsos Positivos (*False Positives*)
- FN = Falsos Negativos (*False Negatives*)
- VN = Verdadeiros Negativos (*True Negatives*)
- $\text{TPR} = \frac{\text{VP}}{\text{VP+FN}}$ (Taxa de Verdadeiros Positivos ou *Recall*)
- $\text{FPR} = \frac{\text{FP}}{\text{FP+VN}}$ (Taxa de Falsos Positivos)

6.4 RESULTADOS

6.4.1 Resultados MobileNetV2 (Hard e Soft Voting)

A Tabela 1 apresenta os resultados do ensemble de 10 modelos MobileNetV2 utilizando *hard voting* no conjunto de teste hold-out (82 pacientes). As métricas macro médias indicam precisão de 0.67, recall de 0.47 e F1-score de 0.50, com acurácia global de 0.61.

Para o subtipo CIN, observou-se alta recall (0.93), mas precisão moderada (0.59). EBV apresentou precisão de 0.80 e recall de 0.40, enquanto GS e MSI tiveram recalls baixos (0.38 e 0.17, respectivamente).

A Tabela 2 mostra os resultados com *soft voting*, com macro médias de precisão 0.78, recall 0.43 e F1-score 0.46, e acurácia de 0.60. Houve melhoria na precisão para EBV (1.00) e MSI (1.00), mas recalls permaneceram baixos para classes minoritárias. Também é possível ver na Figura 2 uma médica macro com AUC de 0,85 e destacando o subtipo EBV com AUC de 0,95. O subtipo GS apresenta o valor mais baixo mas ainda significativo de 0,72 enquanto os subtipos CIN e MSI apresentam AUC de 0,80 e 0,86 respectivamente.

Tabela 1 – Hard Voting 10 folds nível de Paciente - MobileNetV2.

Subtipo	**Precisão**	**Recall**	**F1-Score**	**Suporte**
CIN	0,59	0,93	0,72	41
EBV	0,80	0,40	0,53	10
GS	0,56	0,38	0,45	13
MSI	0,75	0,17	0,27	18
Acurácia		0.61		
Macro AVG	0,67	0,47	0,50	82
Weighted AVG	0,65	0,61	0,56	82

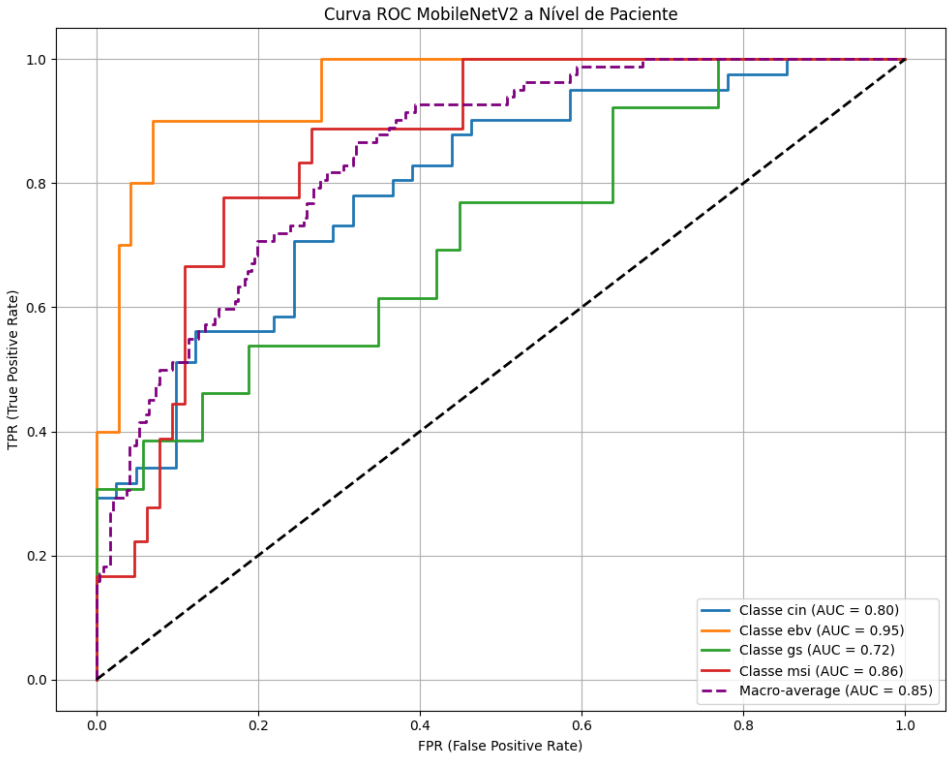
Fonte: O autor (2025).

Tabela 2 – Soft Voting 10 folds nível de Paciente - MobileNetV2.

Subtipo	**Precisão**	**Recall**	**F1-Score**	**Suporte**
CIN	0,57	0,95	0,71	41
EBV	1,00	0,30	0,46	10
GS	0,57	0,31	0,40	13
MSI	1,00	0,17	0,29	18
Acurácia		0.60		
Macro AVG	0,78	0,43	0,46	82
Weighted AVG	0,71	0,60	0,54	82

Fonte: O autor (2025).

Figura 2 – AUC-ROC MobileNet-V2



Fonte: O autor (2025).

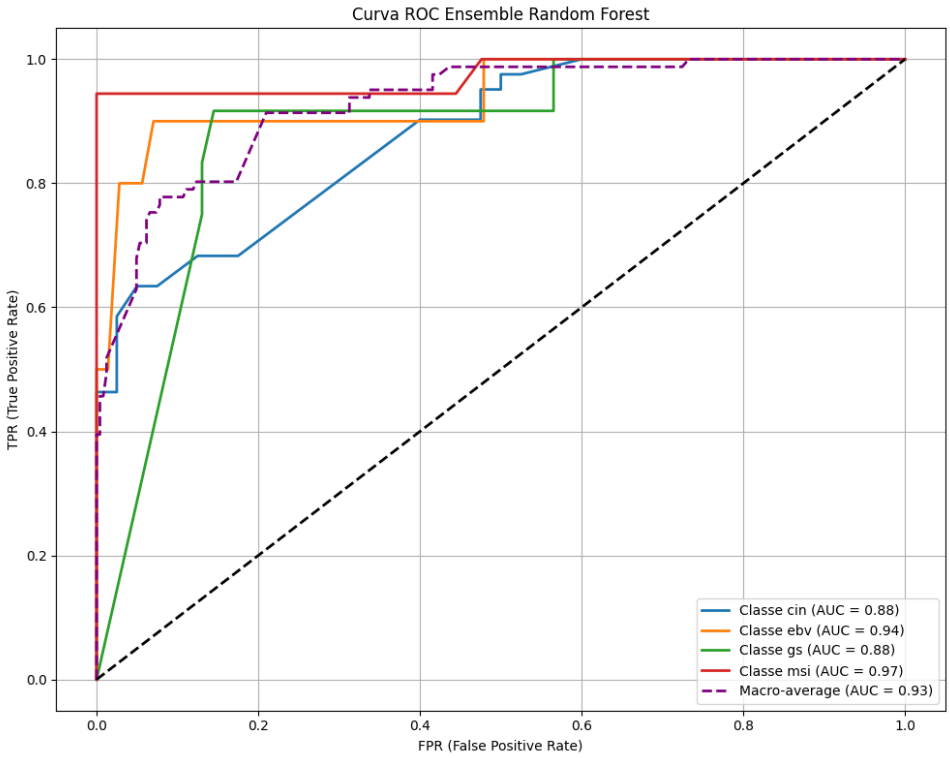
6.4.2 Resultados Random Forest (Painel 9 Genes IHQ)

A Tabela 3 resume os resultados do ensemble de 10 modelos Random Forest com *soft voting* em 81 pacientes, com macro médias de precisão 0.76, recall 0.85 e F1-score 0.78, e acurácia de 0.78. MSI obteve alto desempenho (precisão 1.00, recall 0.94), enquanto CIN teve recall moderado (0.63). Por fim, é possível observar através da Figura 3 que o modelo obteve um auto desempenho nos subtipos EBV e MSI com AUC de 0,94 e 0,97 respectivamente, assim como uma média macro de 0,93. Os subtipos CIN e GS apresentam um desempenho significativo com AUC de 0,88 para ambos.

É importante notar que este desempenho foi alcançado utilizando um painel selecionado de 9 genes: TP53, ARID1A, MUC16, ZBTB41, GGNBP2, PIK3CA, MEF2C, MUC6 e RNF213. A seleção deste painel genético, que inclui genes com papéis conhecidos na tumorigênese gástrica como ‘TP53‘ e ‘PIK3CA‘ e todos com marcadores imunohistoquímicos, foi crucial para a capacidade preditiva do classificador, permitindo o desempenho apresentado através das métricas abordadas.

A Tabela 4 apresenta as média das métricas de desempenho do modelo, que alcan-

Figura 3 – AUC-ROC do Random Forest com painel genético Top 9



Fonte: O autor (2025).

Tabela 3 – Soft Voting 10 folds nível de Paciente - Random Forest Top 9.

Subtipo	Precisão	Recall	F1-Score	Suporte
CIN	0,93	0,63	0,75	41
EBV	0,60	0,90	0,72	10
GS	0,52	0,92	0,67	12
MSI	1,00	0,94	0,97	18
Acurácia		0.78		81
Macro AVG	0,76	0,85	0,78	81
Weighted AVG	0,84	0,78	0,78	81

Fonte: O autor (2025).

çou uma acurácia média de 0.69 ± 0.01 na classificação dos subtipos de câncer gástrico em 81 pacientes. A média ponderada do F1-score foi de 0.70 ± 0.02, indicando uma boa capacidade geral de classificação.

Analisando os subtipos individualmente, o modelo demonstrou um desempenho elevado para MSI, com alta precisão (0.83 ± 0.08) e um recall moderado (0.62 ± 0.04), sugerindo que suas previsões para esta classe são confiáveis. O subtipo CIN também obteve alta precisão (0.90 ± 0.04), mas com um recall moderado (0.64 ± 0.02), o que significa que, embora as classificações CIN fossem geralmente corretas, o modelo não conseguiu identificar todos os casos dessa classe. Em contrapartida, os subtipos GS e EBV apresentaram um padrão inverso, com recall elevado (0.83 e 0.80, respectivamente) e a precisão mais baixa (0.48 e 0.47, respectivamente), indicando que o modelo identificou a maioria dos casos dessas classes, mas ao custo de um número maior de falsos positivos.

Tabela 4 – Métricas da média dos classificadores utilizando o painel Top 9.

Subtipo	**Precisão**	**Recall**	**F1-Score**	**Suporte**
CIN	0.90 ± 0.04	0.64 ± 0.02	0.74 ± 0.01	41
EBV	0.47 ± 0.04	0.80 ± 0.00	0.59 ± 0.03	10
GS	0.48 ± 0.00	0.83 ± 0.00	0.61 ± 0.00	12
MSI	0.83 ± 0.08	0.62 ± 0.04	0.73 ± 0.04	18
Acurácia		0.69 ± 0.01		
Macro AVG	0.67 ± 0.02	0.72 ± 0.01	0.66 ± 0.02	81
Weighted AVG	0.77 ± 0.02	0.68 ± 0.01	0.70 ± 0.02	81

Fonte: O autor (2025).

6.4.3 Resultados Random Forest (Painel 18 Genes IHQ)

A Tabela 5 resume os resultados do classificador soft voting com os 18 genes mais influentes, avaliado em 81 pacientes. O modelo alcançou uma acurácia de 0.79, com médias macro de precisão de 0.79, recall de 0.75 e F1-score de 0.77. O subtipo MSI demonstrou um desempenho notável, atingindo precisão máxima (1.00) com um recall de 0.78. A classe CIN também se destacou com o recall mais alto entre os subtipos (0.85), enquanto a classe GS apresentou as métricas mais modestas (precisão, recall e F1-score de 0.58).

Tabela 5 – Métricas do classificador Soft Voting com painel Top 18.

Subtipo	Precisão	Recall	F1-Score	Suporte
CIN	0,78	0,85	0,81	41
EBV	0,80	0,80	0,80	10
GS	0,58	0,58	0,58	12
MSI	1,00	0,78	0,88	18
Acurácia		0.79		
Macro AVG	0,79	0,75	0,77	81
Weighted AVG	0,80	0,79	0,79	81

Fonte: O autor (2025).

A Tabela 6 apresenta as média das métricas de desempenho do modelo, que alcançou uma acurácia média de 0.74 ± 0.01 na classificação dos subtipos de câncer gástrico em 81 pacientes. A média ponderada do F1-score foi de 0.75 ± 0.02, indicando uma boa capacidade geral de classificação.

Tabela 6 – Métricas da média dos classificadores utilizando o painel Top 18.

Subtipo	Precisão	Recall	F1-Score	Suporte
CIN	0.92 ± 0.03	0.64 ± 0.03	0.75 ± 0.02	41
EBV	0.62 ± 0.02	0.79 ± 0.03	0.70 ± 0.02	10
GS	0.47 ± 0.03	0.81 ± 0.04	0.59 ± 0.02	12
MSI	0.87 ± 0.08	0.88 ± 0.03	0.87 ± 0.04	18
Acurácia		0.74 ± 0.01		
Macro AVG	0.71 ± 0.01	0.78 ± 0.01	0.73 ± 0.01	81
Weighted AVG	0.80 ± 0.02	0.74 ± 0.01	0.75 ± 0.02	81

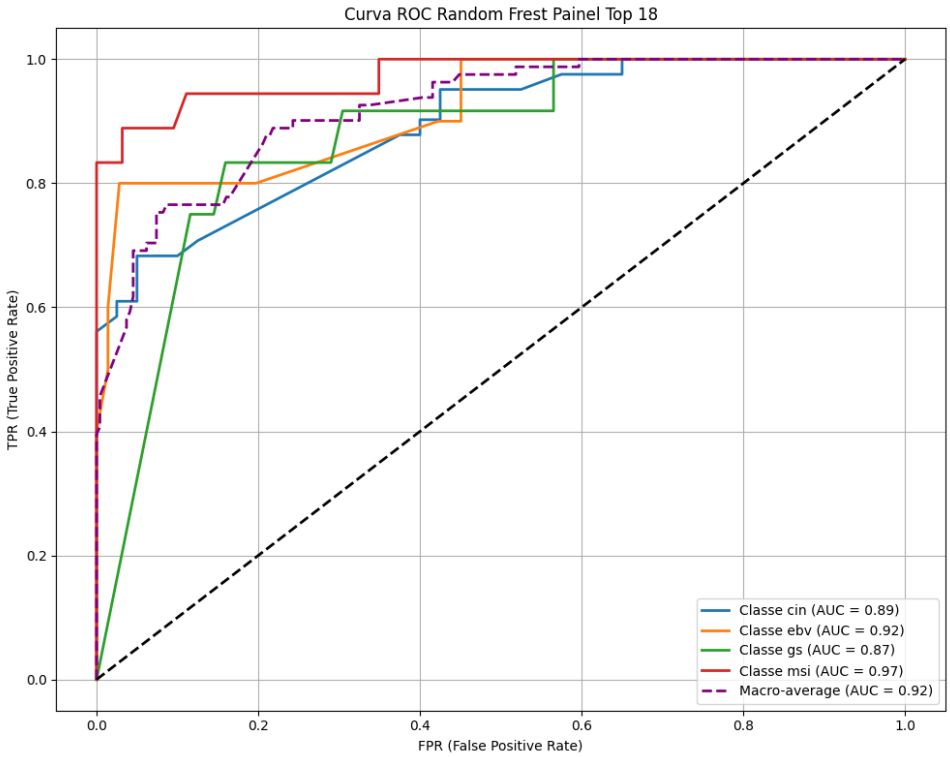
Fonte: O autor (2025).

Analisando os subtipos individualmente, o modelo demonstrou um desempenho excelente para MSI, com alta precisão (0.87 ± 0.08) e recall (0.88 ± 0.03), sugerindo que suas previsões para esta classe são muito confiáveis. O subtipo CIN também obteve alta precisão (0.92 ± 0.03), mas com um recall moderado (0.64 ± 0.03), o que significa que, embora as classificações CIN fossem geralmente corretas, o modelo não conseguiu identificar todos os casos dessa classe. Em contrapartida, os subtipos GS e EBV apresentaram um padrão inverso, com recall elevado (0.81 e 0.79, respectivamente) e a precisão mais baixa

(0.47 e 0.62), indicando que o modelo identificou a maioria dos casos dessas classes, mas ao custo de um número maior de falsos positivos.

Por fim, é possível observar através da Figura 4 que o modelo obteve um auto desempenho nos subtipos EBV e MSI com AUC de 0,92 e 0,97 respectivamente, assim como uma média macro de 0,92. Os subtipos CIN e GS apresentam um desempenho significativo com AUC de 0,89 e 0,87 respectivamente.

Figura 4 – AUC-ROC do Random Forest com painel genético Top 18



Fonte: O autor (2025).

É importante notar que este desempenho foi alcançado utilizando um painel selecionado de 18 genes: TP53, ARID1A, MUC16, ZBTB41, GGNBP2, PIK3CA, SYT17, MEF2C, MUC6, RNF213, SEC31A, BOC, CDH18, NFASC, BHLHB9, FAS, HERC2 e SYNE1. A seleção deste painel genético, que inclui genes com papéis conhecidos na tumorigênese gástrica como ‘TP53‘ e ‘PIK3CA‘, foi crucial para a capacidade preditiva do classificador, permitindo que o modelo discernisse entre os subtipos moleculares com a acurácia reportada.

6.4.4 Resultados Ensemble Multimodal (MobileNet + Random Forest)

Os resultados do ensemble multimodal GSubType-GenoVision, combinando MobileNet (visão computacional) e Random Forest (RF), foram obtidos a partir de 20 modelos treinados com k-fold cross-validation (k=10). As métricas de precisão, recall e F1-score foram calculadas para dois métodos de agregação: *hard voting* (votação majoritária) e *soft voting* (média de probabilidades).

Tabela 7 – SubtGenoVision 9

Subtipo	**Precisão**	**Recall**	**F1-Score**	**Suporte**
CIN	0,77	0,83	0,80	41
EBV	0,73	0,80	0,76	10
GS	0,58	0,58	0,58	12
MSI	0,86	0,67	0,75	18
Acurácia		0.75		
Macro AVG	0,74	0,72	0,72	81
Weighted AVG	0,76	0,75	0,75	81

Fonte: O autor (2025).

Tabela 8 – AUR-ROC: Wang et al. (2022) Vs. modelos desenvolvidos.

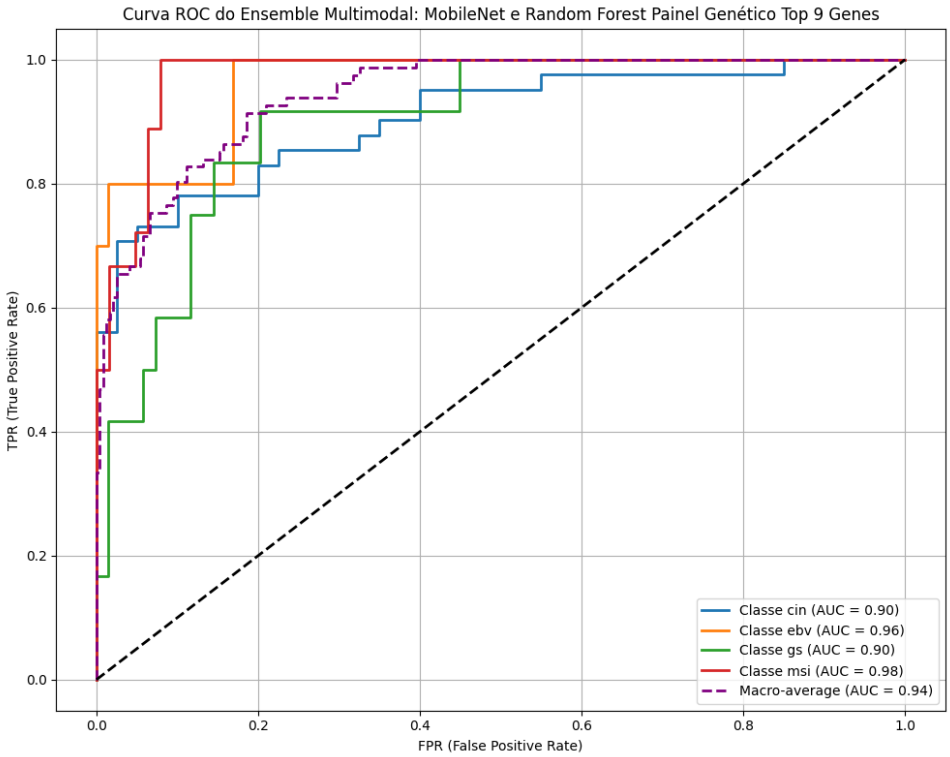
Subtipo	**Wang et al. (2022)**	**G.Subt Forest9**	**G.SubtGeno Vision9**	**G.SubtGeno Vision18**
CIN	0,890	0,87	0,90	0,91
EBV	0,764	0,90	0,96	0,98
GS	0,897	0,84	0,90	0,90
MSI	0,898	0,96	0,98	0,99
Macro AVG	0,840	0,89	0,94	0,95

Fonte: O autor (2025).

Note que a Tabela 8 apresenta uma comparação direta da performance AUR-ROC do modelo G.SubtGenoVision18 com o trabalho de referência de Wang et al. (2022). O modelo proposto demonstra uma superioridade consistente, alcançando uma média macro (Macro AVG) de 0,95, um avanço significativo em relação aos 0,84 reportados pelo modelo base.

Já na Tabela 8 comparamos a performance AUR-ROC dos modelos desenvolvidos com o trabalho de referência de Wang et al. (2022). Observa-se uma melhoria progressiva e

Figura 5 – AUC-ROC G.SubtGenoVision 9



Fonte: O autor (2025).

substancial em todos os modelos propostos, que superam consistentemente o modelo base.

O modelo G.SubtForest9 baseado apenas em características genômicas, já demonstra um avanço significativo, elevando a média macro (Macro AVG) de 0,84 para 0,89. Este modelo se destaca na classificação dos subtipos EBV (0,90) e MSI (0,96), superando com folga o trabalho de referência.

A integração de dados histopatológicos e genômicos nos modelos G.SubtGenoVision resulta em um salto de performance ainda maior. O G.SubtGenoVision9 alcança uma média macro de 0,94, mostrando a força da fusão de dados. O desempenho na classificação de EBV (0,96) e MSI (0,98) é notavelmente alto, indicando uma sinergia eficaz entre as fontes de informação.

Finalmente, o G.SubtGenoVision18 que utiliza um painel genético expandido, firma-se como o modelo de melhor desempenho, atingindo uma média macro de 0,95. Ele obtém resultados quase perfeitos para os subtipos MSI (0,99) e EBV (0,98), e melhora ou iguala a performance em todas as outras classes. Essa evolução demonstra que a combinação de dados de visão computacional com um painel genético otimizado é uma estratégia robusta

e superior para a classificação dos subtipos moleculares do adenocarcinoma gástrico.

O modelo G.SubtGenoVision (Artigo 4) integra visão computacional (MobileNetV2) e dados genéticos (Random Forest com 9 genes) e alcançou um AUC-ROC médio de 0.94 na classificação dos quatro subtipos moleculares (CIN, MSI, EBV, GS) [6].

Para estabelecer uma comparação rigorosa com a literatura unimodal, utilizamos o desempenho do modelo DEMoS de Wang et al. (2022), que se baseia exclusivamente em imagens histopatológicas. Os resultados de DEMoS no nível do paciente são os seguintes (WANG et al., 2022):

A análise comparativa no nível de paciente demonstra a robustez do sistema multimodal, especialmente na melhoria da discriminação das classes minoritárias:

Comparado com Wang et al. (2022) (WANG et al., 2022), o G.SubtGenoVision 6 melhorou os valores AUC-ROC no nível do paciente nos seguintes subtipos:

- CIN (Instabilidade Cromossômica): O G.SubtGenoVision obteve um AUC-ROC de 0.90, superando o AUC-ROC de 0.890 de Wang et al. (2022) em 0.010 ponto percentual. Embora ambos os modelos demonstrem alta capacidade preditiva para esta classe majoritária, a integração multimodal fornece um ganho marginal, mas consistente

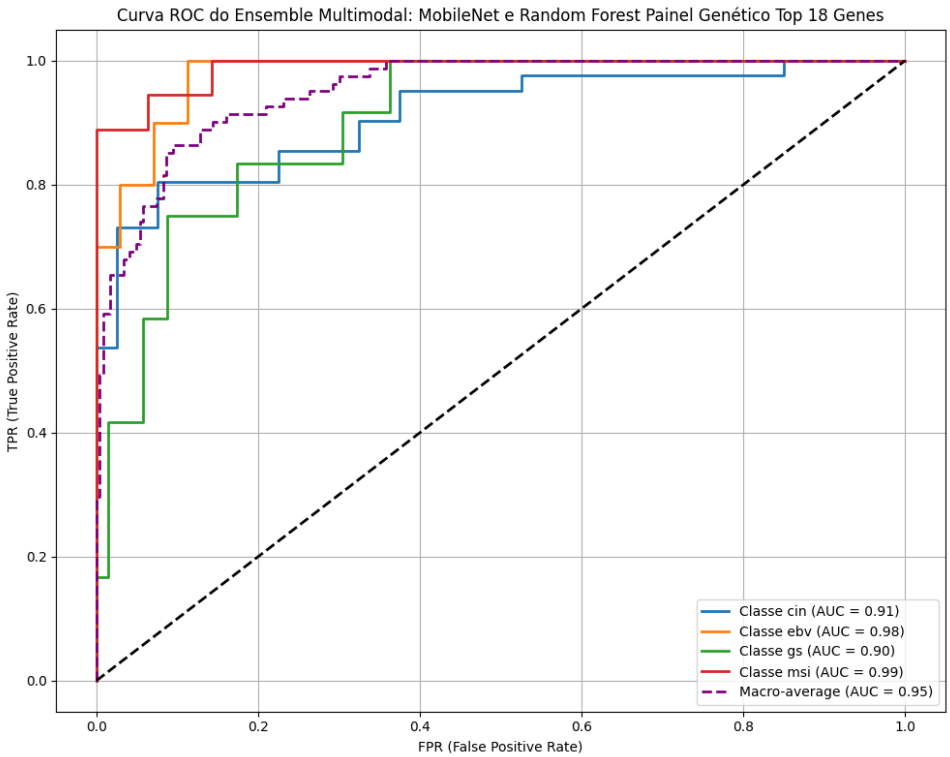
- EBV (Vírus Epstein-Barr): O G.SubtGenoVision alcançou um AUC-ROC de 0.96, representando uma melhoria acentuada de 0.196 ponto percentual sobre o AUC-ROC de 0.764 de Wang et al. (2022). Este ganho substancial é particularmente relevante, pois EBV é uma classe minoritária (9% dos casos [**??**]), e o alto desempenho no AUC-ROC (0.96) sugere que a inclusão do componente genético do Random Forest (otimizado para SNVs influentes) complementa eficazmente a predição da MobileNetV2, que isoladamente alcançou AUC-ROC de 0.95 para EBV

- GS (Genomicamente Estável): O G.SubtGenoVision atingiu um AUC-ROC de 0.90, um aumento de apenas 0.003 ponto percentual em relação ao AUC-ROC de 0.897 de Wang et al. (2022). Isso indica que, para o GS, a abordagem unimodal de Wang já era altamente eficiente, e a integração multimodal manteve esse alto nível de discriminação.

- MSI (Instabilidade Microssatélite): O G.SubtGenoVision obteve um AUC-ROC de 0.98, superando o AUC-ROC de 0.898 de Wang et al. (2022) em 0.082 ponto per-

centual. Este resultado no MSI (uma classe minoritária) demonstra que a fusão de dados genéticos, especialmente através do Random Forest otimizado [5], forneceu a informação essencial para alcançar uma capacidade de discriminação próxima da perfeição (0.98).

Figura 6 – AUC-ROC MobileNet-V2



Fonte: O autor (2025).

Enquanto a média macro do AUC-ROC de Wang et al. (2022) no nível do paciente (calculada em ≈ 0.862) já era considerada um desempenho favorável, o G.SubtGenoVision elevou a métrica global para 0.94.

O aprimoramento do G.SubtGenoVision (0.94) sobre DEMoS (0.862) reside na sua capacidade de balancear o desempenho: ele mantém a alta precisão alcançada pela Visão Computacional em classes como CIN e GS, enquanto usa a informação molecular para amplificar o desempenho nas classes minoritárias MSI e EBV, superando a tendência dos modelos unimodais em classes desbalanceadas.

Tabela 9 – Comparação de AUC-ROC por Subtipo Individual.

Subtipo Molecular	G.SubtGenoVision 18 (Multimodal AUC-ROC [6])	(Wang et al., 2022) (Unimodal AUC-ROC)	Ganho
CIN	0,91	0,890	2,0 %
EBV	0,97	0,764	20,6 %
GS	0,90	0,897	0,3 %
MSI	0,99	0,898	2,0 %

Fonte: O autor (2025).

6.5 CONCLUSÃO

O sistema proposto, **G.SubtGenoVision**, demonstrou desempenho superior aos modelos existentes na literatura e mostrou-se eficiente na classificação dos subtipos moleculares do adenocarcinoma gástrico. A combinação de imagens histopatológicas e dados genéticos foi eficaz em superar limitações de abordagens unimodais, ampliando a acessibilidade à classificação molecular e proporcionando avanços significativos para o diagnóstico e tratamento do câncer gástrico.

7 DISCUSSÃO GERAL

O Artigo 1 “G.SubtVision: Subtipagem Molecular do Câncer Gástrico com Métodos de Ensemble de Redes Neurais Convolucionais (CNNs)” estabelece uma base teórica robusta ao reconstruir com inteligência artificial o conhecimento em patologia, transitando de análises morfológicas baseadas em imagens histopatológicas (Lauren, 1965; OMS, 2019) para a integração de dados moleculares (TCGA, 2014) e, finalmente, ao ensemble de múltiplas arquiteturas de redes neurais convolucionais proposto. A implementação do G.SubtVision, com três arquiteturas de CNNs (MobileNetV2, ShuffleNet e GoogLeNet) é treinada em 263 casos com 476 lâminas, resultou em F1-score macro de 0.48 e precisão macro de 0.56, superando o HEAL de Wang et al. (2022) em 4% na média e 14% em MSI, com desempenho excepcional em EBV (0.48 vs. 0.09). Esses resultados se alinham com os de Flinner et al. (2022), que enfrentaram desafios em subtipos minoritários como GS e MSI. A revisão de Cifici et al. (2022) reforça que a falta de validação externa limita modelos anteriores (FLINNER et al., 2022; CIFCI; FOERSCH; KATHER, 2022).

Essa abordagem técnica encontra respaldo na hipótese de que a combinação de arquiteturas complementares explora a diversidade de features extraídas, superando a dependência de um único modelo, como o EfficientNet em Wang. A superioridade em classes desbalanceadas, como EBV, pode ser atribuída em parte à capacidade do ensemble de ponderar predições de forma distribuída, reduzindo o impacto de amostras escassas.

Além disso, o uso de supervisão molecular, em vez de critérios morfológicos humanos, alinha-se à tendência de Kather et al. (2019) e Coudray et al. (2018) em outras topografias, sugerindo que o G.SubtVision estabelece um marco inicial para a patologia digital no câncer gástrico. Por outro lado, quando se observam os resultados da reprodução do modelo de Wang et al. 2022 com EfficientNet, realizado, percebe-se que a distribuição de casos no grupo teste pode ter sido uma parte importante do melhoramento observado (KATHER et al., 2020; COUDRAY et al., 2018).

Na prática, o G.SubtVision é uma ferramenta que pode auxiliar na triagem de casos para exames moleculares específicos que possam fortalecer o poder preditivo dos subtipos moleculares. Com a expansão da patologia digital com escâner de lâminas acessíveis e plataformas online para diagnóstico patológico, o G.SubtVision pode ser extremamente acessível mesmo em laboratórios com recursos limitados, pois as imagens histopatológi-

cas já fazem parte da rotina e com o scanner de lâminas inteiras digitalizando a imagem que pode ser processada em nuvem.

Já o Artigo 2 “Redes neurais convolucionais classificam subtipo molecular do câncer gástrico em dataset tubular-controlado” aprofunda a investigação ao validar a capacidade das CNNs de identificar fenótipos profundos, definidos como atributos profundos (deep features) diretamente associados a padrões genômicos. O dataset tubular-controlado, composto por 22 casos tubulares do TCGA-STAD categorizados como CIN ou não-CIN, foi projetado para testar a hipótese de que a classificação seria indireta via distribuição histopatológica. Resultados mostram que NASNet-Mobile alcançou AUROC global >0.72, enquanto MobileNetV2 apresentou precisão 0.62, recall 0.73, F1 0.66 e AUROC 0.64, comparáveis ao dataset geral (0.63/0.69/0.66/0.69), rejeitando essa hipótese. Essa consistência alinha-se a Kather et al. (2019), que identificaram MSI em HE, porém esse teste com um dataset construído com tipo histopatológico homogêneo para verificar se mesmo diante dessas condições a CNN continuaria a classificar os subtipos. A persistência da predição em CIN reforça essa capacidade.

Tecnicamente, as métricas do dataset tubular-controlado indicam que a performance independe do tipo histológico tubular, mas decorre de padrões moleculares subjacentes. A comparação com o dataset geral mostra estabilidade (diferença de AUROC <0.05), sugerindo robustez em cenários controlados, uma limitação reconhecida por Wang et al. (2022) em classes desbalanceadas.

Na prática, isso significa que as redes neurais não estão apenas "imitando"o que já se sabe, mas descobrindo novas pistas no tecido, mesmo em amostras pequenas ou heterogêneas. Para um patologista, isso reduz a chance de erros em biópsias desafiadoras do ponto de vista de categorização histopatológica. Já que o reconhecimento dos padrões subjacentes aos subtipos moleculares parece ter independência da associação desses com o tipo histopatológico. (WANG et al., 2022)

A transição para o Artigo 3, “G.SubtForest: Classificador de Subtipos Moleculares do CA Gástrico com TCGA via Random Forest e Painéis Otimizados”, representa um avanço na modalidade genômica, complementando as contribuições visuais. Tecnicamente, o artigo propõe o G.SubtForest, baseado em Random Forest aplicado a 18.600 variantes de nucleotídeo único (SNV) não sinônimas do TCGA-STAD, com k-fold (k=10) para treinamento e SHAP para identificar genes influentes, resultando em painéis otimizados de 18 genes (adequado a NGS) e 9 genes (adequado a IHC), alcançando AUC-ROC média de

0.91 e 0.89, respectivamente.

Essa abordagem supera painéis IHC propostos por Kim et al. (2016), que atingem AUC-ROC 0.73, com ganhos de +0.18 macro, alinhando-se ao TCGA (2014) em genes como TP53 e ARID1A, mas identificando inéditos como ZBTB41 em EBV e GGNBP2 em MSI, hipotetizando que a teoria dos jogos via SHAP captura interações colaborativas melhor que inferências indutivas, mitigando desbalanceamentos em subtipos como EBV (9% no TCGA).

A correlação com artigos anteriores é evidente: enquanto o G.SubtVision (Artigo 1) e o dataset tubular-controlado (Artigo 2) capturam deep features visuais, o G.SubtForest adiciona dados genéticos, preparando a integração multimodal no Artigo 4, superando limitações de imagens isoladas como em Flinner et al. (2022), onde IA em HE alcança AUC-ROC 0.80 para CIN, vs. 0.88 aqui (CHEN et al., 2025; LIU et al., 2025; KIM et al., 2016; Cancer Genome Atlas Research Network, 2014).

Na prática, isso é como criar um "menu genético"acessível: em vez de sequenciar tudo, caro e lento como nos métodos multiômicos do TCGA, usa-se um painel compacto de 9 genes para IHC em rotinas hospitalares, facilitando estratificação rápida de pacientes para terapias-alvo, como inibidores de checkpoint em MSI, sem sobrecarregar sistemas de saúde em países em desenvolvimento. Expandindo nos subitens da Conclusão, o pré-processamento de dados do TCGA (VARSCAN), mutações não-sinônimas, alinha-se à necessidade de foco em variantes funcionais, como destacado por Kim et al. (2016) em IHC, mas a abordagem via SNV aqui supera em precisão (AUC-ROC 0.91 vs. 0.73), hipotetizando que a redução de ruído melhora discriminação em MSI hipermutados, onde hipermutações demandam filtragem robusta.

O treinamento de 10 modelos Random Forest (Tópico 3.2) com k=10 mitiga desbalanceamentos, alcançando macro F1 0.75 para TOP 36, superior a Lian et al. (2020) em metilação (0.70), sugerindo que ensembles elevam robustez em GS (F1 0.60), melhor que Flinner et al. (2022) em IHC (0.50) (FLINNER et al., 2022; KIM et al., 2016).

A revisão de importância de variáveis via SHAP e teoria dos jogos, resulta em painéis com TP53 e ARID1A, alinhados ao TCGA (2014), mas hipotetizando que SHAP supera importância permutada ao capturar colaborações, explicando genes inéditos como ZBTB41 (1º em EBV, ligado a repressão epigenética). A aplicação de SHAP para pontuação ponderada (Tópico 3.4) constrói painéis acessíveis, superando Kim (2016) em AUC-ROC (+0.16), priorizando MUC16 (ausente em Kim), melhorando acessibilidade em contextos limitados,

como Röcken (2022) em biomarcadores preditivos. Genes relevantes inéditos (Tópico 3.5), como GGNBP2 (2º em MSI) e BHLHB9 (5º em MSI), contrastam com foco em TP53 do TCGA (2014), hipotetizando papéis em instabilidade e proliferação, expandindo repertório para terapias.

Os sistemas G.SubtForest 18 (NGS) e 9 (IHC) (Tópico 3.8) obtêm AUC-ROC 0.91 e 0.89, superando Flinner (2022) em CIN (0.80), hipotetizando que SHAP-RF eleva translação, como Lian (2020) em metilação, facilitando rotina clínica. Na prática, é como oferecer "opções econômicas": o painel 9 para IHC em hospitais sem NGS, reduzindo custos vs. multiômicos (TCGA, 2014), superando inferências probabilísticas de Kim (2016), promovendo medicina de precisão acessível (FLINNER et al., 2022; LIU et al., 2024).

A culminância da tese se dá no Artigo 4, G.SubtGenoVision: Sistema ensemble multimodal para classificação dos subtipos moleculares do adenocarcinoma gástrico com Imagens histopatológicas e painel de mutações, que integra as modalidades visuais e genômicas desenvolvidas nos artigos anteriores. Tecnicamente, o G.SubtGenoVision concatena 10 modelos MobileNetV2 (desenvolvidos para o artigo 1) com o G.SubtForest 9 (do Artigo 3), utilizando dados do TCGA-STAD (476 lâminas e SNVs de 18.600 genes em 290 pacientes), com pré-processamento de *tiling* e normalização de cor para imagens, e tabulação por caso/gene para SNVs (SANDLER et al., 2018; Cancer Genome Atlas Research Network, 2014).

O G.SubtGenoVision alcançou AUC-ROC médio de 0.94, demonstrando superioridade notável sobre as abordagens unimodais. A comparação no nível de paciente entre as abordagens de Visão Computacional isolada (G.SubtVision/Artigo 1 e Wang et al./DEMoS) e o G.SubtGenoVision multimodal revela o impacto da integração:

- Comparado com G.SubtVision (CNN ensemble unimodal, AUC-ROC 0.85), o G.SubtGenoVision melhorou os valores AUROC:
  - CIN: (0.82 para 0.90), resultando em uma melhoria de 0.08 ponto percentual.
  - EBV: (0.94 para 0.96), um ganho modesto de 0.02 ponto percentual, reforçando a alta capacidade preditiva da MobileNetV2 isolada para esta classe
  - GS: (0.71 para 0.90), um ganho substancial de 0.19 ponto percentual. Este aumento significativo demonstra que a adição da informação genética (Random Forest) foi crucial para resgatar a baixa performance da CNN (MobileNetV2) na classificação do subtipo Genomicamente Estável

  - MSI: (0.86 para 0.98), uma melhoria robusta de 0.12 ponto percentual.

- Comparado com Wang et al. (2022) (DEMoS, AUC-ROC médio ≈ 0.86), o G.SubtGenoVision melhorou os valores AUROC no nível do paciente:

  - CIN: (0.890 para 0.90), com um ganho marginal de 0.010 ponto percentual, refletindo a eficácia de ambas as abordagens em uma classe majoritária
  - (0.764 para 0.96), um ganho substancial de 0.196 ponto percentual. Este ganho valida a estratégia multimodal, superando a dificuldade do DEMoS em classes minoritárias
  - GS: (0.897 para 0.90), com um ganho mínimo de 0.003 ponto percentual. Embora Wang et al. (2022) já tivessem alcançado alta performance para GS, o G.SubtGenoVision conseguiu igualar esse patamar robusto
  - MSI: (0.898 para 0.98), com um ganho de 0.082 ponto percentual, destacando a capacidade da fusão de dados genéticos em MSI, um subtipo hipermutado

Na prática, é como montar uma triangulação de modos de conhecimento para o diagnóstico: imagens e genes colaboram para dar respostas mais precisas e rápidas, ajudando a identificar subtipos que guiam tratamentos sem depender de testes caros e demorados.

Por fim, é importante enfatizar que a tecnologia necessária à translação dos resultados foi desenvolvida em paralelo, com projeto de inovação (descrito no tópico "Outras Produções Durante o Vínculo com o PPGGBM") que culminou na criação do Pathoscope. Uma solução nacional integral em patologia digital com escâner de lâminas, plataforma online e modelos de IA, viabilizando aplicação prática dos avanços já que a infraestrutura acelera translação clínica, integrando ensembles multimodais em fluxos reais, permitindo o avanço para estudos clínicos prospectivos.

Tabela 1 – AUC-ROC: Wang et al. (2022) vs. Modelos desenvolvidos em nível de pacientes

SM	Wang et al. (2022)	G.SubtVision	G.SubtForest9	G.SubtGenoVision9	G.SubtGenoVision18
CIN	0,890	0,82	0,87	0,90	0,91
EBV	0,764	0,94	0,90	0,96	0,98
GS	0,897	0,71	0,84	0,90	0,90
MSI	0,898	0,86	0,96	0,98	0,99
Macro AVG	0,84	0,85	0,89	0,94	0,95

Fonte: O autor (2025).

Tabela 2 – Precisoin: Wang et al. (2022) vs. Modelos desenvolvidos em nível de pacientes

SM	Wang et al. (2022)	G.SubtVision	G.SubtForest9	G.SubtiGenoVision9	G.SubtiGenoVision18
CIN	0,58	0,57	0,90±0,04	0,77	0,78
EBV	1,00	1,00	0,47±0,04	0,73	0,80
GS	0,83	0,62	0,48±0,00	0,58	0,58
MSI	0,65	1,00	0,83±0,08	0,86	1,00
Macro AVG	0,77	0,80	0,67±0,02	0,74	0,79

Fonte: O autor (2025).

8 CONCLUSÃO

A presente tese:

- Desenvolveu **G.SubtVision**: sistema preditivo para imagens histopatológicas na identificação de subtipos moleculares:
  - Propôs um novo método, fundamentado em ensemble de três redes neurais convolucionais: **G.SubtVision**, demonstrando resultados superiores à literatura (Artigo 1).
  - Demonstrou que as redes neurais foram capazes de classificar o subtipo molecular, mesmo em dataset tubular-controlado, apontando para a identificação de fenótipo profundo (atributo profundo decorrente de treinamento supervisionado com dados genéticos) (Artigo 2).
- Desenvolveu **G.SubtForest**, sistemas preditivos no modo de conhecimento das mutações somáticas de variação de nucleotídeo único (SNV):
  - Identificou painéis de mutações para a classificação do subtipo molecular com 18 e 9 genes que podem ser aplicados de acordo com a acessibilidade a métodos diagnósticos.
  - Identificou genes relevantes na diferenciação entre subtipos moleculares que não estavam previamente descritos na literatura sobre câncer gástrico.
  - Desenvolveu sistemas preditivos **G.SubtForest 18** para painel com NGS e **G.SubtForest 9** para painel com imuno-histoquímica (Artigo 3).
- Desenvolveu o **G.SubtGenovision**, sistema de ensemble multimodal integrando padrões de imagens histopatológicas e mutações somáticas:
  - Desenvolveu sistema integrando em ensemble 10 modelos de MobileNetV2 e G.SubtForest 9, obteendo resultados significativamente superiores aos descritos na literatura. Contribuindo para a ampliação da acessibilidade à classificação dos subtipos moleculares.
- Desenvolveu sistema aplicável na prática médica já que foi realizado paralelamente ao projeto de inovação que resultou no **Pathoscope**, que confere a infraestrutura

tecnológica necessária para viabilizar, em curto prazo, a aplicação prática dos avanços aqui descritos, por meio de scanner de lâminas, plataforma diagnóstica online e modelos de inteligência artificial, conforme descrito no tópico "Produção Durante o Vínculo com o PPGGBM".

REFERÊNCIAS

AMIN, M. B.; EDGE, S. B.; GREENE, F. L.; BYRD, D. R.; BROOKLAND, R. K.; WASHINGTON, M. K.; GERSHENWALD, J. E.; COMPTON, C. C.; HESS, K. R.; SULLIVAN, D. C.; JESSUP, J. M.; BRIERLEY, J. D.; GASPAR, L. E.; SCHILSKY, R. L.; BALCH, C. M.; WINCHESTER, D. P.; ASARE, E. A.; MADERA, M.; GRESS, D. M.; MEYER, L. R. *AJCC Cancer Staging Manual*. 8.. ed. New York: Springer, 2017. ISBN 978-3-319-40617-6.

BAAN, R. A.; STEWART, B. W.; STRAIF, K. (Ed.). *Tumour Site Concordance and Mechanisms of Carcinogenesis*. Lyon, France: International Agency for Research on Cancer, 2019. v. 165. (IARC Scientific Publications, v. 165). ISBN 978-92-832-2165-6. Disponível em: <https://publications.iarc.fr/Book-And-Report-Series/Iarc-Scientific-Publications/Tumour-Site-Concordance-And-Mechanisms-Of-Carcinogenesis-2019>.

BALTRUŠAITIS, T.; AHUJA, C.; MORENCY, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 41, n. 4, p. 755–779, 2018.

BARCHI, L. C.; RAMOS, M. F. K. P.; YAGI, O. K.; MUCERINO, D. R.; BRESCIANI, C. J. C.; JÚNIOR, U. R.; ANDREOLLO, N. A.; ASSUMPÇÃO, P. P.; WESTON, A. C.; NETO, R. C. Brazilian Gastric Cancer Association guidelines (part 1): an update on diagnosis, staging, endoscopic treatment and follow-up. *ABCD. Arquivos Brasileiros de Cirurgia Digestiva (São Paulo)*, SciELO Brasil, v. 33, p. e1535, 2020.

BENITES-GOÑI, H.; CABRERA-HINOJOSA, D.; LATORRE, G.; HERNANDEZ, A. V.; UCHIMA, H.; RIQUELME, A. OLGA and OLGIM staging systems on the risk assessment of gastric cancer: a systematic review and meta-analysis of prospective cohorts. *Therapeutic Advances in Gastroenterology*, SAGE Publications Sage UK: London, England, v. 18, p. 17562848251325461, 2025.

BOSSUYT, P. M.; REITSMA, J. B.; BRUNS, D. E.; GATSONIS, C. A.; GLASZIOU, P. P.; IRWIG, L.; LIJMER, J. G.; MOHER, D.; RENNIE, D.; VET, H. C. W. D. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Radiology*, Radiological Society of North America, v. 277, n. 3, p. 826–832, 2015.

BREIMAN, L.; FRIEDMAN, J.; OLSHEN, R.; STONE, C. *Classification and Regression Trees*. Belmont, CA, USA: Wadsworth International Group, 1984. ISBN 978-0412048418.

Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, Nature Publishing Group, v. 513, n. 7517, p. 202, 2014.

CHEN, B.; HUAN, L.; LU, J.; YUAN, J. Shapley Additive Explanations (SHAP) based feature selection reveals CXCL14 as a key immune-related gene in predicting idiopathic pulmonary fibrosis. *Frontiers in Medicine*, Frontiers, v. 12, p. 1608078, 2025.

CHEN, Y.; WANG, L.; LI, Z. MUC16 mutation is associated with tumor mutation burden and stromal infiltration in gastric cancer. *Frontiers in Genetics*, v. 12, p. 743519, 2021.

CIFCI, D.; FOERSCH, S.; KATHER, J. N. Artificial intelligence to identify genetic alterations in conventional histopathology. *The Journal of Pathology*, Wiley Online Library, v. 257, n. 4, p. 430–444, 2022.

COGLIANO, V. J.; BAAN, R.; STRAIF, K.; GROSSE, Y.; SECRETAN, B.; GHISSASSI, F. E. Preventable exposures associated with human cancers. *Journal of the National Cancer Institute*, Oxford University Press, v. 103, n. 24, p. 1827–1839, dez. 2011. Disponível em: <https://doi.org/10.1093/jnci/djr483>.

COMPTON, C. C.; ROBB, J. A.; ANDERSON, M. W.; BERRY, A. B.; BIRDSONG, G. G.; BLOOM, K. J.; BRANTON, P. A.; CROTHERS, J. W.; CUSHMAN-VOKOUN, A. M.; HICKS, D. G. Preanalytics and precision pathology: pathology practices to ensure molecular integrity of cancer patient biospecimens for precision medicine. *Archives of pathology & laboratory medicine*, the College of American Pathologists, v. 143, n. 11, p. 1346–1363, 2019.

COUDRAY, N.; OCAMPO, P. S.; SAKELLAROPOULOS, T.; NARULA, N.; SNUDERL, M.; FENYÖ, D.; MOREIRA, A.; RAZAVIAN, N.; TSIRIGOS, A. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*, Nature Publishing Group, v. 24, n. 10, p. 1559–1567, 2018.

DENG, J.; DONG, W.; SOCHER, R.; LI, L.-J.; LI, K.; FEI-FEI, L. Imagenet: A large-scale hierarchical image database. In: IEEE. *2009 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.], 2009. p. 248–255.

DURAN, R. The accumulation of CNVs over time plays a key role in genomic instability associated with cellular aging and malignant transformation. *Epigenomics Insights*, v. 15, p. 72–79, 2023.

EVANS, A. J.; SALAMA, M. E.; HENRICKS, W. H.; PANTANOWITZ, L. Implementation of whole slide imaging for clinical purposes: issues to consider from the perspective of early adopters. *Archives of Pathology & Laboratory Medicine*, College of American Pathologists, v. 142, n. 3, p. 306–312, 2018.

FLINNER, N.; GRETSER, S.; QUAAS, A.; BANKOV, K.; STOLL, A.; HECKMANN, L. E.; MAYER, R. S.; DOERING, C.; DEMES, M. C.; BUETTNER, R. Deep learning based on hematoxylin–eosin staining outperforms immunohistochemistry in predicting molecular subtypes of gastric adenocarcinoma. *The journal of pathology*, Wiley Online Library, v. 257, n. 2, p. 218–226, 2022.

FUKAYAMA, M.; RUGGE, M.; WASHINGTON, M. (Ed.). *WHO Classification of Tumours of the Digestive System*. 5th. ed. [S.l.]: International Agency for Research on Cancer (IARC), 2019. ISBN 978-92-832-4499-8.

GENECARDS. *GeneCards: the human gene database*. 2025. Acesso em: 23 set. 2025. Disponível em: <https://www.genecards.org>.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>.

GUO, B. *Prediction of cancer driver genes with graph neural networks: a case study in glioblastoma*. Tese (Doutorado) — Universidade Federal do Rio Grande do Sul, 2022. Disponível em: <https://www.lume.ufrgs.br/handle/10183/261761>. Acesso em: 2 set. 2025.

HAMASHIMA, C. Update version of the Japanese Guidelines for Gastric Cancer Screening. *Japanese Journal of Clinical Oncology*, Oxford University Press, v. 48, n. 7, p. 673–683, 2018.

HE, P.; LI, X.; ZOU, D.; TANG, F.; CHEN, H.; LI, Y. Environmental factors inducing gastric cancer: insights into risk and prevention strategies. *Discover oncology*, Springer, v. 16, n. 1, p. 25, 2025.

HOWSON, C. P.; HIYAMA, T.; WYNDER, E. L. The decline in gastric cancer: epidemiology of an unplanned triumph. *Epidemiologic Reviews*, v. 8, p. 1–27, 1986. Disponível em: <https://doi.org/10.1093/oxfordjournals.epirev.a036428>.

HUANG, X. Deep learning-based histopathological analysis for cancer diagnosis: Progress and challenges. *Cancer Letters*, v. 497, p. 1–9, 2021.

IIZUKA, O.; KANAVATI, F.; KATO, K.; RAMBEAU, M.; ARIHIRO, K.; TSUNEKI, M. Deep learning models for histopathological classification of gastric and colonic epithelial tumours. *Scientific reports*, Nature Publishing Group UK London, v. 10, n. 1, p. 1504, 2020.

Instituto Nacional de Câncer José Alencar Gomes da Silva. *Estimativa 2023: incidência de câncer no Brasil*. Rio de Janeiro: INCA, 2022. Disponível em: <https://www.inca.gov.br/estimativa>. Acesso em: 19 maio 2025.

International Agency for Research on Cancer. *Preamble to the IARC Monographs (2019)*. 2019. Disponível em: <https://monographs.iarc.who.int/wp-content/uploads/2019/07/Preamble-2019.pdf>. Acesso em: 2 set. 2025.

JANG, B. et al. Machine-learning model derived gene signature predictive of paclitaxel response in gastric cancer. *Gut*, v. 71, n. 4, p. 676–685, 2023.

JANG, H.-J.; SONG, I.-H.; LEE, S.-H. Deep learning for automatic subclassification of gastric carcinoma using whole-slide histopathology images. *Cancers*, MDPI, v. 13, n. 15, p. 3811, 2021.

KAMANGAR, F.; DAWSEY, S. M.; BLASER, M. J. Opposing risks of gastric cardia and noncardia gastric adenocarcinomas associated with Helicobacter pylori seropositivity. *Journal of the National Cancer Institute*, v. 98, p. 1445–1452, 2006. Disponível em: <https://doi.org/10.1093/jnci/djj393>.

KANAVATI, F.; ICHIHARA, S.; RAMBEAU, M.; IIZUKA, O.; ARIHIRO, K.; TSUNEKI, M. Deep learning models for gastric signet ring cell carcinoma classification in whole slide images. *Technology in Cancer Research & Treatment*, SAGE Publications Sage CA: Los Angeles, CA, v. 20, p. 15330338211027901, 2021.

KATHER, J. N.; PEARSON, A. T.; HALAMA, N.; JÄGER, D.; KRAUSE, J.; LOOSEN, S. H.; MARX, A.; BOOR, P.; TACKE, F.; NEUMANN, U. P.; LUEDDE, T.; HÜLSKEN, J.; ZÖLLNER, F. G.; FERBER, D.; BOOR, P. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *The Lancet Oncology*, Elsevier, v. 21, n. 11, p. 1618–1629, 2020.

KIM, H. S.; SHIN, S.-J.; BEOM, S.-H.; JUNG, M.; CHOI, Y. Y.; SON, T.; KIM, H.-I.; CHEONG, J.-H.; HYUNG, W. J.; NOH, S. H. Comprehensive expression profiles of gastric cancer molecular subtypes by immunohistochemistry: implications for individualized therapy. *Oncotarget*, Impact Journals, LLC, v. 7, n. 28, p. 44608, 2016.

KIN, T.-a.; WANG, X.; STOYANOVA, R.; FLICK, M. J.; LUIDENS, M. K.; REFIGUREDO, M.; MORAN, M.; FALCIONI, R.; LANGUINO, L. R.; PRENDERGAST, G. C. et al. Comprehensive molecular characterization of clinical outcomes in human gastric cancer using a TCGA-based classification. *Cancer Research*, v. 76, n. $14_{S}upplement, p.209 - -209, 2016.$

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015. Disponível em: <https://arxiv.org/abs/1412.6980>.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2012. v. 25, p. 1097–1105.

LAURÉN, P. The two histological main types of gastric carcinoma: diffuse and so-called intestinal-type carcinoma. *Acta Pathologica et Microbiologica Scandinavica*, Wiley Online Library, v. 64, n. 1, p. 31–49, 1965.

LEE, J.; KIM, H.; LEE, H. S. ARID1A mutation in gastric cancer: A systematic review and meta-analysis. *Cancers*, v. 15, n. 3, p. 895, 2023.

LEE, J.; WARNER, E.; SHAIKHOUNI, S.; BITZER, M.; KRETZLER, M.; GIPSON, D.; PENNATHUR, S.; BELLOVICH, K.; BHAT, Z.; GADEGBEKU, C. et al. Unsupervised machine learning for identifying important visual features through bag-of-words using histopathology data from chronic kidney disease. *Scientific reports*, Nature Publishing Group UK London, v. 12, n. 1, p. 4832, 2022.

LI, J.; HE, L.; ZHANG, X.; LI, X.; WANG, L.; ZHU, Z.; SONG, K.; WANG, X. GCclassifier: An R package for the prediction of molecular subtypes of gastric cancer. *Computational and Structural Biotechnology Journal*, Elsevier, v. 23, p. 752–758, 2024.

LINS, J. A. B.; MENEZES, F. S.; LIRA, P. V.; MELO, H. M. A. Objetivos de desenvolvimento sustentável: Uma abordagem multidisciplinar dos desafios e soluções. In: ____. [S.l.]: Editora UFPE, 2024. cap. Saúde, Bem-Estar e Busca por Sentido: Pensando Criativamente as Interrelações para a Prática da Sustentabilidade.

LIU, X. et al. Multiomics integration and machine learning reveal prognostic signature for gastric cancer. *Scientific Reports*, v. 14, n. 82233, p. 1–15, 2024.

LIU, X.; TAO, P.; SU, H.; LI, Y. Machine learning-random forest model was used to construct gene signature associated with cuproptosis to predict the prognosis of gastric cancer. *Scientific Reports*, Nature Publishing Group UK London, v. 15, n. 1, p. 4170, 2025.

LOPES, M. Multiomics allowed mapping molecular and metabolic patterns characteristic of tumor subtypes, promoting a more precise direction of therapies. *Advances in Molecular Oncology*, v. 40, p. 304–321, 2024.

MACENKO, M.; NIETHAMMER, M.; MARRON, J. S.; BORLAND, D.; WOOSLEY, J. T.; GUAN, X.; SCHMITT, C.; THOMAS, N. E. A method for normalizing histology slides for quantitative analysis. In: IEEE. *2009 IEEE international symposium on biomedical imaging: from nano to macro*. [S.l.], 2009. p. 1107–1110.

MARTEL, C.; PARSONNET, J. Stomach cancer. In: THUN, M.; LINET, M. S.; CERHAN, J. R.; HAIMAN, C. A.; SCHOTTENFELD, D. (Ed.). *Cancer Epidemiology and Prevention*. 4.. ed. [S.l.]: Oxford University Press, 2018. p. 593–610.

MAZUREK, M.; SZEWC, M.; SITARZ, M. Z.; DUDZIŃSKA, E.; SITARZ, R. Gastric cancer: an up-to-date review with new insights into early-onset gastric cancer. *Cancers*, MDPI, v. 16, n. 18, p. 3163, 2024.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, v. 5, p. 115–133, 1943. Disponível em: <https://link.springer.com/article/10.1007/BF02478259>.

MITHANY, R. H.; SHAHID, M. H.; MANASSEH, M.; SAEED, M. T.; ASLAM, S.; MOHAMED, M. S.; DANIEL, N. Gastric cancer: a comprehensive literature review. *Cureus*, Cureus, v. 16, n. 3, 2024.

MONJO, A.; GARCÍA, A.; HERNÁNDEZ, F.; LÓPEZ, J. Molecular supervision for training deep neural networks for histopathological image recognition. *Nature Scientific Reports*, Nature Publishing Group, v. 12, n. 1, p. 1234–1240, 2022.

NAGTEGAAL, I. D.; ODZE, R. D.; KLIMSTRA, D.; PARADIS, V.; RUGGE, M.; SCHIRMACHER, P.; WASHINGTON, K. M.; CARNEIRO, F.; CREE, I. A. The 2019 who classification of tumours of the digestive system. *Histopathology*, Wiley Online Library, v. 76, n. 2, p. 182–188, 2020.

National Cancer Institute. *Genomic Data Commons Data Portal (GDC Data Portal)*. 2025. Disponível em: <https://portal.gdc.cancer.gov/>. Acesso em: 2 set. 2025. RRID:SCR_014514.

OLEFSON, S.; MOSS, S. F. Obesity and related risk factors in gastric cardia adenocarcinoma. *Gastric cancer*, Springer, v. 18, n. 1, p. 23–32, 2015.

OpenSlide. *OpenSlide Python API*. 2023. <https://openslide.org/api/python/>. Acessado em: 09 set. 2025.

PARK, J.-Y.; CHOI, Y.-D.; KIM, T.-H. The hedgehog receptor BOC is a prognostic marker and correlates with microsatellite instability in gastric cancer. *Journal of Pathology and Translational Medicine*, v. 57, n. 4, p. 221–230, 2023.

PEARCE, N.; BLAIR, A.; VINEIS, P.; AHRENS, W.; ANDERSEN, A.; ANTO, J. M.; ARMSTRONG, B. K.; BACCARELLI, A. A.; BELAND, F. A.; BERRINGTON, A. IARC monographs: 40 years of evaluating carcinogenic hazards to humans. *Environmental health perspectives*, NLM-Export, v. 123, n. 6, p. 507–514, 2015.

PEZZOTTI, L. Hallmarks of aging-based dual-purpose disease and age-associated targets predicted using PandaOmics AI-powered discovery engine. *Aging*, Impact Journals LLC, v. 14, n. 6, p. 2475–2506, 2022.

PIMENTEL-NUNES, P.; LIBÂNIO, D.; MARCOS-PINTO, R.; AREIA, M.; LEJA, M.; ESPOSITO, G.; GARRIDO, M.; KIKUSTE, I.; MEGRAUD, F.; MATYSIAK-BUDNIK, T. Management of epithelial precancerous conditions and lesions in the stomach (MAPS II): European Society of gastrointestinal endoscopy (ESGE), European Helicobacter and microbiota Study Group (EHMSG), European Society of pathology (ESP), and Sociedade

Portuguesa de Endoscopia Digestiva (SPED) guideline update 2019. *Endoscopy*, © Georg Thieme Verlag KG, v. 51, n. 04, p. 365–388, 2019.

POWELL, J.; MCCONKEY, C. C. Increasing incidence of adenocarcinoma of the gastric cardia and adjacent sites. *British Journal of Cancer*, v. 62, p. 440–443, 1990. Disponível em: <https://www.nature.com/articles/bjc1990164>.

PRECHELT, L. Early stopping—but when? In: *Neural Networks: Tricks of the trade*. [S.l.]: Springer, 1997. p. 55–69.

RAJAGOPALAN, H.; NOWAK, M. A.; VOGELSTEIN, B.; LENGAUER, C. The significance of unstable chromosomes in colorectal cancer. *Nature Reviews Cancer*, Nature Publishing Group, v. 3, p. 695–701, 2003.

RAMOS, M. F. K. P. *Caracterização dos subtipos moleculares do câncer gástrico por expressão gênica e proteica*. Tese (Doutorado) — Universidade de São Paulo, 2019.

RODRIGUES, T.; SILVA, G.; FONSECA, M. Slide-seq: A technique for spatial transcriptomics using location-specific barcodes. *Cell*, Cell Press, v. 178, n. 5, p. 1001–1012, 2019.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, v. 65, n. 6, p. 386–408, 1958. Disponível em: <https://doi.org/10.1037/h0042519>.

RUGGE, M.; CORREA, P.; MARIO, F. D.; EL-OMAR, E.; FIOCCA, R.; GEBOES, K.; GENTA, R. M.; GRAHAM, D. Y.; HATTORI, T.; MALFERTHEINER, P. OLGA staging for gastritis: a tutorial. *Digestive and liver disease*, Elsevier, v. 40, n. 8, p. 650–658, 2008.

SANDLER, M.; HOWARD, A.; ZHU, M.; ZHMOGINOV, A.; CHEN, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [s.n.], 2018. p. 4510–4520. Disponível em: <https://arxiv.org/abs/1801.04381>.

SENA, E.; ROCHA, H. Implementação de rede neural convolucional para predição de COVID-19 através de imagens de raio x. In: *Congresso Brasileiro de Informática na Saúde*. [S.l.: s.n.], 2021. p. 1–8.

SHAH, D.; BENTREM, D. Environmental and genetic risk factors for gastric cancer. *Journal of surgical oncology*, Wiley Online Library, v. 125, n. 7, p. 1096–1103, 2022.

SHAPLEY, L. S. A value for n-person games. In: KUHN, H. W.; TUCKER, A. W. (Ed.). *Contributions to the Theory of Games*. Princeton: Princeton University Press, 1953. v. 2, p. 307–317.

SHAUKAT, A.; WANG, A.; ACOSTA, R. D.; BRUINING, D. H.; CHANDRASEKHARA, V.; CHATHADI, K. V.; ELOUBEIDI, M. A.; FANELLI, R. D.; FAULX, A. L.; FONKALSRUD, L. The role of endoscopy in dyspepsia. *Gastrointestinal endoscopy*, Elsevier, v. 82, n. 2, p. 227–232, 2015.

SOBIN, L.; GOSPODAROWICZ, M.; WITTEKIND, C. *TNM Classification of Malignant Tumours*. [S.l.]: Wiley-Blackwell, 2017.

SOBIN, L.; GOSPODAROWICZ, M.; WITTEKIND, C. *TNM Classification of Malignant Tumours*. [S.l.]: UICC, 2017.

SRINIVASAN, M.; SEDMAK, D.; JEWELL, S. Effect of fixatives and tissue processing on the content and integrity of nucleic acids. *The American journal of pathology*, Elsevier, v. 161, n. 6, p. 1961–1971, 2002.

STELZER, G.; ROSEN, N.; PLASCHKES, I.; ZIMMERMAN, S.; TWIK, M.; FISHILEVICH, S.; STEIN, T. I.; NUDEL, R.; LIEDER, I.; MAZOR, Y.; KAPLAN, S.; DAHARY, D.; WARSHAWSKY, D.; GUAN-GOLAN, Y.; KOHN, A.; RAPPAPORT, N.; SAFRAN, M.; LANCET, D. The genecards suite: From gene data mining to disease genome sequence analyses. *Current Protocols in Bioinformatics*, v. 54, n. 1, p. 1.30.1–1.30.33, 2016. Acessado em 03/09/2025; base de dados disponível em https://www.genecards.org/.

STOLTE, M.; MEINING, A. The updated Sydney system: classification and grading of gastritis as the basis of diagnosis and treatment. *Canadian Journal of Gastroenterology and Hepatology*, Wiley Online Library, v. 15, n. 9, p. 591–598, 2001.

SUNG, H.; FERLAY, J.; SIEGEL, R. L.; LAVERSANNE, M.; SOERJOMATARAM, I.; JEMAL, A.; BRAY, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, Wiley, Hoboken, v. 71, n. 3, p. 209–249, fev. 2021.

The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, v. 513, p. 202–209, 2014.

THOMPSON, S. L.; BAKHOUM, S. F.; COMPTON, D. A. Mechanisms of chromosomal instability. *Current Biology*, Elsevier, v. 20, n. 6, p. R285–R295, 2010.

VAPNIK, V. *The nature of statistical learning theory*. [S.l.]: Springer science & business media, 2013.

WANG, L.; CHEN, Y.; LI, Z.; LIU, J.; LI, A.; WANG, H. The role of tp53 mutations in gastric cancer: A review of the literature. *Frontiers in Oncology*, v. 14, 2024. ISSN 2234-943X.

WANG, Y.; HU, C.; KWOK, T.; BAIN, C. A.; XUE, X.; GASSER, R. B.; WEBB, G. I.; BOUSSIOUTAS, A.; SHEN, X.; DALY, R. J. DEMoS: a deep learning-based ensemble approach for predicting the molecular subtypes of gastric adenocarcinomas from histopathological images. *Bioinformatics*, Oxford University Press, v. 38, n. 17, p. 4206–4213, 2022.

WILLIAMS, C.; PONTÉN, F.; MOBERG, C.; SÖDERKVIST, P.; UHLÉN, M.; PONTÉN, J.; SITBON, G.; LUNDEBERG, J. A high frequency of sequence alterations is due to formalin fixation of archival specimens. *The American journal of pathology*, Elsevier, v. 155, n. 5, p. 1467–1471, 1999.

WU, Y.; XIE, L. AI-driven multi-omics integration for multi-scale predictive modeling of causal genotype-environment-phenotype relationships. *arXiv preprint arXiv:2407.06405*, 2024.

XU, S. et al. Forestsubtype: a cancer subtype identifying approach based on holistic tumor phenotype. *BMC Bioinformatics*, v. 24, n. 1, p. 1–17, 2023.

YURKOVICH, J. T.; TIAN, Q.; PRICE, N. D.; HOOD, L. A systems approach to clinical oncology uses deep phenotyping to deliver personalized care. *Nature Reviews Clinical Oncology*, Nature Publishing Group UK London, v. 17, n. 3, p. 183–194, 2020.

ZHANG, H.; LI, Y.; WANG, C.; LIU, Y.; SONG, Y.; WANG, Z. ZBTB41 is a prognostic biomarker and therapeutic target in epstein-barr virus-associated gastric cancer. *Journal of Medical Virology*, v. 95, n. 1, p. e28438, 2023.

9 PRODUÇÃO DURANTE O VÍNCULO COM O PPGGBM

Nesse capítulo serão encontradas outras produções realizadas durante o processo de doutoramento que não se relacionam diretamente com o encadeamento lógico dos objetivos principais da presente tese. Destaca-se o projeto de inovação diretamente associado a acessibilidade dos resultados da presente tese. Outras produções foram capítulo de livro publicado e experimentos realizados que não entraram na presente tese.

Coordenação de projeto de inovação: Sistema de detecção precoce do câncer por inteligência artificial.

O autor durante o desenvolvimento da presente tese escreveu e coordenou o projeto "Sistema de detecção precoce do câncer por inteligência artificial aplicada à patologia digital: prevenção do câncer de colo do útero e de estômago".

Submeteu o projeto de sua autoria à chamada pública para empresas do setor saúde desenvolverem tecnologia 4.0. imediatamente antes da seleção ao doutorado no PPGGBM. Dentre as tecnologias habilitadoras consideradas nessa chamada pública como tecnologias 4.0 estavam a Inteligência Artificial e a computação em nuvem.

A digitalização da patologia passou a ser possível há menos de uma década. Embora a fotografia digital de partes pequenas da lâmina uma de cada vez em conformidade com o aumento do microscópio já existissem desde o surgimento da fotografia digital, foi apenas com o aumento do poder computacional e do avanço nos algorítimos que a costura automatizada dos pedaços para compor digitalmente a lâmina inteira passou a ser possível. Por ter menos de uma década a patologia digital está ainda em sua infância, sem ainda haver em 2020 uma solução nacional. Os custos com scanners de lâminas inteiras é ainda elevado a ponto de ser proibitivo para pequenos laboratórios. Os aparelhos disponíveis apenas para compra elevam em demasia o custo da imagem em um mercado de margens pequenas e em tendência de queda por competição acirrada de preços entre os laboratórios de patologia. Essa tem sido a principal barreira à ampla adoção da tecnologia. O acesso a essa tecnologia, portanto, permanece restrito às grandes empresas. O que é uma limitação à acessibilidade das Inteligências Artificiais (IA) que já demonstraram eficiência significativa no reconhecimento de imagens como maior probabilidade de pre-

sença de um tipo específico de tumor e, portanto, com potencial de otimizar o suporte aos patologistas.

Foi aprovado na linha temática SAÚDE 4.0 propondo o desenvolvimento de uma solução que integre a captura de imagens de lâminas inteiras com câmeras digitais acopladas a microscópios automatizados usando processamento interno de costuras das imagens microscópicas para a exportação da *WSI* imagem de lâmina inteira. Sistema web de visualização online com segurança e integração com Sistemas de Informação Laboratorial. Processamento de imagens em nuvem com modelos de inteligência artificial para a seleção de campos de maior probabilidade de câncer, aumentando a sensibilidade. Um salto rumo a um futuro mais acessível e eficiente para a patologia digital.

Objetivo Geral Desenvolver uma solução integral de patologia digital ampliada por inteligência artificial.

Objetivos específicos: 1- Desenvolver *scanner* de lâminas para digitalização de lâminas histológicas e citológicas 2- Desenvolver sistema *web* para patologia digital que permita interface com o usuário patologista e colaboração entre usuários patologistas 3- Treinar inteligência computacional para identificar alterações morfológicas em amostras de citologia oncótica vaginal e histologia de mucosa gástrica; Desenvolver aplicação considerando a experiência do patologista;

4- Validar a aplicação em grupos de pacientes em estudos retrospectivos e prospectivos controlados; Validar a aplicação em ambiente operacional; Implantar a aplicação validada na rotina operacional de laboratórios patologia credenciados.

Da perspectiva de Experiência do Usuário, a construção da plataforma do PATHOSCOPE foi realizada em conjunto com um time de patologistas do Ampliar o que permitiu uma melhor compreensão de usabilidade bem como a definição das melhores ferramentas de manipulação de imagens para a área.

A inovação desenvolvida aqui é disruptiva no âmbito nacional pois altera significativamente a maneira como as análises patológicas são feitas e atende às principais demandas estratégicas dos laboratórios especializados na área. Ainda hoje a informatização é parcial nos laboratórios de patologia, sendo usada apenas como ferramenta de gestão e edição de texto, todo o trabalho de reconhecimento de padrões de imagens é manual, utilizando microscópio ótico e sem qualquer sistema de apoio ao diagnóstico. O presente projeto desenvolveu um sistema de apoio diagnóstico ao câncer desenvolvendo a primeira tecnologia de patologia digital ampliada por inteligência computacional no país. A proposta do

projeto redesenha todo o modelo de negócio, aumentando a escalabilidade e atendendo uma demanda reprimida para diagnóstico de câncer, em um momento que, caso nada seja feito, se tornará crítica, não havendo número suficiente de patologista para os diagnósticos necessários. A inteligência artificial aplicada à patologia digital utilizando computação em nuvem é uma inovação disruptiva no âmbito nacional e internacional, já que está presente na prática médica, somente em uns poucos centros nos países desenvolvidos.

Níveis de Maturidade Tecnológica (TRLs) abrangidos nível 3 Imagens processada já validada nível 4 Curadoria das imagens ra ser realizada com banco balanceado e marcado nível 5 resultados estatísticos favoráveis nível 6 testes em ambiente virtual nível 7 protótipo em ambiente operacional validado

A capacidade do novo processo alterar o paradigma técnico-econômico vigente está em que infelizmente o patologista atua na maior parte do tempo como microscopista. O conhecimento fisiopatológico crescente já há muito tempo é extenso demais para que os clínicos e cirurgiões os dominem sem auxílio especializado. Por outro lado o médico patologista frequentemente não recebe as informações completas dos casos e se encontra sobrecarregado com o rastreamento microscópico manual. O rastreamento pode ser comparado a procurar por uma moeda em um gramado. Para poder afirmar a ausência de moeda é necessário um rastreamento minucioso. Afirmar a ausência de pequenas células em uma imagem com milhares de quadros é muitas vezes uma atividade, monótona, delongada e extenuante que exige longos períodos de imobilidade diante do microscópio. Esse paradigma leva ao aumento de falso-negativo por pequenas alterações passarem desapercebidas.

Nos laboratórios pequenos e médios, que é o caso de todos do norte-nordeste, o patologista lauda grande diversidade de topografias, exercendo a chamada patologia geral. Fazem todo o rastreamento sozinhos com grande consumo de tempo. O que é um incentivo à redução do tempo despendido em uma lâminas. Seria de grande interesse dos patologistas um processo de trabalho que permitisse uma maior especialização dos exames laudados e o auxílio de IA supervisionada, permitindo mais precisão nos exames e menor estresse para o patologista. O avanço possibilitará a patologia digital nacional e assim permitirá melhor organização do fluxo na rede de laboratórios credenciados. Permitirá que os patologistas possam trabalhar a distância, com colaboração técnico-científica entre serviços de diversas regiões, permitirá assim que os patologistas atuem de maneira mais especializada, mesmo pertencendo a serviços menores. Promoverá aumento da qualidade

nos diagnósticos devido ao uso de ferramentas de computação para desenhar, destacar imagens, medir, contar e cooperar em tempo real com outros patologistas. Padrões que não são fáceis de serem percebidos por seres humanos, mas que são reconhecidos por inteligência artificial são um exemplo inequívoco da contribuição das redes neurais para o diagnóstico.

A patologia digital vai acessibilizar diagnósticos mais rápidos e precisos à milhares de pessoas, promovendo a colaboração entre médicos patologistas nas regiões e médicos patologistas hiper especializados em grande variedade de topografias. Rede neural desenvolvida para citologia auxiliará no diagnóstico precoce de casos de milhares de casos de câncer de colo do útero evitando morbidade e mortalidade Rede neural desenvolvida para histologia auxiliará no diagnóstico precoce de milhares de casos de câncer do estômago. O scanner de Imagens de Lâminas Inteiras devido a computação em nuvem poderá ser um serviço acessível aos laboratórios do norte-nordeste que continuarão competitivos. O novos processos organizarão os dados nos arquivos da beneficiária proponente para deixá-los limpos e preparados para novos desenvolvimentos de redes neurais.

Impacto Tecnológico: 1. Avanço no âmbito do diagnóstico patológico com utilização da patologia digital com computação em nuvem e assistida por inteligência artificial 2. Introdução na rotina diagnóstica Sistema inteligente de apoio ao diagnóstico precoce do câncer de estômago 3. Introdução na rotina diagnóstica Sistema inteligente de apoio ao diagnóstico precoce do câncer de colo do útero 4. Avanço da Telemedicina no apoio ao diagnóstico anatomopatológico do câncer. 5. Redução de doenças ocupacionais da coluna nos médicos patologistas com desenvolvimento de estação de trabalho ergonomicamente apropriada.

O autor da presente atuou como primeiro autor e coordenador geral do projeto, coordenando o desenvolvimento do *scanner* de lâminas (microscópio automatizado), do sistema web de visualização de *WSI* imagens de lâminas inteiras, do desenvolvimento dos modelos de visão computacional e da coordenação financeira, de contratações e demissões de desenvolvedores de tecnologia da informação, da prestação de contas.

O resultado desse trabalho foi a criação da primeira empresa especializada em patologia digital do Brasil. A PATHOSCOPE, comprometida com a ampliação da visão da patologia por inteligência artificial para o diagnóstico mais precisos e velozes.

CAPITULO PUBLICADO: SAÚDE, BEM-ESTAR E BUSCA POR SENTIDO: PENSANDO CRIATIVAMENTE AS INTERRELAÇÕES PARA A PRÁTICA DA SUSTENTABILIDADE

Capítulo de livro indexado e com comitê de avaliação publicada pela editora UFPE na série livro texto: Objetivos do desenvolvimento sustentável, uma abordagem multidisciplinar dos desafios e soluções (LINS et al., 2024)

Esse capítulo fala sobre a importância da promoção da saúde no contexto do envelhecimento populacional, destacando estratégias para garantir uma longevidade ativa e saudável. Ele aborda a diferença entre expectativa de vida e tempo de saúde, enfatizando o papel dos hábitos saudáveis e da prevenção no envelhecimento biológico. Além disso, apresenta a promoção da saúde como uma abordagem integrada, que vai além do tratamento de doenças, abrangendo o bem-estar físico, mental, social e comunitário. O capítulo também incentiva o uso do pensamento criativo e da colaboração interdisciplinar para enfrentar os desafios de saúde e alcançar as metas de desenvolvimento sustentável, promovendo empreendimentos transformadores e sustentáveis na sociedade.