Universidade Estadual de Feira de Santana

Programa de Pós-Graduação em Ciência da Computação

Avaliação de Aprendizado Auto-Supervisionado na Classificação de Glomérulos Renais em Cenário de Forte Desbalanceamento de Dados

Gabriel Silva de Azevedo

Feira de Santana

Fevereiro 2026

Universidade Estadual de Feira de Santana

Programa de Pós-Graduação em Ciência da Computação

Gabriel Silva de Azevedo

Avaliação de Aprendizado Auto-Supervisionado na Classificação de Glomérulos Renais em Cenário de Forte Desbalanceamento de Dados

Dissertação apresentada à Universidade Estadual de Feira de Santana como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Orientador: Angelo Amâncio Duarte

Feira de Santana

2026

Ficha catalográfica - Biblioteca Central Julieta Carteado - UEFS

Azevedo, Gabriel Silva de
A987a Avaliação de aprendizado Auto-supervisionado na classificação de glomérulos renais em cenário de forte desbalanceamento de dados / Gabriel Silva de Azevedo. - 2026.
90f.: il.

Orientador: Angelo Amâncio Duarte

Dissertação (mestrado) - Universidade Estadual de Feira de Santana. Programa de Pós-Graduação em Ciência da Computação, 2026.

1. Aprendizado Auto-supervisionado. 2. Patologia computacional. 3. Histopatologia renal. 4. Classificação de imagens. 5. Deep learning. I. Duarte, Angelo Amâncio, orient. II. Universidade Estadual de Feira de Santana. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDU: 004.85:616.61

Rejane Maria Rosa Ribeiro – Bibliotecária CRB-5/695

Gabriel Silva de Azevedo

Avaliação de Aprendizado Auto-Supervisionado na Classificação de Glomérulos Renais em Cenário de Forte Desbalanceamento de Dados

Dissertação apresentada à Universidade Estadual de Feira de Santana como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Feira de Santana, 8 de dezembro de 2025

BANCA EXAMINADORA

gov.br Documento assinado digitalmente
ANGELO AMANCIO DUARTE
Data: 18/12/2025 16:31:30-0300
Verifique em https://validar.iti.gov.br

Angelo Amâncio Duarte (Orientador(a))
Universidade Estadual de Feira de Santana

gov.br Documento assinado digitalmente
ROBERTA BARBOSA OLIVEIRA
Data: 08/12/2025 19:43:36-0300
Verifique em https://validar.iti.gov.br

Roberta Barbosa Oliveira
Universidade de Brasília

gov.br Documento assinado digitalmente
MICHELE FULVIA ANGELO
Data: 10/12/2025 09:21:01-0300
Verifique em https://validar.iti.gov.br

Michele Fúlvia Angelo
Universidade Estadual de Feira de Santana

Abstract

The diagnosis of chronic kidney diseases through histopathological analysis involves challenges such as inter-observer variability and the high time required for analysis. From a computational perspective, the automatic modeling of these lesions is further hindered by severe class imbalance and the scarcity of representative samples of rare lesions. This dissertation investigates the use of Self-Supervised Learning (SSL) techniques to build an automatic classifier of glomerular lesions, aiming to overcome limitations of supervised methods that depend on large amounts of labeled data. A total of 12,524 renal tissue images were used, distributed across eleven distinct lesions and obtained from multiple histological stains: hematoxylin and eosin (HE), periodic acid–Schiff (PAS), and Azan–Mallory (Azan). SSL methods such as Self-Distillation with No Labels (DINO) and Simple Framework for Contrastive Learning of Visual Representations (SimCLR) were evaluated with different backbones (ResNet and Vision Transformers), considering data augmentation and balancing strategies. The results demonstrate that SSL methods outperformed supervised approaches in multiclass classification tasks under imbalanced data scenarios. The DINO method, in particular, stood out in this context, achieving an F1-micro score of 58.27% compared to 43.47% for the ImageNet-pretrained baseline. In contrast, SimCLR showed superior performance in binary classification for several classes, highlighting its discriminative power in specific lesion classification tasks.

Keywords: self-supervised learning, computational pathology, renal histopathology, image classification, deep learning.

Resumo

O diagnóstico de doenças renais crônicas por meio da análise histopatológica envolve desafios como a variabilidade interobservador e o elevado tempo de análise. Do ponto de vista computacional, a modelagem automática dessas lesões é adicionalmente dificultada pelo desbalanceamento severo entre classes e pela escassez de amostras representativas de lesões raras. Esta dissertação investiga o uso de técnicas de Aprendizado Auto-Supervisionado (Self-Supervised Learning, SSL) para construir um classificador automático de lesões glomerulares, buscando superar limitações dos métodos supervisionados que dependem de muitos dados rotulados. Foram utilizadas 12.524 imagens de tecido renal, distribuídas em onze lesões distintas e obtidas a partir de múltiplas colorações histológicas: hematoxilina e eosina (HE), ácido periódico de Schiff (PAS) e Azan–Mallory (Azan). Avaliaram-se métodos de SSL, como Self-Distillation with No Labels (DINO) e Simple Framework for Contrastive Learning of Visual Representations (SimCLR), com diferentes backbones (ResNet e Vision Transformers), considerando aumentos e balanceamento de dados. Os resultados demonstram que os métodos de SSL superaram as abordagens supervisionadas em tarefas de classificação multiclasse em cenários de dados desbalanceados. O método DINO, em particular, destacou-se nesse contexto, alcançando um F1-micro de 58,27% em comparação com 43,47% do baseline pré-treinado no ImageNet. Em contrapartida, o SimCLR apresentou um desempenho superior na classificação binária para diversas classes, evidenciando seu poder discriminativo em tarefas específicas de classificação de lesões.

Palavras-chave: aprendizado auto-supervisionado, patologia computacional, histopatologia renal, classificação de imagens, deep learning.

Prefácio

Esta dissertação de mestrado foi submetida à Universidade Estadual de Feira de Santana (UEFS) como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

A dissertação foi desenvolvida no Programa de Pós-Graduação em Ciência da Computação (PGCC), tendo como orientador o Prof. Dr. **Angelo Amâncio Duarte**.

"A busca da verdade começa com a dúvida."

– Platão

Sumário

Alinhamento com a Linha de Pesquisa

Linha de Pesquisa: *Software* e Sistemas Computacionais

A presente dissertação está alinhada à linha de pesquisa *Software* e Sistemas Computacionais, pois investiga o desenvolvimento e a aplicação de métodos avançados de aprendizado auto-supervisionado no contexto da classificação de imagens histopatológicas. O trabalho propõe soluções computacionais que envolvem desde a modelagem e implementação de arquiteturas de redes neurais profundas até a análise de desempenho e robustez em cenários reais. Com isso, contribui para o aprimoramento de sistemas inteligentes de apoio ao diagnóstico médico, evidenciando a relevância da pesquisa na construção de sistemas computacionais eficientes, escaláveis e voltados para aplicações críticas.

Produções Bibliográficas, Produções Técnicas e Premiações

Azevedo, G. S.; Santos, A. S.; de Oliveira, L. R.; dos Santos, W. L. C.; Duarte, A. A. (2025). Dealing With Strong Imbalance to Classifying Glomerular Lesions via Self Supervised Learning. In *Anais do 38th Conference on Graphics, Patterns and Images (SIBGRAPI) - Workshop of Works in Progress (WIP)*. Paper ID 277.

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

"A medicina é uma ciência da incerteza e uma arte da probabilidade."

– William Osler

As doenças renais crônicas afetam cerca de 10% da população mundial, representando um importante desafio de saúde pública (Haas et al., 2020). Dentro desse contexto, a análise histopatológica dos glomérulos renais desempenha papel central no diagnóstico das nefropatias, impactando diretamente as condutas terapêuticas adotadas.

Apesar de sua relevância, o diagnóstico histopatológico enfrenta obstáculos importantes. Entre os principais desafios estão a variabilidade interobservador que compromete a consistência entre diferentes especialistas, o tempo elevado de análise por amostra, e a dificuldade na identificação de lesões raras.

Adicionalmente, a coloração das lâminas histológicas, etapa essencial na preparação das amostras, introduz uma variação significativa entre os dados. Esse processo, embora fundamental para realçar estruturas específicas do tecido renal, pode variar de acordo com o protocolo adotado em cada laboratório. Tal variabilidade influencia diretamente na aparência das imagens, dificultando a padronização e, consequentemente, impactando a consistência dos diagnósticos. Para modelos computacionais, essa diversidade representa um desafio adicional, pois aumenta a complexidade dos dados usados no treinamento e compromete a capacidade de generalização das soluções desenvolvidas (Tiard et al., 2022).

Nesse cenário, o uso de técnicas de aprendizado de máquina tem se mostrado promissor, especialmente no contexto da Patologia Computacional — campo voltado ao desenvolvimento de soluções computacionais para a análise de imagens histopa-

tológicas. Iniciativas como o PathoSpotter[1] (Oliveira et al., 2022) exemplificam esse avanço, ao buscarem reduzir a subjetividade diagnóstica por meio de classificadores automatizados de lesões glomerulares. No entanto, a eficácia desses sistemas depende fortemente da disponibilidade de bases de dados extensivamente anotadas, o que representa um obstáculo significativo à sua implementação em larga escala.

A construção dessas bases de dados anotadas representa um processo particularmente oneroso para patologistas, demandando: (1) conhecimento especializado para identificação precisa de padrões histológicos, (2) análise morfológica detalhada de cada lâmina digitalizada, e (3) tempo considerável para marcação pixel-a-pixel de estruturas relevantes. O problema se agrava pela escassez de patologistas especializados e pelos altos custos associados, tornando o desenvolvimento de sistemas robustos financeiramente inviável sem alternativas. Esse esforço manual se intensifica exponencialmente no contexto de lesões raras, onde a escassez de casos clinicamente confirmados obriga os especialistas a revisarem centenas de amostras para identificar poucos exemplos válidos, frequentemente com necessidade de consenso entre múltiplos avaliadores para garantir confiabilidade.

Adicionalmente, a natureza morfológica complexa e, por vezes, sutil de algumas lesões raras, que podem se assemelhar a padrões normais ou a outras patologias mais comuns, torna sua identificação um desafio significativo, mesmo para patologistas experientes. Para modelos computacionais, isso adiciona uma camada de dificuldade na extração de características discriminativas.

O aprendizado auto-supervisionado (*Self-Supervised Learning* — SSL) surge como uma abordagem promissora para superar esses desafios. Por não depender diretamente de rótulos manuais, o SSL permite o uso eficiente de grandes volumes de dados não anotados, extraindo representações úteis que podem ser aplicadas em tarefas específicas de classificação (Stacke et al., 2021b; Taleb et al., 2020). Embora já aplicado com sucesso em outras áreas da medicina que utilizam imagens, seu uso na nefropatologia digital ainda é incipiente, especialmente no que diz respeito à robustez frente à variação de colorações e à eficácia na detecção de lesões pouco frequentes (Huang et al., 2023; Wang et al., 2023).

Dessa forma, este trabalho propõe a investigação e o desenvolvimento de métodos baseados em SSL para a classificação automática de lesões em glomérulos renais, considerando os desafios relacionados à coloração histológica, ao desbalanceamento severo de dados e à necessidade de maior precisão na identificação de lesões raras. Este trabalho buscou preencher essa lacuna, oferecendo uma abordagem inovadora para superar as barreiras de anotação manual e da variabilidade de dados no domínio da nefropatologia.

As principais perguntas que norteiam este estudo são: Quais técnicas de SSL são mais eficazes na detecção de lesões glomerulares em cenários com desbalanceamento

[1] https://pathospotter.bahia.fiocruz.br/

de dados? E ainda, como diferentes estratégias de aumento de dados afetam o desempenho desses modelos na detecção de lesões raras?

1.1 Objetivos

1.1.1 Objetivo Geral

Investigar e implementar técnicas baseadas em aprendizado auto-supervisionado para a classificação de lesões em glomérulos renais em imagens histopatológicas com diferentes colorações, buscando superar as limitações das abordagens clássicas baseadas em redes convolucionais supervisionadas e alcançar melhor desempenho na detecção de lesões em cenários com forte desbalanceamento de classes.

1.1.2 Objetivos Específicos

- Estruturar e preparar um conjunto de dados contendo amostras representativas de doze tipos distintos de lesões glomerulares.
- Implementar e comparar diferentes técnicas de SSL em cenários com dados desbalanceados.
- Avaliar o impacto de estratégias de aumento de dados no desempenho dos modelos, com foco no F1-score.
- Analisar as causas das diferenças de desempenho entre diferentes técnicas de *SSL* aplicadas à classificação de lesões glomerulares, visando identificar fatores que possam orientar o aprimoramento dos métodos de classificação.
- Avaliar o impacto no uso de diferentes corantes para realizar o treinamento, averiguando se um corante em específico possui melhor desempenho que o uso de todos os corantes.

1.2 Contribuições

Este trabalho traz uma contribuição científica para o campo da Patologia Computacional e do Aprendizado Auto-supervisionado (SSL), ao realizar uma comparação sistemática de diferentes métodos avançados de SSL em um conjunto de dados histopatológicos desafiador, caracterizado pela presença de múltiplas colorações e pelo severo desbalanceamento entre classes. A relevância dessa contribuição está associada à complexidade diagnóstica das imagens analisadas, bem como à proposta de estratégias específicas para normalização digital e ponderação das classes, visando melhorar a generalização e robustez dos modelos implementados. Essas abordagens buscam preencher lacunas metodológicas frequentemente destacadas na literatura especializada. As principais contribuições incluem:

- Indicação do melhor entre os diferentes métodos de *SSL*, aplicados à tarefa de classificação de lesões glomerulares, analisando seu desempenho em cenários multiclasse e binários, e destacando sua robustez em condições de desbalanceamento severo de dados..
- Organização e curadoria de um subconjunto representativo de imagens histopatológicas com múltiplas colorações, focando em lesões de alta relevância clínica — como necrose fibrinoide, amiloidose, acúmulo de neutrófilos e depósitos hialinos — devido à sua gravidade, baixa incidência e complexidade diagnóstica.
- Proposição e avaliação de estratégias de normalização digital e técnicas de ponderação de classes, com o objetivo de reduzir os efeitos do desbalanceamento entre classes e da variabilidade introduzida pelas diferentes colorações.
- Disponibilização dos experimentos e códigos em um repositório de acesso aberto, promovendo a reprodutibilidade e incentivando a continuidade das pesquisas por parte da comunidade científica.

1.3 Organização do Trabalho

O Capítulo 2 apresenta uma revisão dos fundamentos teóricos sobre aprendizado auto-supervisionado e sua aplicação em histopatologia. O Capítulo 3 descreve os métodos propostos, a construção do *dataset* e os procedimentos de avaliação. Os resultados experimentais são discutidos no Capítulo 4, seguidos pela análise crítica no Capítulo 5. Por fim, o Capítulo 6 apresenta as conclusões alcançadas neste trabalho.

Capítulo 2

Revisão Bibliográfica

"Knowing how things work is the basis for appreciation, and is thus a source of civilized delight."
– William Safire

Este capítulo apresenta uma revisão abrangente dos fundamentos, avanços e aplicações do aprendizado auto-supervisionado (*Self-Supervised Learning* — SSL) no contexto da análise de imagens médicas, com ênfase na Patologia Computacional. Inicialmente, são discutidos os princípios teóricos que sustentam o SSL, seguidos pela descrição de suas principais abordagens e, por fim, suas aplicações específicas na patologia computacional.

2.1 Patologia Computacional

A Patologia Computacional é uma área interdisciplinar emergente que integra técnicas avançadas de ciência da computação, processamento de imagens digitais e análise quantitativa para revolucionar a prática da patologia diagnóstica. Esta área representa uma evolução natural da patologia tradicional, aproveitando o poder computacional para extrair informações objetivas e quantitativas de imagens histopatológicas que podem não ser perceptíveis ao olho humano. Diferentemente da patologia digital, que se limita à digitalização e visualização de lâminas, a patologia computacional engloba o desenvolvimento de algoritmos sofisticados capazes de realizar análises automatizadas, identificar padrões complexos e fornecer suporte inteligente ao diagnóstico médico.

O surgimento da patologia computacional foi impulsionado pela convergência de múltiplos fatores tecnológicos: o desenvolvimento de escâneres de alta resolução capazes de digitalizar lâminas inteiras (*Whole Slide Imaging*, WSI), o aumento exponencial da capacidade computacional, e os avanços em algoritmos de aprendizado

de máquina e visão computacional. Esta transformação digital permitiu que laboratórios de patologia processem e analisem grandes volumes de dados histológicos de forma sistemática e reprodutível, estabelecendo novos paradigmas para a prática diagnóstica.

A importância da patologia computacional transcende a mera automação de processos. Ela oferece a possibilidade de identificar biomarcadores quantitativos sutis, correlacionar características morfológicas com desfechos clínicos, e desenvolver sistemas de apoio à decisão que podem reduzir significativamente a variabilidade inter e intra-observador. Além disso, esta abordagem computacional permite a análise de características em escalas múltiplas, desde detalhes celulares microscópicos até padrões arquiteturais complexos em toda a extensão do tecido.

2.1.1 Processo de Obtenção das Amostras

O fluxo de trabalho em patologia computacional inicia-se com a obtenção adequada das amostras teciduais. No contexto da patologia renal, este processo geralmente envolve a realização de biópsias renais, procedimentos minimamente invasivos que permitem a coleta de fragmentos representativos do tecido renal para análise diagnóstica. A biópsia renal é frequentemente realizada sob orientação ultrassonográfica ou tomográfica, utilizando agulhas especializadas que preservam a arquitetura tecidual e minimizam artefatos que poderiam comprometer a análise posterior Haas et al. (2020); Fogo e Kashgarian (2003). Um exemplo visual deste procedimento pode ser observado na Figura 2.1.

O material obtido por meio de biópsia passa por um rigoroso processo de preparo histotécnico, como apresentado por Fogo e Kashgarian (2003); De Ruijter et al. (2015). Inicialmente, as amostras são fixadas em formalina tamponada a 10%, um processo que interrompe a degradação tecidual e preserva as estruturas celulares e extracelulares. Esta etapa de fixação é crítica, pois determina a qualidade final das imagens e a confiabilidade dos resultados analíticos, influenciando diretamente a consistência e a generalização de modelos computacionais. Subsequentemente, o tecido é processado através de etapas sequenciais de desidratação em álcoois de concentrações crescentes, clarificação em xilol, e emblocamento em parafina, resultando em blocos sólidos que podem ser seccionados com precisão micrométrica.

2.1.2 Técnicas de Coloração e Suas Aplicações

A coloração histológica representa uma etapa fundamental na preparação das amostras para análise computacional Gurina e Simms (2023); Fallatah et al. (2024). Diferentes técnicas de coloração revelam aspectos específicos da morfologia tecidual, sendo essencial para o diagnóstico diferencial de lesões renais. A coloração por Hematoxilina e Eosina (HE) constitui o método padrão, onde a hematoxilina cora núcleos celulares em azul-violeta e a eosina cora citoplasmas e matriz extracelular em tons rosados Gurina e Simms (2023); Fallatah et al. (2024). Esta coloração fornece uma

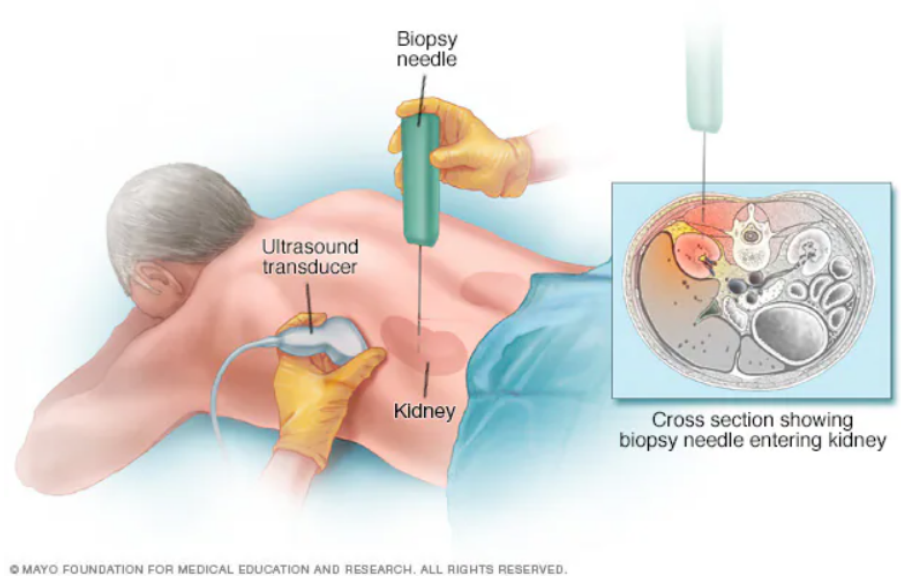


Figura 2.1: Representação esquemática do procedimento de biópsia renal, destacando a utilização de ultrassonografia para orientação da agulha de biópsia e a coleta de fragmentos do tecido renal. Este procedimento invasivo é crucial para a obtenção de amostras representativas para análise diagnóstica em patologia computacional Schwarz et al. (2022). Esta figura foi adaptada a partir de uma representação didática de procedimento clínico publicada pela Mayo Clinic may (2023)

visão geral da arquitetura tecidual e permite a identificação de alterações celulares básicas.

A coloração pelo Ácido Periódico de Schiff (PAS) é particularmente valiosa na patologia renal, pois destaca especificamente carboidratos complexos, glicoproteínas e mucinas presentes na membrana basal glomerular e na matriz mesangial Gurina e Simms (2023); Fallatah et al. (2024). Esta técnica permite a avaliação detalhada da integridade da barreira de filtração glomerular e a identificação de espessamentos ou duplicações da membrana basal, características de diversas glomerulopatias. A coloração AZAN (Azul de Anilina) ou Tricrômico de Masson complementa este arsenal, evidenciando fibras colágenas em azul brilhante e permitindo a quantificação precisa de fibrose intersticial e esclerose glomerular Gurina e Simms (2023); Fallatah et al. (2024).

Colorações especiais adicionais podem ser empregadas conforme a suspeita diagnóstica. A coloração com Vermelho Congo é específica para depósitos amiloides, exibindo birrefringência verde-maçã característica sob luz polarizada Gurina e Simms (2023); Fallatah et al. (2024). A coloração com Prata Metenamina destaca fungos e a membrana basal glomerular, sendo útil no diagnóstico diferencial de glomerulopatias Gurina e Simms (2023); Fallatah et al. (2024). A imuno-histoquímica, embora tecnicamente mais complexa, permite a identificação específica de proteínas e antígenos, fornecendo informações moleculares complementares à análise morfológica

tradicional Gurina e Simms (2023); Fallatah et al. (2024). Apesar de sua importância diagnóstica, a variabilidade inerente aos protocolos de coloração representa um desafio significativo para a padronização e a robustez de sistemas de patologia computacional Brixtel et al. (2022).

2.1.3 Digitalização e Aquisição de Imagens

O processo de digitalização em patologia computacional utiliza escâneres especializados de alta resolução, capazes de capturar lâminas histológicas completas com magnificações equivalentes a 20× ou 40× de microscopia óptica tradicional Kumar et al. (2020). Estes equipamentos, conhecidos como escâneres de lâmina inteira (WSI), empregam sistemas ópticos sofisticados com câmeras de alta sensibilidade e iluminação controlada para garantir uniformidade na aquisição das imagens. O resultado são arquivos digitais de grandes dimensões, frequentemente ultrapassando vários gigabytes por lâmina, que preservam todos os detalhes morfológicos necessários para análise diagnóstica Kumar et al. (2020).

Os escâneres WSI modernos incorporam tecnologias avançadas como autofoco automático, correção de distorções ópticas, e balanceamento automático de cores, garantindo qualidade consistente entre diferentes aquisições Kumar et al. (2020); Brixtel et al. (2022). Alguns sistemas permitem a captura em múltiplos planos focais (*z-stacking*), possibilitando a reconstrução tridimensional de estruturas teciduais complexas Brixtel et al. (2022). A resolução típica destes sistemas varia entre 0,25 a 0,5 micrômetros por pixel, permitindo a visualização de detalhes subcelulares e a análise quantitativa precisa de características morfométricas Kumar et al. (2020).

2.1.4 Desafios e Limitações na Patologia Computacional

A implementação efetiva da patologia computacional em nefropatologia enfrenta desafios técnicos e práticos significativos. A variabilidade técnica introduzida por diferentes protocolos de fixação, processamento e coloração pode resultar em inconsistências na aparência das imagens, afetando a reprodutibilidade dos algoritmos de análise. Esta heterogeneidade técnica é particularmente problemática quando sistemas computacionais treinados em um laboratório são aplicados em amostras de outros centros com protocolos diferentes.

O desbalanceamento inerente entre diferentes tipos de lesões constitui outro desafio fundamental. Lesões comuns como esclerose glomerular e fibrose intersticial são abundantemente representadas nas bases de dados, enquanto condições raras como amiloidose renal ou certas formas de glomerulonefrite aparecem com frequência muito baixa. Este desbalanceamento pode resultar em modelos computacionais com desempenho assimétrico, excellente para lesões comuns mas inadequado para condições raras que podem ter implicações clínicas críticas.

A complexidade morfológica das lesões renais apresenta desafios adicionais para análise automatizada. Muitas condições patológicas manifestam-se através de al-

terações sutis que requerem avaliação de múltiplas características morfológicas simultaneamente. Por exemplo, o diagnóstico diferencial entre diferentes formas de glomerulonefrite pode depender de nuances na distribuição de depósitos, padrões de proliferação celular e alterações da membrana basal que são facilmente percebidas por patologistas experientes mas desafiadoras para algoritmos computacionais.

A necessidade de anotações especializadas representa talvez o maior obstáculo para o desenvolvimento de sistemas de patologia computacional robustos. A criação de bases de dados anotadas requer tempo considerável de patologistas especialistas, profissionais cuja disponibilidade é limitada. Além disso, a concordância inter-observador para certas lesões pode ser subótima, introduzindo ruído nos dados de treinamento que pode comprometer o desempenho dos modelos resultantes.

2.2 Fundamentos do Aprendizado Auto-Supervisionado

O aprendizado auto-supervisionado (SSL — *Self-Supervised Learning*) é uma abordagem para treinamento de modelos de aprendizado profundo sem a necessidade de rótulos explícitos fornecidos por humanos. O modelo aprende representações através da resolução de tarefas auxiliares, conhecidas como tarefas pretextuais. Essas tarefas exploram propriedades intrínsecas dos dados, permitindo ao modelo adquirir uma compreensão estrutural e semântica que pode ser posteriormente aplicada em tarefas reais, como classificação ou segmentação.

2.2.1 Funcionamento Geral do SSL

Em linhas gerais, métodos SSL seguem uma abordagem estruturada por etapas fundamentais:

1. **Definição da tarefa pretextual**: Uma tarefa artificial é elaborada com base exclusivamente nas características intrínsecas dos dados não rotulados.
2. **Treinamento do modelo**: O modelo é treinado para resolver a tarefa pretextual, minimizando uma função de perda específica associada à tarefa proposta.
3. **Transferência das representações aprendidas**: As representações internas obtidas são transferidas para serem utilizadas ou refinadas em tarefas reais com poucos ou nenhum dado rotulado adicional.

2.2.2 Aprendizado Contrastivo

O aprendizado contrastivo, representado por métodos como SimCLR (Chen et al., 2020), busca aprender representações discriminativas aproximando pares de amostras semelhantes (positivas) e afastando pares de amostras diferentes (negativas) no espaço de representação. Como mostrado na Figura 2.2.

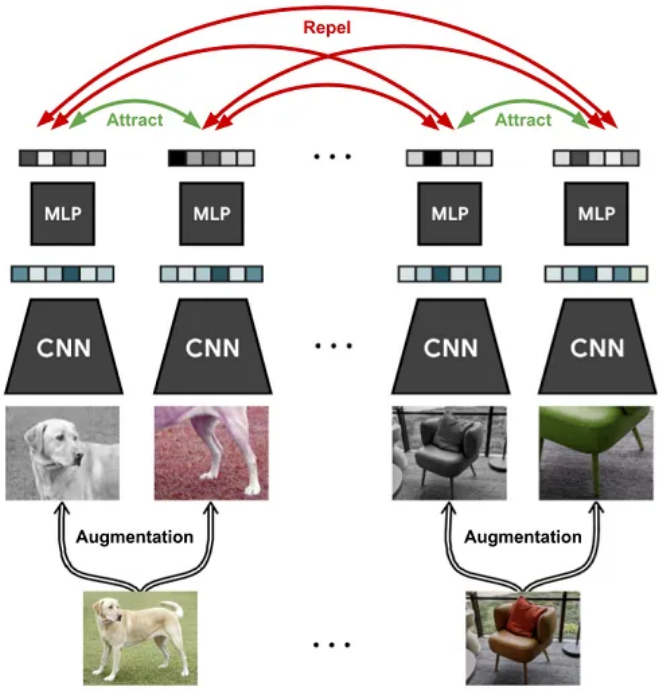


Figura 2.2: Ilustração do funcionamento do SimCLR. A mesma imagem passa por duas diferentes transformações (*augmentation*) gerando pares positivos, enquanto imagens distintas formam pares negativos. A rede é treinada para aproximar os positivos e afastar os negativos.Chen et al. (2020)

A função de perda contrastiva (InfoNCE) é formalizada como:

$$\mathcal{L}\text{contrastiva} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum k = 1^N \exp(\text{sim}(z_i, z_k)/\tau)} \tag{2.1}$$

onde:

- z_i e z_j são representações de um par positivo,
- $\text{sim}(\cdot, \cdot)$ é uma função de similaridade, tipicamente o produto escalar,
- τ é a temperatura, um hiperparâmetro,
- N é o número total de amostras no lote.

Esse processo garante que o modelo aprenda representações robustas e discriminativas.

2.2.3 DINO: Auto-Destilação sem Rótulos

O DINO (*Self-Distillation with No Labels*), proposto por Caron et al. (2021a), é um método de aprendizado auto-supervisionado que se destaca por combinar arquite-

turas modernas como o *Vision Transformer* (ViT) com uma estratégia de autodestilação eficaz, mesmo na ausência total de rótulos. Seu diferencial reside na capacidade de produzir representações semânticas robustas e transferíveis, explorando propriedades emergentes que não são observadas com frequência em arquiteturas convolucionais ou mesmo em ViTs treinados de forma supervisionada.

O método é estruturado sobre um arcabouço de professor e aluno. Ambos os modelos compartilham a mesma arquitetura, mas o professor é atualizado através de uma média móvel exponencial dos pesos do aluno, o que garante estabilidade e consistência durante o treinamento. Enquanto o professor é responsável por gerar distribuições de probabilidade a partir de transformações globais da imagem, o aluno é treinado para prever essas distribuições, processando tanto *crops* globais quanto locais. A diferença nas escalas das imagens utilizadas por cada rede obriga o modelo a aprender representações invariantes e consistentes em múltiplos níveis de granularidade.

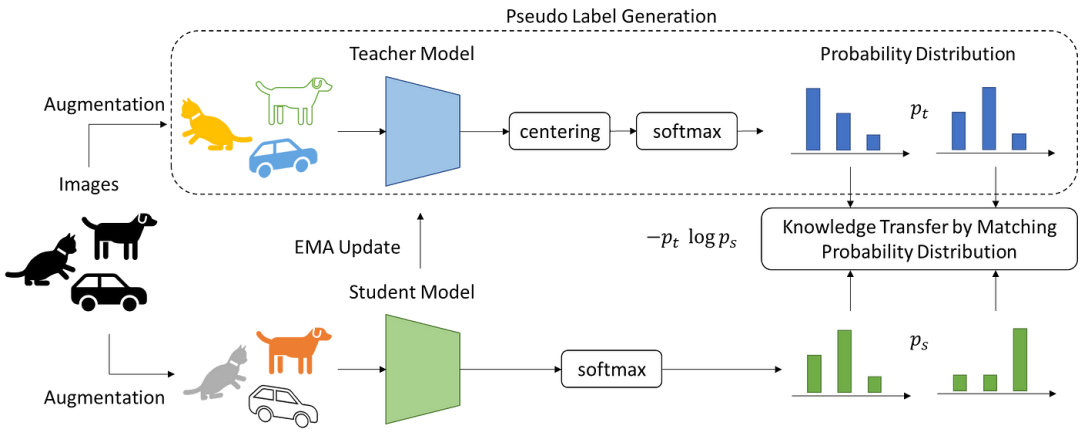


Figura 2.3: Ilustração do mecanismo de auto-destilação no DINO. O professor, atualizado por média móvel exponencial (EMA), recebe *crops* da imagem e produz uma distribuição de saída suavizada, que passa por centralização. O aluno, ao receber *crops* diferentes, aprende a imitar essa distribuição via entropia cruzada, promovendo a transferência de conhecimento sem rótulos Caron et al. (2021a).

Uma das inovações centrais do DINO é o uso do *multi-crop training*, uma técnica que consiste em aplicar múltiplos recortes (*crops*) da imagem de entrada em diferentes escalas. Tipicamente, dois *crops* maiores (abrangendo mais de 50% da imagem original) são utilizados como entrada para o professor, enquanto o aluno recebe esses mesmos *crops*, além de vários recortes menores. Essa estratégia força o modelo a alinhar suas representações entre diferentes vistas de uma mesma imagem, promovendo uma aprendizagem robusta e contextualizada.

A função de perda utilizada pelo DINO é formulada a partir da divergência entre as distribuições de probabilidade produzidas pelas duas redes — a do aluno e a do professor, conforme ilustrado na Equação 2.3. Cada uma dessas distribuições é obtida

aplicando-se a função *softmax* sobre os vetores de saída (*logits*) correspondentes, modulada por um parâmetro denominado **temperatura** (T). Esse parâmetro controla a "nitidez" ou a "suavidade" da distribuição de probabilidade, sendo incorporado na forma:

$$p_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}, \tag{2.2}$$

onde z_i representa o logit associado à classe i. Quando o valor de T é baixo, a função *softmax* acentua as diferenças entre os logits, produzindo uma distribuição mais afiada (ou concentrada), isto é, com probabilidades mais próximas de 0 ou 1. Por outro lado, valores de T altos atenuam essas diferenças, resultando em uma distribuição mais suave, em que as probabilidades das classes tendem a se aproximar umas das outras.

No contexto do DINO, o *professor* utiliza uma temperatura **baixa** (T_t), gerando distribuições mais determinísticas e concentradas, enquanto o *aluno* emprega uma temperatura **alta** (T_s), promovendo uma distribuição mais difusa. Essa diferença tem o propósito de estabilizar o aprendizado: o aluno tenta aproximar sua distribuição mais incerta daquela fornecida pelo professor, que atua como uma referência mais confiante. A função de perda é, portanto, a entropia cruzada entre essas duas distribuições:

$$\mathcal{L}_{DINO} = -\sum_i p_i^{(t)} \log p_i^{(s)}, \tag{2.3}$$

em que $p_i^{(t)}$ e $p_i^{(s)}$ correspondem, respectivamente, às probabilidades produzidas pelo professor e pelo aluno após a normalização com suas temperaturas específicas.

Para evitar o colapso representacional, situação em que todas as amostras são mapeadas para o mesmo vetor, o DINO adota mecanismos adicionais, como a centralização das saídas do professor e o uso de diferentes visões da mesma imagem, garantindo que a rede preserve a diversidade das representações aprendidas.

O DINO compartilha a filosofia de distilação sem rótulo com o método BYOL (*Bootstrap Your Own Latent*), introduzido por Grill et al. (2020b). Ambos evitam o uso de amostras negativas e trabalham com um mecanismo de professor-aluno. No BYOL, no entanto, o objetivo do aluno é aproximar diretamente os vetores de *embeddings* gerados pelo professor, utilizando uma projeção vetorial e a função de perda L2. Já no DINO, o foco está no alinhamento de distribuições de probabilidade inteiras, por meio de *softmax* e entropia cruzada, o que permite explorar estruturas mais ricas e globais no espaço latente.

Além disso, o DINO introduz o uso explícito de distribuições com diferentes temperaturas, bem como estratégias como centering e *multi-crop training*, que não são

parte do BYOL. Como consequência, o DINO tende a ser mais estável e menos propenso ao colapso representacional — um problema recorrente nos primeiros experimentos com o BYOL sem um regulador adicional como a normalização de batch. Por outro lado, o BYOL é mais simples em sua formulação e frequentemente utilizado com ResNets, enquanto o DINO demonstra propriedades emergentes mais marcantes quando acoplado a Vision Transformers.

2.2.4 Escolha de *Backbones* para SSL

No contexto do aprendizado profundo, o termo *backbone* refere-se à arquitetura central de extração de características do modelo. Em métodos de aprendizado auto-supervisionado (SSL), o *backbone* é responsável por transformar as imagens de entrada em representações latentes — ou *embeddings* — que capturam propriedades relevantes dos dados sem depender de rótulos. A eficácia das representações aprendidas, e sua transferibilidade para tarefas supervisionadas, está diretamente relacionada à capacidade do *backbone* de modelar padrões visuais de maneira robusta e contextualizada.

Neste estudo, foram utilizados três tipos principais de *backbones*: ResNet, *Vision Transformer* (ViT) e *Swin Transformer*. A seguir, detalham-se suas particularidades e justificativas técnicas, com ênfase na interação com métodos SSL como SimCLR e DINO.

SSL com ResNet: Invariâncias Locais e Robustez Textural A arquitetura ResNet (He et al., 2016), baseada em convoluções e conexões residuais, é amplamente utilizada como *backbone* em métodos contrastivos como o SimCLR. Sua capacidade de preservar e propagar gradientes por redes profundas permite a aprendizagem eficiente mesmo com conjuntos de dados limitados, como é comum em domínios biomédicos. As ResNets são particularmente eficazes para capturar padrões locais — como texturas e contornos — que são essenciais para diferenciar lesões histológicas Chen et al. (2020). Além disso, sua estrutura convolucional favorece a aquisição de invariâncias locais, úteis para a generalização em imagens com grande variação morfológica.

SSL com *Vision Transformers*: Atenção Global e Propriedades Emergentes Os *Vision Transformers* (ViT) (Dosovitskiy et al., 2020) introduzem um paradigma alternativo à convolução, tratando a imagem como uma sequência de patches e utilizando atenção global para modelar dependências espaciais de longo alcance. Essa abordagem é especialmente adequada ao método DINO, cujo mecanismo de autodestilação entre *crops* da imagem se beneficia da percepção global proporcionada pelos ViTs. No contexto da patologia computacional, os ViTs são valiosos por sua capacidade de capturar padrões semânticos amplos, permitindo reconhecer alterações morfológicas que se distribuem de maneira difusa pelo tecido.

SSL com *Swin Transformer*: Hierarquia e Multiescala O *Swin Transformer* (Liu et al., 2021) representa uma evolução dos ViTs ao incorporar atenção hierárquica por meio de janelas deslizantes deslocadas. Isso permite capturar tanto contextos locais quanto globais com eficiência computacional. Em métodos SSL como SimCLR e DINO, o *Swin Transformer* oferece vantagens por operar de forma multiescalar, promovendo a aprendizagem de representações robustas em diferentes níveis de granularidade. Essa característica é especialmente relevante para imagens médicas, onde lesões de diferentes tamanhos e complexidades coexistem.

Impacto na Dinâmica de Treinamento e Transferência A escolha do *backbone* também influencia diretamente a dinâmica de otimização dos modelos SSL. *backbones* convolucionais como a ResNet tendem a apresentar treinamento mais estável e rápido, com menor sensibilidade a hiperparâmetros He et al. (2016). Em contrapartida, arquiteturas baseadas em atenção, como ViT e Swin, exigem ajustes finos, como *learning rate warmup*, controle de temperatura no *softmax* e técnicas como *output centering* Dosovitskiy et al. (2020); Caron et al. (2021a). Apesar disso, essas arquiteturas frequentemente produzem representações mais semânticas e transferíveis Azizi et al. (2021), com melhor desempenho em tarefas complexas de classificação, especialmente em contextos biomédicos.

2.2.5 Síntese dos Passos Fundamentais do SSL

Resumidamente, o método SSL pode ser descrito pelas seguintes etapas:

- **Criação da tarefa pretextual**;
- **Treinamento para resolver essa tarefa**;
- **Transferência ou adaptação das representações obtidas** para tarefas específicas.

De forma geral, o aprendizado auto-supervisionado (SSL) constitui uma abordagem alternativa promissora aos métodos supervisionados tradicionais, pois dispensa a necessidade de rótulos manuais durante a etapa de pré-treinamento. Essa característica o torna particularmente atrativo em domínios como a Patologia Computacional, em que a anotação de grandes conjuntos de dados demanda tempo, recursos e conhecimento especializado. Assim, o SSL surge como uma estratégia eficaz para explorar informações intrínsecas das imagens médicas e reduzir a dependência de dados rotulados, conforme discutido a seguir.

2.3 SSL em Patologia Computacional

A aplicação de aprendizado auto-supervisionado (SSL) na patologia computacional tem crescido nos últimos anos, motivada pelas dificuldades associadas à anotação

de grandes volumes de dados médicos, que exige conhecimento especializado e enfrenta limitações relacionadas à disponibilidade de patologistas experientes (Tajbakhsh et al., 2020). Esse cenário é evidente em histopatologia, onde a análise de lâminas digitalizadas envolve esforço considerável e conhecimento técnico.

O SSL parte do princípio de que é possível aprender representações informativas a partir da estrutura dos próprios dados, sem o uso de rótulos manuais. Diferentemente de domínios como a visão computacional geral, onde conjuntos como o ImageNet (Deng et al., 2009) fornecem milhões de imagens anotadas, o contexto médico é marcado pela escassez de dados rotulados (He et al., 2022). Isso torna o SSL uma alternativa relevante, considerando os custos associados à anotação em aplicações clínicas.

Diversas abordagens de SSL vêm sendo utilizadas em tarefas envolvendo imagens histológicas Jin et al. (2022b). Métodos como MoCo v2, SwAV, Barlow Twins, SimCLR, BYOL e DINO foram empregados com o objetivo de reduzir a dependência de anotações manuais e lidar com a variabilidade visual presente nesses dados. A Tabela 2.1 apresenta uma síntese de estudos recentes que exploraram essas estratégias na patologia computacional, com diferentes combinações de bases de dados, protocolos e arquiteturas.

Apesar dos avanços, a literatura ainda apresenta lacunas na avaliação sistemática de métodos SSL em cenários de patologia renal caracterizados por múltiplas colorações e forte desbalanceamento de classes, fatores que este trabalho aborda.

Tabela 2.1: Resumo de estudos recentes que utilizaram SSL na Patologia Computacional

Referência	**Ano**	**Dataset**	**Método SSL**
Stacke et al. (2021b)	2021	Histológico	SimCLR
Azizi et al. (2021)	2021	Multimodal	SimCLR (MICLe)
Ciga et al. (2022)	2022	Histológico	SimCLR
Jin et al. (2022a)	2022	ImageNet/Histológico	SimCLR
Yang et al. (2022)	2022	H&E Breast	CS-CO (Híbrido)
Gadermayr et al. (2022)	2022	Renal Pathology	Stain-Adaptive SSL
Bauer e Augenstein (2023)	2023	Prostate Cancer	DINO
Wang et al. (2023)	2023	TCGA (WSI)	Pyramid SSL
Filiot et al. (2023)	2023	Camelyon16	Masked Image Modeling
Oquab et al. (2023a)	2023	NIH Pan-Câncer	DINOv2

Este trabalho adota o método DINO como foco de análise, com base em sua frequência de uso na literatura, desempenho relatado em estudos prévios e aderência às características dos dados utilizados. A decisão de concentrar-se nesse método também considera sua capacidade de aprender representações a partir de diferentes transformações de imagem, sem depender de rótulos explícitos, o que é compatível com os desafios apresentados por imagens histológicas de tecido renal. A seguir, são discutidos em maior detalhe os principais trabalhos relacionados que aplicaram técnicas de

SSL em contextos diversos da patologia computacional, organizados por categorias metodológicas e desafios específicos abordados.

2.3.1 Aprendizado Contrastivo Adaptado

O framework HistoSSL (Jin et al., 2022a) representa uma das primeiras adaptações bem-sucedidas de métodos contrastivos para histopatologia, demonstrando que o pré-treinamento auto-supervisionado pode melhorar significativamente o desempenho na classificação de tecidos histopatológicos mesmo com quantidades limitadas de exemplos anotados. O método adapta técnicas como SimCLR (Chen et al., 2020) e BYOL (Grill et al., 2020b) para capturar as características específicas de imagens histológicas. Estudos posteriores de Stacke et al. (2021b) e Ciga et al. (2022) aprofundaram a análise sobre como ajustes específicos nos aumentos de dados — incluindo rotações, inversões, modificações de cor e transformações estocásticas — são cruciais para que o SSL capture adequadamente as estruturas morfológicas relevantes em imagens histológicas. Estas investigações revelaram que transformações tradicionais de visão computacional podem não ser ideais para tecidos biológicos, onde certas orientações e padrões de coloração carregam significado diagnóstico específico.

2.3.2 Estratégias Invariante à Coloração

Um desafio particular em histopatologia computacional é a variabilidade na coloração entre diferentes laboratórios e protocolos de preparação. Gadermayr et al. (2022) propuseram métodos de SSL adaptativos à coloração, enquanto Tiard et al. (2022) desenvolveram abordagens invariantes à coloração que mantêm a capacidade de generalização entre diferentes protocolos de preparação histológica. Estas contribuições são fundamentais para aplicações clínicas reais, onde a robustez à variação de coloração é essencial. Yang et al. (2022) introduziram o *Color-Sensitive Contrastive Learning with Co-Training* (CS-CO), um método híbrido que combina aprendizado contrastivo e co-treinamento para imagens HE, demonstrando melhorias consistentes em múltiplas tarefas de classificação histopatológica. O método aborda especificamente os desafios relacionados à heterogeneidade visual típica de amostras histológicas.

2.3.3 *Masked Image Modeling* em Histopatologia

Seguindo tendências em processamento de linguagem natural, Filiot et al. (2023) exploraram o uso de *Masked Image Modeling* (MIM) para histopatologia, demonstrando que técnicas de mascaramento podem capturar padrões estruturais complexos em imagens de alta resolução. Esta abordagem mostra-se particularmente promissora para *whole-slide images* (WSIs), onde a análise multi-escala é fundamental.

2.3.4 Análise de Tecido Renal

Hermsen et al. (2019) pioneiramente aplicaram *deep learning* para avaliação histopatológica de tecido renal, estabelecendo bases metodológicas que posteriormente foram expandidas com técnicas SSL. O trabalho demonstrou que métodos automáticos podem atingir concordância substancial com patologistas experientes na identificação de lesões glomerulares, seguindo diretrizes como as estabelecidas por Haas et al. (2020).

2.3.5 Detecção de Mitoses e Análise de Proliferação

Tellez et al. (2019) desenvolveram métodos para detecção de mitoses em histologia mamária usando *Phosphohistone H3* (PHH3) como referência, criando redes invariantes à coloração. Esta linha de pesquisa estabeleceu precedentes importantes para o uso de SSL em tarefas de detecção específicas em patologia.

2.3.6 Classificação Multi-classe e *Few-shot Learning*

Wang et al. (2023) propuseram uma abordagem baseada em pirâmides para SSL em classificação histopatológica, demonstrando eficácia particular em cenários de *few-shot learning* — situação comum em patologia devido à raridade de certas condições. O método mostrou-se eficaz na captura de características em múltiplas escalas, aspecto crucial para análise histológica.

2.3.7 *Big Self-Supervised Models*

Azizi et al. (2021) apresentaram uma estratégia baseada em grupos patológicos (MICLe - *Medical Image Contrastive Learning*) que melhora o contraste sem depender de rótulos manuais, com resultados competitivos não apenas em histopatologia, mas também em radiografias e imagens dermatológicas. O estudo demonstrou que modelos SSL de grande escala podem transferir conhecimento efetivamente entre diferentes modalidades de imagem médica.

2.3.8 *Benchmarking* e Avaliação Sistemática

Kang et al. (2023) conduziram uma avaliação abrangente de métodos SSL em múltiplos datasets patológicos, revelando que a performance varia significativamente dependendo do tipo de tecido e tarefa específica. Este trabalho estabeleceu *benchmarks* importantes e identificou lacunas metodológicas que precisam ser endereçadas.

2.3.9 Representação Distribucional e *Augmentation*

Tang et al. (2024) introduziram métodos de aprendizado de distribuição de representações para *Augmentation* confiável de dados em classificação de WSIs, abordando o

problema de como gerar aumentos de dados realisticamente úteis para treinamento em contextos médicos.

2.3.10 Desafios e Limitações Identificadas

O fenômeno conhecido como *domain shift* refere-se à diferença de distribuição entre os dados utilizados no treinamento e aqueles encontrados durante a inferência. Em outras palavras, um modelo treinado em um determinado domínio — por exemplo, imagens provenientes de um laboratório específico — pode apresentar degradação de desempenho quando exposto a dados oriundos de outras fontes com características distintas. Essas diferenças podem incluir variações de coloração, resolução, iluminação ou mesmo discrepâncias nos protocolos de preparo e digitalização das lâminas histológicas.

Stacke et al. (2021a) conduziram uma análise detalhada sobre *domain shift* em histopatologia, identificando que variações entre instituições, protocolos de preparação e equipamentos de digitalização podem comprometer significativamente a generalização de modelos SSL. Desafios, como o *domain shift* causado pela variabilidade de coloração e a dificuldade de lidar com classes minoritárias, são precisamente o foco de investigação deste trabalho. Ao empregar técnicas de SSL e estratégias de balanceamento de dados, buscou-se desenvolver modelos mais robustos e generalizáveis, para lidar com a heterogeneidade inerente aos dados histopatológicos no contexto da patologia renal.

Revisões sistemáticas como a de Huang et al. (2023) sintetizam contribuições da área, classificando métodos SSL em categorias como aprendizado contrastivo, predição pretextual e destilação de conhecimento. O estudo destaca criticamente a carência de *benchmarks* padronizados e a predominância inadequada de métricas como acurácia simples, que podem ser inadequadas em contextos com forte desbalanceamento de classes.

Maier-Hein et al. (2024b) enfatizaram a necessidade de métricas mais adequadas para validação em análise de imagens médicas, propondo diretrizes específicas que consideram as particularidades do domínio clínico.

Strubell et al. (2019) levantaram questões importantes sobre custos computacionais de métodos SSL, particularmente relevantes em contextos médicos onde recursos computacionais podem ser limitados. Esta consideração é crucial para tradução clínica de métodos SSL.

A integração de SSL com métodos de *few-shot learning* (Snell et al., 2017) representa uma direção promissora, especialmente para diagnósticos de condições raras. Adicionalmente, a combinação de SSL com arquiteturas modernas como *Vision Transformers* (Dosovitskiy et al., 2020) e suas variantes hierárquicas (Liu et al., 2021) oferece potencial para capturar relações espaciais complexas em imagens histológicas.

A aplicação de métodos como DINOv2 (Oquab et al., 2023a) especificamente adaptados para histopatologia representa outra fronteira de pesquisa, com potencial para melhorar a qualidade das representações aprendidas sem supervisão.

Finalmente, a integração de SSL com ferramentas de apoio diagnóstico como o PathoSpotter (Oliveira et al., 2022; FIOCRUZ, 2024) demonstra o potencial de tradução clínica dessas tecnologias, especialmente em cenários com limitação de expertise patológica especializada.

2.4 *Few-Shot Learning* (FSL) e sua Aplicação em Domínios com Poucos Rótulos

Few-shot learning refere-se à capacidade de um modelo de generalizar para novas classes ou tarefas com apenas algumas amostras rotuladas por classe (por exemplo, $k = 1, 5, 10$ exemplos). Frequentemente, utiliza-se o esquema n-way k-shot, onde n representa o número de classes novas e k o número de exemplos de suporte por classe (Pachetti e Colantonio, 2023). Entre os métodos mais comuns destacam-se abordagens baseadas em métrica, como redes siamesas ou redes protótipo, que comparam *embeddings* de suporte e consulta; técnicas de meta-aprendizado ou otimização, como *Model-Agnostic Meta-Learning* (MAML), que ajustam rapidamente os parâmetros do modelo para novas tarefas; e estratégias de transferência, em que *embeddings* pré-treinados são adaptados com poucas amostras rotuladas (VuNguyen et al., 2024; ZhengruiGuo et al., 2024).

Em domínios com escassez de anotação, como a medicina e, em particular, a patologia digital, o FSL tem se mostrado relevante por reduzir a dependência de grandes bases rotuladas. Revisões recentes indicam que o meta-aprendizado e as técnicas baseadas em *embeddings* são escolhas frequentes em imagens médicas, devido à sua capacidade de adaptação rápida a novas classes (Pachetti e Colantonio, 2023).

2.4.1 *Few-Shot Learning* Integrado a SSL em Patologia Computacional

A combinação de aprendizado auto-supervisionado com *few-shot learning* tem se mostrado promissora na classificação de imagens histopatológicas, especialmente quando há poucas amostras rotuladas disponíveis. Nessa abordagem, inicialmente realiza-se pré-treinamento via SSL para extrair representações gerais e robustas a partir de dados não rotulados. Em seguida, essas representações são adaptadas para novas classes ou tarefas específicas utilizando apenas algumas amostras rotuladas (*k-shot*). Estudos recentes indicam que essa combinação permite superar limitações de variabilidade intra-classe e heterogeneidade de lâminas digitais, fatores que dificultam a generalização de modelos supervisionados convencionais (HaoQuan et al., 2023).

Na prática, modelos pré-treinados via SSL podem ser facilmente adaptados a tarefas de classificação *few-shot* em imagens patológicas, utilizando protótipos de classe ou classificadores leves, reduzindo o custo de anotação e de ajuste de hiperparâmetros. Trabalhos como o modelo *Universal Network for Histopathology Image Representation* (UNI) demonstram que *embeddings* extraídos de milhões de *patches* podem ser utilizados para classificação rápida de novas classes através de *few-shot learning*, mantendo desempenho competitivo em múltiplas tarefas clínicas (Chen et al., 2024). Outra abordagem recente, o *Dual-channel Prototype Network*, combina SSL e *few-shot* para classificação de imagens patológicas com amostras limitadas, evidenciando a aplicabilidade prática desse método (HaoQuan et al., 2023).

Para estudos como o da presente dissertação, que compara métodos SSL como SimCLR, BYOL e DINO com baselines supervisionados, a integração de um protocolo *few-shot* aliado ao pré-treinamento auto-supervisionado pode potencializar o desempenho das tarefas de classificação, especialmente em classes minoritárias, como depósitos de amiloidose ou necrose fibrinoide, mesmo com poucas amostras rotuladas. Essa abordagem reforça a relevância da investigação em cenários de anotação restrita, típicos da patologia computacional, ao evidenciar como representações robustas obtidas via SSL podem ser exploradas para melhorar a acurácia e a confiabilidade das previsões com custo reduzido de anotação.

2.5 Avaliação de desempenho

2.5.1 Matriz de Confusão

A matriz de confusão é uma ferramenta fundamental para a avaliação do desempenho de modelos de classificação, especialmente em cenários onde a distribuição das classes é desbalanceada. Ela oferece uma visão detalhada dos acertos e erros do modelo, discriminando os tipos de predições realizadas em relação aos valores reais.

Uma matriz de confusão para um problema de classificação binária é tipicamente estruturada da seguinte forma:

Predito \ Real	**Classe Positiva**	**Classe Negativa**
Classe Positiva	Verdadeiros Positivos (VP)	Falsos Positivos (FP)
Classe Negativa	Falsos Negativos (FN)	Verdadeiros Negativos (VN)

Os elementos da matriz são definidos como:

- **Verdadeiros Positivos (VP)**: Casos em que o modelo previu corretamente a classe positiva. São as instâncias positivas que foram classificadas corretamente como positivas.
- **Falsos Positivos (FP)**: Casos em que o modelo previu a classe positiva, mas o valor real era negativo. Também conhecidos como erro Tipo I, representam instâncias negativas que foram incorretamente classificadas como positivas.

- **Falsos Negativos (FN)**: Casos em que o modelo previu a classe negativa, mas o valor real era positivo. Também conhecidos como erro Tipo II, representam instâncias positivas que foram incorretamente classificadas como negativas.
- **Verdadeiros Negativos (VN)**: Casos em que o modelo previu corretamente a classe negativa. São as instâncias negativas que foram classificadas corretamente como negativas.

A partir da matriz de confusão, é possível derivar diversas métricas de desempenho, como acurácia, precisão, revocação e *F1-score*, que fornecem diferentes perspectivas sobre a qualidade do classificador. A análise desses componentes permite identificar se o modelo está tendendo a falsos positivos ou falsos negativos, o que é crucial para ajustar seu comportamento de acordo com os custos associados a cada tipo de erro em um determinado domínio, como o médico.

2.5.2 *F1 Score*

A escolha de métricas adequadas para avaliação de desempenho é essencial em tarefas de classificação, especialmente quando há desbalanceamento entre classes — uma característica recorrente em bases de dados médicos. Nesses contextos, a acurácia pode se tornar uma métrica enganosa, pois modelos tendem a obter altos valores simplesmente por acertarem as classes majoritárias, negligenciando o desempenho nas classes minoritárias.

Para reduzir esse problema, é comumente adotado o F1 score, uma métrica que combina precisão e revocação de maneira harmônica, proporcionando uma visão mais equilibrada do desempenho do modelo.

O F1 score é definido como:

$$F1 = 2 \times \frac{\text{precisão} \times \text{revocação}}{\text{precisão} + \text{revocação}} \tag{2.4}$$

em que:

$$\text{Precision} = \frac{VP}{VP + FP} \tag{2.5}$$

$$\text{Recall} = \frac{VP}{VP + FN} \tag{2.6}$$

O valor de F1 varia de 0 a 1, penalizando desequilíbrios entre precisão e revocação. Essa métrica é especialmente relevante quando se deseja garantir que o modelo não apenas detecte a maioria dos casos positivos (alta revocação), mas também que os casos positivos detectados sejam corretos (alta precisão).

O F1 score é aplicável em problemas de classificação binária, multiclasse e multilabel, com variações no modo de agregação dos resultados conforme descrito a seguir.

2.5.3 F1-micro e F1-macro: Estratégias de Agregação

Em cenários com múltiplas classes ou rótulos simultâneos, o F1 score pode ser agregado de diferentes formas. As duas abordagens mais comuns são o F1-micro e o F1-macro.

O F1-micro calcula os valores de verdadeiros positivos (VP), falsos positivos (FP) e falsos negativos (FN) somando os resultados de todas as classes ou rótulos. A partir desses totais, são computadas a precisão e a revocação globais, e então o F1 score correspondente. Essa abordagem favorece classes mais frequentes e reflete o desempenho geral do modelo, como apresentado na Equação 2.7.

$$F1_{\text{micro}} = \frac{2 \cdot \text{VPtotal}}{2 \cdot \text{VPtotal} + \text{FPtotal} + \text{FNtotal}} \tag{2.7}$$

O F1-micro é adequado quando se deseja priorizar a performance geral, considerando o impacto proporcional das classes com base em sua frequência.

Já o F1-macro calcula o F1 score individualmente para cada classe e, em seguida, realiza a média aritmética dos resultados. Cada classe tem o mesmo peso, independentemente de sua representatividade no conjunto de dados, como apresentado na Equação 2.8.

$$F1_{\text{macro}} = \frac{1}{C} \sum_{i=1}^{C} F1_i \tag{2.8}$$

onde C é o número de classes.

O F1-macro é mais apropriado em contextos onde todas as classes são igualmente importantes, mesmo aquelas com menor frequência, como costuma ocorrer em tarefas de detecção de anomalias ou identificação de lesões raras.

Em classificação binária, utiliza-se o F1 score padrão, ou seja, para a classe positiva principal, sendo útil para avaliar o equilíbrio entre precisão e revocação quando um dos rótulos é mais relevante (por exemplo, “lesão presente” versus “ausente”). Em classificação multiclasse, emprega-se F1-micro ou F1-macro para obter uma medida agregada de desempenho entre várias categorias mutuamente exclusivas. Em classificação multilabel, onde múltiplas classes podem coexistir por amostra (como em imagens médicas com múltiplas lesões), o F1-micro e F1-macro podem ser adaptados para considerar cada rótulo como uma tarefa binária individual. Essa avaliação permite capturar desequilíbrios tanto entre classes quanto entre amostras.

2.5.4 Validação Cruzada

A validação cruzada (cross-validation) é uma técnica estatística fundamental para avaliar o desempenho de modelos de aprendizado de máquina de forma robusta

e imparcial. Seu principal objetivo é estimar como um modelo se comportará em dados não vistos, mitigando problemas de sobreajuste (overfitting) e fornecendo uma avaliação mais confiável da capacidade de generalização.

O princípio básico da validação cruzada consiste em dividir o conjunto de dados em múltiplas partições, utilizando algumas para treinamento e outras para validação, repetindo esse processo de forma sistemática. Essa abordagem permite que todos os dados sejam utilizados tanto para treino quanto para teste em diferentes iterações, maximizando o aproveitamento das informações disponíveis.

A variante mais comum é a validação cruzada k-fold, onde o conjunto de dados é dividido em k subconjuntos (folds) de tamanhos aproximadamente iguais. O processo é executado em k iterações, onde em cada uma delas $k-1$ folds são usados para treinamento e o fold restante para validação (Browne (2000)). O desempenho final é calculado como a média das métricas obtidas em todas as iterações, conforme a Equação 2.9:

$$\text{Desempenho}_{\text{CV}} = \frac{1}{k}\sum_{i=1}^{k}\text{Métrica}_i \tag{2.9}$$

onde Métrica_i representa o valor da métrica (por exemplo, F1 score) obtido na i-ésima iteração.

Valores típicos para k incluem 5 ou 10, representando um compromisso entre viés e variância na estimativa do desempenho. A validação cruzada tende a produzir estimativas com menor viés em relação ao desempenho real do modelo, porém é computacionalmente mais custosa e pode apresentar alta variância, especialmente quando aplicada a conjuntos de dados pequenos.

Para dados estratificados, como é comum em aplicações médicas com classes desbalanceadas, utiliza-se a validação cruzada estratificada, que preserva a proporção de cada classe em todos os folds. Dessa forma, garante-se que cada partição mantenha aproximadamente a mesma distribuição das classes observada no conjunto original. Matematicamente, essa propriedade pode ser expressa pela Equação 2.10, que assegura que a proporção de amostras da classe c em um determinado fold i seja próxima à proporção global dessa classe no conjunto total:

$$\frac{n_{c,i}}{n_i} \approx \frac{n_c}{n} \quad \forall c, i \tag{2.10}$$

onde $n_{c,i}$ representa o número de amostras da classe c no fold i, n_i é o tamanho do fold i, n_c é o número total de amostras da classe c e n é o número total de amostras no conjunto de dados.

A validação cruzada é particularmente valiosa em contextos médicos onde os dados são limitados e cada amostra é preciosa. Além de fornecer uma estimativa mais

confiável do desempenho, permite calcular medidas de variabilidade (por exemplo, desvio padrão) das métricas, oferecendo insights sobre a estabilidade do modelo.

2.5.5 Significância Estatística

A análise de significância estatística é um componente essencial na avaliação de modelos de aprendizado de máquina, pois permite verificar se as diferenças observadas nas métricas de desempenho (como F1-score, acurácia ou AUC) refletem variações reais entre os modelos ou se podem ser atribuídas ao acaso (Maier-Hein et al. (2024a); Dietterich (1998)). Em experimentos envolvendo aprendizado supervisionado ou auto-supervisionado, pequenas flutuações nos conjuntos de treino e validação podem gerar variações consideráveis no resultado, especialmente em cenários com amostras limitadas ou dados desbalanceados. Por isso, a inferência estatística fornece uma base formal para distinguir diferenças genuínas de ruído experimental.

A comparação entre modelos é normalmente estruturada como um teste de hipóteses:

- **Hipótese nula (H_0):** não existe diferença estatisticamente significativa entre os desempenhos médios dos modelos comparados;
- **Hipótese alternativa (H_1):** existe diferença estatisticamente significativa entre os modelos.

O resultado é interpretado com base no valor-p, que representa a probabilidade de observar uma diferença tão grande quanto a encontrada (ou maior), assumindo que a hipótese nula (H_0) seja verdadeira. Para decidir se essa diferença é estatisticamente significativa, comparamos o valor-p com um nível de significância, denotado por α (alfa). Esse nível é um limite pré-estabelecido que define quão improvável uma diferença precisa ser para considerarmos que não ocorreu por acaso. Comum na prática é usar $\alpha = 0,05$. Se o valor-p for menor que α, rejeitamos H_0, indicando que a diferença entre os modelos provavelmente não é fruto do acaso.

Em experimentos de validação cruzada, cada modelo é avaliado em múltiplos *folds*, e as métricas obtidas em cada partição podem ser tratadas como observações pareadas, isto é, cada fold fornece uma comparação direta entre dois modelos nas mesmas condições de dados. Nesses casos, utilizam-se testes específicos para dados dependentes:

- **Teste t pareado:** apropriado quando as diferenças entre os resultados dos modelos seguem uma distribuição aproximadamente normal e apresentam variâncias semelhantes. O valor do teste é calculado por:

$$t = \frac{\bar{d}}{s_d/\sqrt{k}} \tag{2.11}$$

  onde $\bar{d}$ é a média das diferenças entre os modelos em cada fold, s_d é o desvio padrão dessas diferenças e k é o número de folds (Dietterich (1998)). O teste avalia se a média das diferenças é estatisticamente diferente de zero.

- **Teste de postos com sinais de Wilcoxon:** alternativa não paramétrica que não exige a suposição de normalidade. Ele ordena as diferenças absolutas entre os pares de observações e analisa os sinais (positivos ou negativos) dessas diferenças. É especialmente útil em conjuntos pequenos ($k \leq 30$) ou quando há *outliers*, pois oferece maior robustez (Demšar (2006)).

Quando mais de dois modelos são comparados simultaneamente, por exemplo, em experimentos envolvendo várias arquiteturas de aprendizado auto-supervisionado, deve-se ajustar o nível de significância para evitar o aumento da probabilidade de falsos positivos (erro Tipo I). Para isso, aplicam-se procedimentos de correção múltipla:

- **Correção de Bonferroni:** divide o limiar α pelo número total de comparações m, resultando em $\alpha_{\text{ajustado}} = \alpha/m$;
- **Método de Holm-Bonferroni:** procedimento sequencial menos conservador, que ajusta os valores-p de forma progressiva, aumentando o poder do teste;
- **Controle da Taxa de Falsas Descobertas (FDR – Benjamini-Hochberg):** prioriza o controle da proporção esperada de falsos positivos, sendo mais apropriado quando o número de comparações é grande.

Em síntese, a significância estatística fornece a base para afirmar, com confiança, se o desempenho superior de um modelo sobre outro decorre de um ganho real de generalização e não de variações aleatórias inerentes ao processo de treinamento ou amostragem.

Capítulo 3

Metodologia

Este capítulo descreve, de forma detalhada, a metodologia empregada para investigar a aplicabilidade de técnicas de aprendizado auto-supervisionado (SSL) na classificação de lesões em imagens histopatológicas renais. São apresentados os dados utilizados, incluindo a origem, organização por corantes e classes patológicas, e os procedimentos de balanceamento e transformação das imagens. Em seguida, detalha-se a infraestrutura computacional adotada, os métodos SSL utilizados (SimCLR, BYOL e DINO), bem como os *backbones* avaliados (ResNet, ViT e Swin Transformer). A estrutura experimental é explicada com base em um pipeline dividido em pré-treinamento auto-supervisionado, *fine-tuning* supervisionado e avaliação com validação cruzada estratificada. São discutidas também as abordagens multiclasse e binária (one-vs-all), os critérios de avaliação com ênfase no F1-score e as estratégias de visualização com t-SNE para interpretação das representações aprendidas 4.2.1. Por fim, empregou-se análise de significância estatística (teste t) para comparar os resultados dos modelos e verificar se as diferenças observadas entre os métodos pré-treinados são estatisticamente relevantes. Essa fundamentação fornece suporte à análise comparativa entre os métodos e arquiteturas, apresentada no capítulo seguinte.

3.1 Descrição Geral do Método

A Figura 3.1 apresenta uma visão geral do método experimental desenvolvido neste estudo, estruturado em quatro etapas principais: preparação dos dados, pré-treinamento auto-supervisionado, ajuste supervisionado (*fine-tuning*) e avaliação dos modelos. Essa estrutura foi definida para permitir a comparação entre diferentes métodos de aprendizado auto-supervisionado, arquiteturas de rede e estratégias de classificação.

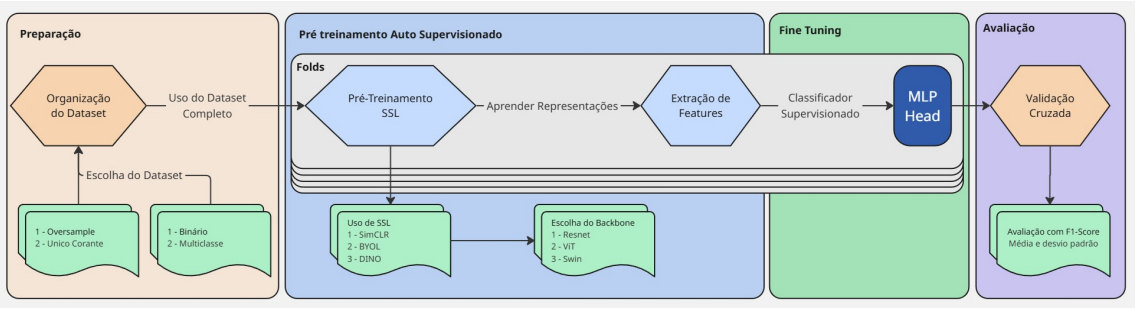


Figura 3.1: Fluxograma do método utilizado, da preparação dos dados à avaliação final. Os blocos em verde indicam escolhas experimentais (e.g., tipo de tarefa, técnica SSL e backbone). O processo é conduzido com validação cruzada estratificada, refletida pela sobreposição de camadas no fluxo principal.

1. Preparação dos Dados: O conjunto de dados foi inicialmente organizado de acordo com as classes diagnósticas e os corantes histológicos. O total de 12.521 imagens foi particionado por meio de validação cruzada estratificada em 5 *folds*, preservando a proporção original das classes em cada divisão, com aproximadamente 80% das imagens destinadas ao treinamento e 20% ao teste em cada *fold*. A distribuição global das imagens por classe, bem como suas respectivas proporções, é apresentada na Tabela 3.1.

Tabela 3.1: Distribuição global das imagens por classe no conjunto de dados.

Classe	**Nº de imagens**	**Proporção (%)**
Hipercelularidade	3134	25.0
Normal	2695	21.5
Membranous	1539	12.3
Esclerose pura sem crescente	1482	11.8
Crescente	1104	8.8
Acúmulo de neutrófilos	719	5.7
Esclerose	617	4.9
Podocitopatia	505	4.0
Amiloidose	374	3.0
Hipercelularidade pura sem crescente	224	1.8
Depósitos hialinos	94	0.8
Necrose fibrinoide	34	0.3
Total	12.521	100.0

Foram consideradas duas configurações experimentais: (i) aplicação de *oversampling* apenas no conjunto de treinamento para mitigar o desbalanceamento entre classes; e (ii) experimentos conduzidos utilizando imagens de um único corante por vez. A partir dessa organização, os dados foram preparados para cenários de classificação binária (*one-vs-all*) e multiclasse.

2. Pré-Treinamento Auto-Supervisionado: As imagens organizadas foram utilizadas em uma etapa de pré-treinamento não supervisionado, com o objetivo de aprender representações latentes sem uso de rótulos. Foram empregados três métodos: SimCLR, BYOL e DINO, combinados com diferentes arquiteturas de backbone (ResNet, ViT e Swin Transformer). Esta etapa foi realizada em múltiplos folds, respeitando a estrutura de validação cruzada estratificada.

3. Extração de Representações e Fine-Tuning: Após o pré-treinamento, os vetores de características foram extraídos pelas redes e utilizados como entrada para um classificador supervisionado do tipo *MLP Head*. Essa etapa foi realizada separadamente para os experimentos binários e multiclasse, mantendo-se a divisão dos dados definida pelos folds.

4. Avaliação: Os modelos foram avaliados em cada subdivisão da validação cruzada, utilizando o F1-score como métrica principal. Os valores de média e desvio padrão foram computados para cada experimento, permitindo a análise comparativa entre as diferentes configurações testadas.

Essa sequência de etapas viabilizou a aplicação controlada dos métodos de aprendizado auto-supervisionado no contexto da classificação de lesões renais, considerando múltiplos fatores como coloração, estrutura arquitetural e natureza do problema de classificação.

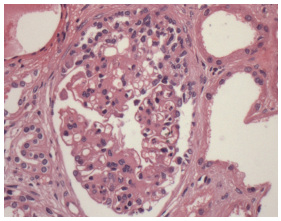
3.2 Dataset

Para este estudo, utilizamos um conjunto de imagens fornecido pelo Dr. Washington Luis Conrado dos Santos, médico patologista do Instituto Gonçalo Moniz da FIOCRUZ. O acervo é composto por 12.524 imagens histopatológicas de tecidos renais, obtidas com diferentes técnicas de coloração, incluindo Hematoxilina-eosina (*Hematoxylin and Eosin* – HE), Ácido Periódico de Schiff (*Periodic Acid–Schiff* – PAS), Tricrômico de Azan (*Azan Trichrome* – AZAN), bem como corantes adicionais, como Ácido Periódico de Prata Metenamina de Gomori (*Gomori Methenamine Silver* – PAMS) e Índice de Solvatação Preferencial (*Preferential Solvation Index* – PSI). Essas colorações têm como objetivo destacar estruturas e lesões específicas do tecido renal, proporcionando subsídios valiosos para a análise computacional.

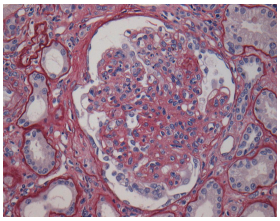
As imagens apresentam uma ampla variação dimensional, com larguras variando de 219 a 4128 pixels e alturas entre 205 e 3286 pixels, refletindo a diversidade de métodos de aquisição e digitalização empregados no acervo. Essas imagens contemplam diferentes tipos de lesões, conforme ilustrado nas Figuras 3.2, 3.3 e 3.4, tais como acúmulo de neutrófilos, depósitos hialinos, necrose fibrinoide, amiloidose, esclerose, hipercelularidade, crescente, membranosa e podocitopatia. Na Tabela 3.2, apresentamos a distribuição quantitativa das imagens por tipo de lesão, a fim de facilitar o controle e a análise do conjunto de dados.

Tabela 3.2: Distribuição de classes no *dataset* de imagens histopatológicas da FIOCRUZ.

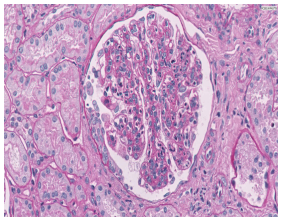
Classe	Total	AZAN	HE	PAMS	PAS	PSI
Total de Amostras	12524	1171	6369	1254	3668	52
Normal	2695	223	1585	345	542	0
Amiloidose	374	31	145	96	102	0
Esclerose Pura Sem Crescente	1481	233	672	104	472	0
Hipercelularidade	3134	257	1890	0	987	0
Hipercelularidade Pura Sem Crescente	224	60	0	0	164	0
Crescente	1104	121	467	157	359	0
Membranosa	1539	136	712	324	367	0
Esclerose	617	0	276	122	219	0
Podocitopatia	505	90	65	106	244	0
Acumulo de neutrofilos	713	0	486	0	175	52
Depositos hialinos	94	14	51	0	29	0
Necrose fibrinoide	34	6	20	0	8	0



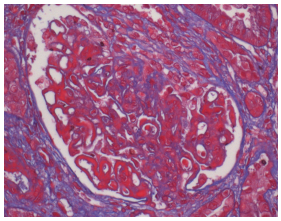
(a) Acumulo de Neutrófilos HE



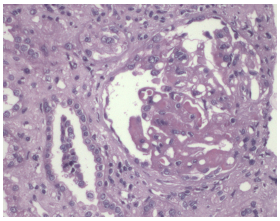
(b) Acumulo de Neutrófilos PAS



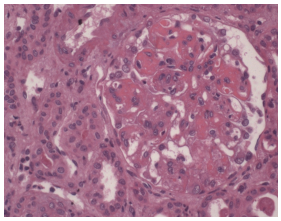
(c) Acumulo de Neutrófilos PSI



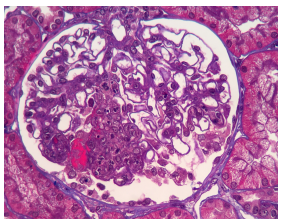
(d) Depositos Hialinos AZAN



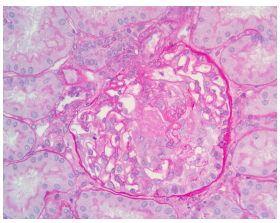
(e) Depositos Hialinos PAS



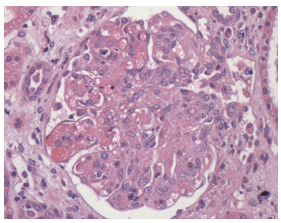
(f) Depositos Hialinos HE



(g) Necrose Fibrinoide AZAN

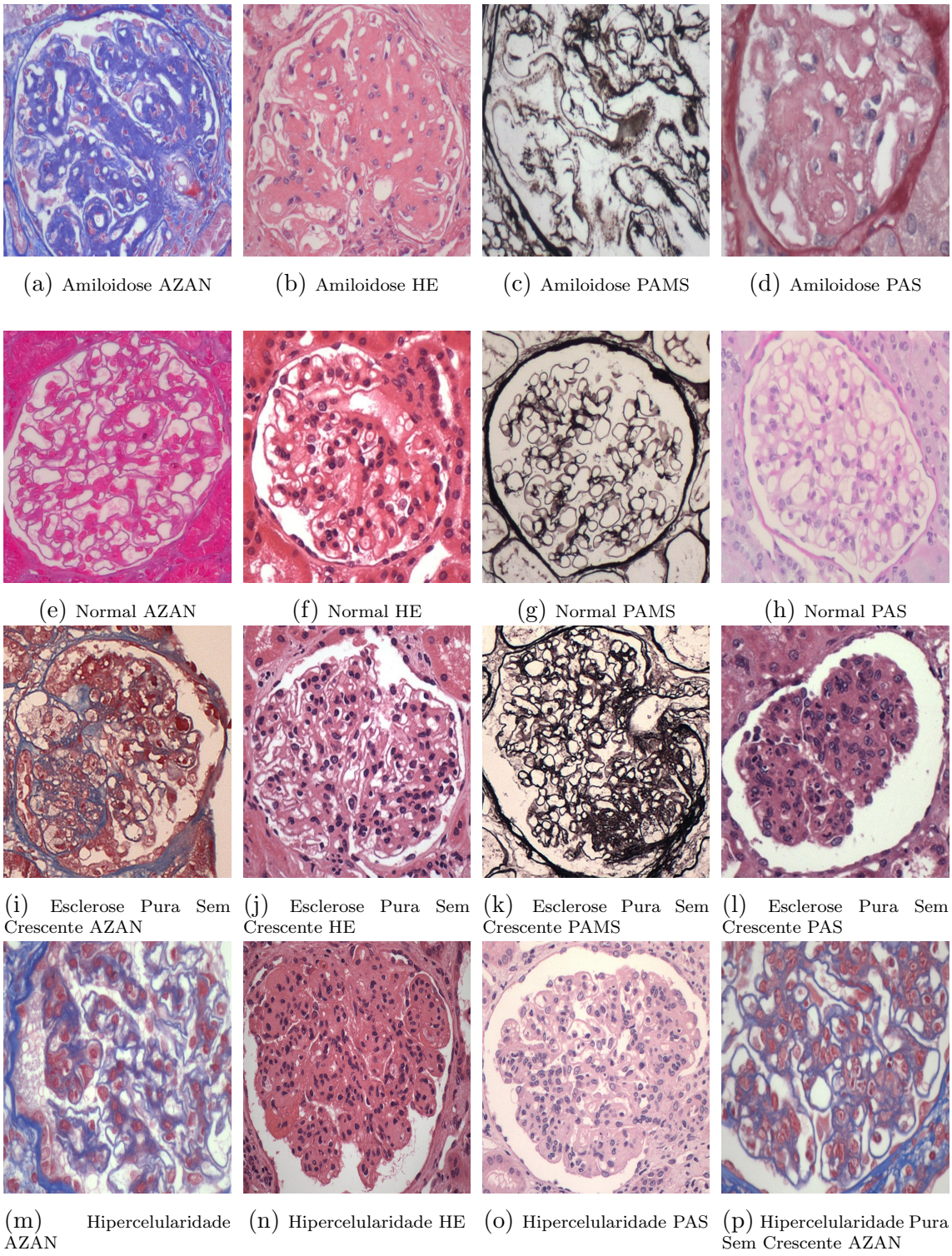


(h) Necrose Fibrinoide PAS



(i) Necrose Fibrinoide HE

Figura 3.2: Amostras do *Dataset* para Acúmulo de Neutrófilos, Depósitos Hialinos e Necrose Fibrinoide com corantes (HE, PSi, PAS, AZAN).



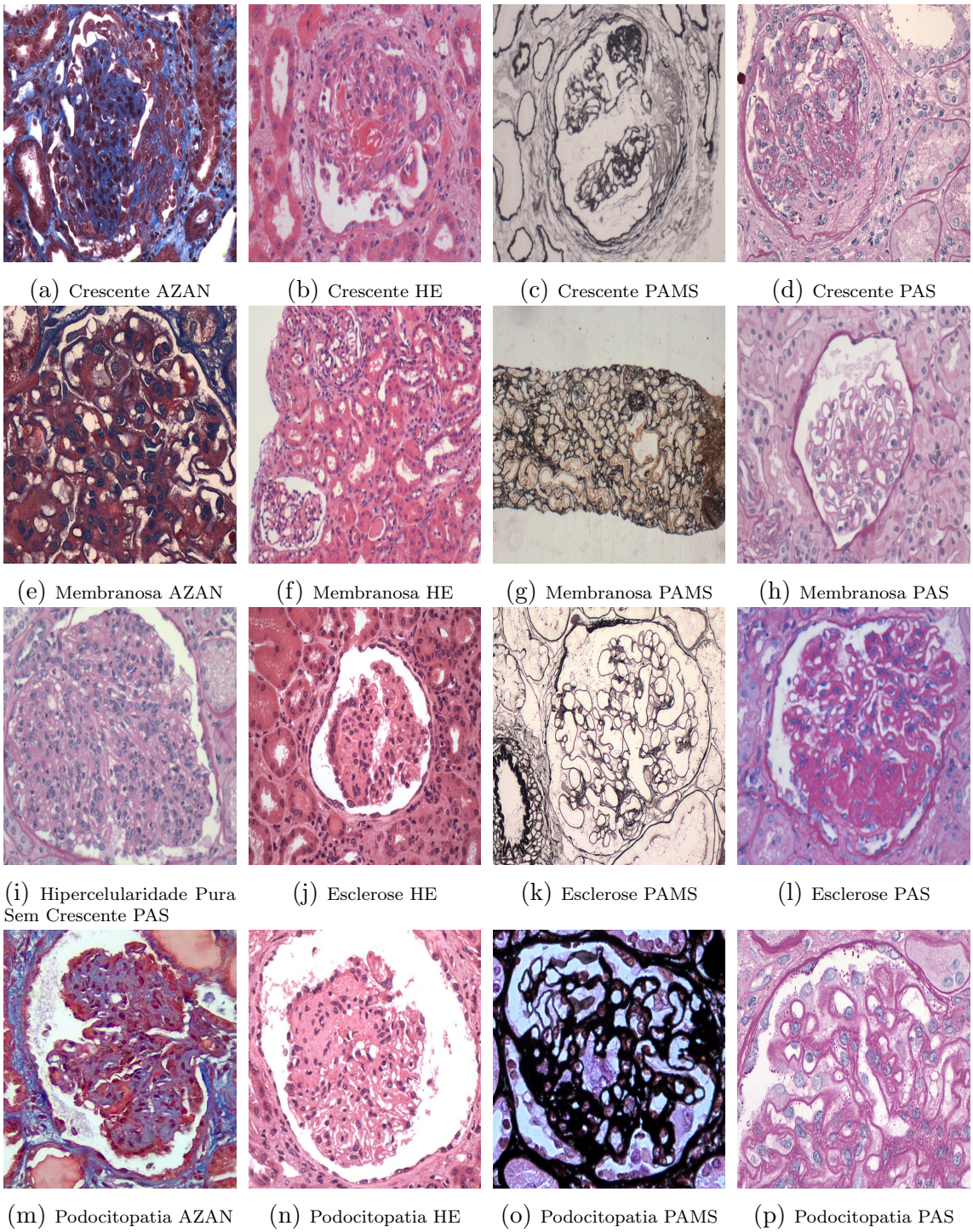
(a) Amiloidose AZAN (b) Amiloidose HE (c) Amiloidose PAMS (d) Amiloidose PAS

(e) Normal AZAN (f) Normal HE (g) Normal PAMS (h) Normal PAS

(i) Esclerose Pura Sem Crescente AZAN (j) Esclerose Pura Sem Crescente HE (k) Esclerose Pura Sem Crescente PAMS (l) Esclerose Pura Sem Crescente PAS

(m) Hipercelularidade AZAN (n) Hipercelularidade HE (o) Hipercelularidade PAS (p) Hipercelularidade Pura Sem Crescente AZAN

Figura 3.3: Amostras do *Dataset* para Amiloidose, Normal, Esclerose Pura sem crescente e Hipercelularidade com corantes (HE, PAS, AZAN, PAMS).



(a) Crescente AZAN (b) Crescente HE (c) Crescente PAMS (d) Crescente PAS

(e) Membranosa AZAN (f) Membranosa HE (g) Membranosa PAMS (h) Membranosa PAS

(i) Hipercelularidade Pura Sem Crescente PAS (j) Esclerose HE (k) Esclerose PAMS (l) Esclerose PAS

(m) Podocitopatia AZAN (n) Podocitopatia HE (o) Podocitopatia PAMS (p) Podocitopatia PAS

Figura 3.4: Amostras do *Dataset* para Crescente, Membranosa, Hipercelularidade, Esclerose e Podocitopatia com corantes (HE, PAS, AZAN, PAMS).

3.3 Infraestrutura de *Hardware* e *Software*

Os experimentos descritos neste trabalho foram executados sobre uma máquina HGX Next, composta por 8 GPUs NVIDIA A100 e 40GB de memória dedicada por GPU, 2 processadores AMD EPYC 7742 com 64 núcleos de 2,25GHz, 1TB de memória RAM e 20TB de armazenamento secundário. O código foi desenvolvido utilizando-se a linguagem Python versão 3.9.5, utilizando-se as bibliotecas PyTorch 2.4.0, Torchvision 0.19.0, NVIDIA Cuda 11.8.87 e Scikit Learn 1.5.1.

3.4 Preparação e Organização do Conjunto de Dados

Esta seção descreve os procedimentos utilizados para organizar o conjunto de dados antes do treinamento dos modelos. As etapas contemplaram o balanceamento das classes por duplicação, a aplicação de transformações durante o pré-treinamento auto-supervisionado, e a divisão estratificada dos dados para avaliação supervisionada.

3.4.1 Aumento de Dados e Balanceamento das Classes

Devido à distribuição desigual das classes no conjunto de dados, foi necessário empregar estratégias de aumento de dados com o objetivo de reduzir o impacto do desbalanceamento durante o treinamento. Três configurações distintas foram comparadas neste estudo: o treinamento sem qualquer aumento (baseline), o aumento obtido por replicação direta das amostras minoritárias e o aumento por transformações controladas das imagens.

Na primeira configuração (baseline), o modelo foi treinado diretamente sobre a distribuição original das classes, permitindo avaliar o desempenho sem intervenções artificiais no conjunto de treinamento. Na segunda, foi aplicado um aumento de dados baseado na simples replicação das amostras pertencentes às classes minoritárias, de modo a equilibrar a frequência entre as categorias. Essa abordagem, embora não introduza novas variações visuais, contribui para mitigar o viés decorrente do desbalanceamento sem modificar a morfologia original das estruturas histológicas, preservando padrões sutis relevantes ao diagnóstico (Hermsen et al., 2019; Stacke et al., 2021b).

Por fim, avaliou-se uma terceira configuração que empregou técnicas de aumento de dados baseadas em transformações controladas das imagens, a fim de ampliar a diversidade visual do conjunto de treinamento. Essa abordagem permitiu comparar o impacto de diferentes estratégias de balanceamento, considerando tanto a preservação morfológica quanto a robustez dos modelos diante de variações estruturais e cromáticas. A decisão por controlar rigorosamente o uso dessas transformações também levou em conta as características dos métodos de aprendizado auto-supervisionado adotados neste estudo, os quais requerem consistência e reprodutibilidade nas trans-

formações aplicadas durante o pré-treinamento (Chen et al., 2020; Grill et al., 2020a; Oquab et al., 2023b).

3.4.2 Transformações e Data Augmentation no Pré-Treinamento e Fine-Tuning

As transformações de imagem, conhecidas como técnicas de *data augmentation*, desempenham papel fundamental tanto no pré-treinamento auto-supervisionado quanto na fase de ajuste fino supervisionado dos modelos utilizados neste estudo. No contexto do aprendizado auto-supervisionado, especialmente em métodos como o DINO, as *augmentations* são responsáveis por gerar múltiplas "vistas" de uma mesma imagem, permitindo que o modelo aprenda representações invariantes a variações irrelevantes — como iluminação, contraste ou pequenas distorções — sem comprometer a semântica morfológica essencial dos glomérulos renais (Chen et al., 2020; Grill et al., 2020a; Caron et al., 2021a; Oquab et al., 2023b). Essa estratégia promove o aprendizado de características robustas e generalizáveis, que servem de base para o posterior fine-tuning supervisionado.

Durante o pré-treinamento com o DINO, foram aplicadas modificações artificiais de imagem para aumento de variabilidade. Entre as principais operações utilizadas destacam-se o *color jittering*, o desfoque gaussiano e a solarização. O *color jittering* introduz variações controladas no brilho, contraste, saturação e matiz, simulando diferenças de coloração entre lâminas histológicas que podem surgir por variações nos processos de fixação e coloração. O desfoque gaussiano, por sua vez, atua reduzindo ruídos de textura e promovendo uma maior ênfase nas estruturas de maior escala, enquanto a solarização inverte parcialmente valores de intensidade de pixels acima de um limiar, estimulando o modelo a tornar-se menos dependente de contrastes absolutos e mais sensível à estrutura contextual (Caron et al., 2021a; Huang et al., 2023).

Essas transformações foram combinadas a recortes aleatórios, redimensionamentos e inversões horizontais, compondo o conjunto de vistas múltiplas exigidas pelo mecanismo de distilação auto-supervisionada proposto pelo DINO. A Figura 3.5 apresenta as variações em função de uma amostra do dataset.

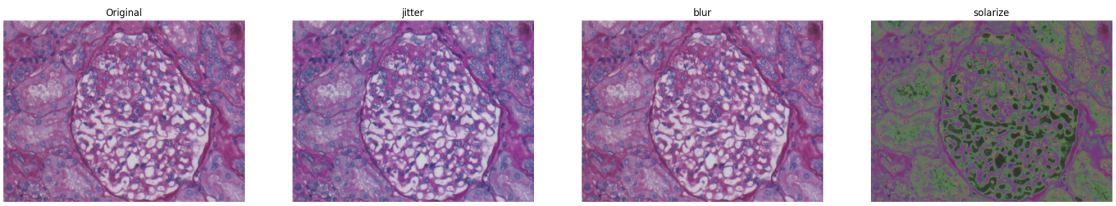


Figura 3.5: Exemplo das transformações aplicadas a uma amostra do Dataset

3.4.3 Divisão dos Dados e Validação Cruzada

Para garantir a comparabilidade entre os experimentos e reduzir a variância nas estimativas de desempenho, adotou-se um protocolo de validação cruzada com cinco subdivisões estratificadas (*5-fold stratified cross-validation*). A escolha por cinco folds é particularmente adequada ao contexto deste estudo, que lida com um conjunto de dados contendo um número reduzido de amostras para várias de suas classes. Esta abordagem assegura que a proporção original entre as classes seja preservada em cada subconjunto, tanto para treinamento quanto para validação, prevenindo vieses de avaliação. (Refaeilzadeh et al., 2009; Arlot e Celisse, 2010).

Cada subdivisão foi utilizada alternadamente como conjunto de validação, enquanto as demais compuseram o conjunto de treinamento. Essa estratégia é amplamente empregada na literatura para avaliar o desempenho generalizável de modelos de aprendizado de máquina, especialmente em cenários com conjuntos de dados limitados ou desbalanceados (Arlot e Celisse, 2010).

Além da avaliação por métricas de desempenho, foram conduzidos testes de significância estatística para comparar os modelos dos experimentos. Especificamente, utilizou-se o teste t de Student para amostras pareadas quando as premissas de normalidade eram atendidas, e o teste de Wilcoxon para amostras pareadas em casos nos quais a normalidade não poderia ser assumida. Esses testes permitem verificar se as diferenças observadas entre os desempenhos dos modelos são estatisticamente significativas, oferecendo maior rigor na interpretação dos resultados.

3.5 Estratégias de Classificação: Abordagens Multiclasse e Binária por Classe

Conforme discutido na revisão bibliográfica, diversas abordagens recentes em aprendizado de máquina aplicado à histopatologia optam por estratégias de classificação multiclasse, nas quais cada imagem é rotulada diretamente com uma entre várias categorias patológicas Ciga et al. (2022); Esteva et al. (2017). No entanto, esse tipo de abordagem tende a ser sensível a desequilíbrios severos entre classes, comuns em bases clínicas reais, dificultando o reconhecimento de padrões relacionados a lesões menos frequentes.

Para investigar a capacidade dos modelos auto-supervisionados em representar adequadamente cada lesão de forma independente, este trabalho também avaliou uma estratégia baseada em classificadores binários especializados, utilizando o esquema um contra todos (*one-vs-all*). Nessa configuração, para cada lesão de interesse, foi treinado um classificador binário no qual a classe positiva correspondia às imagens com a lesão em questão, enquanto todas as demais imagens (de outras lesões ou normais) eram tratadas como classe negativa.

O objetivo dessa abordagem não foi criar um sistema de decisão unificado, mas sim avaliar diretamente a qualidade das representações extraídas por modelos auto-

supervisionados para distinguir cada lesão individualmente. Assim, cada classificador binário serviu como uma ferramenta diagnóstica para estimar o quão bem as features extraídas no pré-treinamento (sem rótulos) capturavam as variações morfológicas relevantes de cada categoria patológica.

Para avaliar o desempenho dos classificadores binários, foi utilizado o F1-score, métrica que computa os verdadeiros positivos, falsos positivos e falsos negativos de forma agregada entre todas as classes antes de calcular a média harmônica entre precisão e revocação. Essa escolha foi motivada por seu comportamento mais robusto em cenários desbalanceados, pois reflete a performance global do modelo considerando o número total de acertos e erros, independentemente da distribuição das classes.

Essa análise permitiu avaliar o potencial discriminativo das representações extraídas por métodos como SimCLR, BYOL e DINO para cada categoria específica de lesão, sem depender exclusivamente de classificações multiclasses complexas e potencialmente enviesadas por classes dominantes.

3.6 Arquiteturas de Backbones Utilizadas

Além das metodologias de aprendizado auto-supervisionado avaliadas neste estudo, a escolha das arquiteturas de *backbone* teve um papel central na definição dos experimentos. Três arquiteturas foram selecionadas por suas características distintas e relevância na literatura biomédica recente: ResNet (He et al., 2016), Vision Transformer (ViT) (Dosovitskiy et al., 2020) e Swin Transformer (Liu et al., 2021).

A *ResNet* foi incluída por ser uma arquitetura clássica baseada em redes convolucionais profundas, amplamente reconhecida por sua robustez em tarefas de classificação de imagens médicas. Seu principal diferencial é o uso de conexões residuais (*skip connections*), que permitem o fluxo direto de gradientes através de múltiplas camadas, mitigando o problema do desaparecimento do gradiente em redes muito profundas (He et al., 2016).

O *Vision Transformer* (*ViT*) foi incorporado por empregar mecanismos de *auto-attention* global, capazes de modelar relações espaciais de longo alcance entre regiões da imagem. Essa característica mostrou-se vantajosa em contextos médicos ao permitir que o modelo capture padrões morfológicos complexos distribuídos em diferentes regiões da imagem (Dosovitskiy et al., 2020; Chen et al., 2020).

Já o *Swin Transformer* foi adicionado à análise por combinar atenção local e global de forma hierárquica, por meio de janelas deslizantes com deslocamento (*shifted windows*). Essa estrutura reduz a complexidade computacional e é especialmente eficaz na análise de imagens de alta resolução, mantendo representações contextualmente ricas e multiescalares (Liu et al., 2021).

A comparação entre essas arquiteturas permitiu avaliar não apenas o impacto das metodologias de aprendizado *auto-supervisionado*, mas também como diferentes pa-

radigmas arquiteturais — convolucionais versus baseados em atenção — influenciam o desempenho em cenários *multiclasse* e *binários*. Tal comparação é relevante, sobretudo em tarefas que demandam sensibilidade à qualidade das representações espaciais e morfológicas, como é o caso da análise de imagens histopatológicas renais.

3.7 Descrição dos Procedimentos Experimentais

Os experimentos seguiram a estrutura apresentada na Seção 3.1, abrangendo as seguintes etapas: (i) preparação dos dados, (ii) pré-treinamento auto-supervisionado, (iii) ajuste supervisionado (*fine-tuning*) e (iv) avaliação com validação cruzada estratificada.

As imagens histopatológicas foram organizadas em diretórios estruturados por classe de lesão (por exemplo, amiloidose e hipercelularidade) e por corante histológico (HE, PAS, AZAN, PAMS). Essa organização permitiu variações experimentais em dois eixos principais: (i) experimentos com um único corante ou com a junção de todos os corantes disponíveis, e (ii) classificações binárias (um-vs-todos) ou multiclasse. A separação dos dados foi feita por meio de validação cruzada estratificada com $k = 5$ folds, aplicada de forma consistente em todos os experimentos supervisionados (ver Seção 2.5.4).

A escolha de $k = 5$ representa um equilíbrio entre custo computacional e confiabilidade estatística da estimativa de desempenho. Valores mais altos de k poderiam oferecer menor viés, porém aumentariam significativamente o tempo de treinamento e agravariam problemas decorrentes do desbalanceamento entre classes. Em particular, o número reduzido de amostras em determinadas categorias limita a viabilidade de divisões mais numerosas, pois a estratificação passaria a gerar folds com número insuficiente de exemplos representativos. Dessa forma, a utilização de cinco folds assegura uma avaliação estável, estatisticamente consistente e adequada à distribuição assimétrica das classes, sendo também uma prática amplamente adotada na literatura de aprendizado de máquina aplicada a dados biomédicos.

As experimentações foram divididas em dois blocos principais:

- **Comparação entre métodos de aprendizado auto-supervisionado:** Nesta etapa, os modelos SimCLR, BYOL e DINO foram avaliados utilizando a arquitetura ResNet-18 como backbone. O objetivo foi comparar a capacidade de extração de representações de cada método mantendo a arquitetura fixa. O mesmo esquema de validação cruzada foi aplicado para os três métodos, assegurando que os conjuntos de treinamento e validação fossem idênticos.
- **Comparação entre arquiteturas de backbone:** Após a identificação do DINO como método com melhor desempenho médio 4.3, ele foi mantido como base para avaliação do impacto de diferentes arquiteturas. Foram avaliadas três famílias de backbones: (i) ResNet, com três variações de profundidade (ResNet-18, ResNet-50 e ResNet-101), (ii) Vision Transformer (ViT-B/16), e (iii) Swin

Transformer, com três versões: Swin-Tiny (Swin-T), Swin-Small (Swin-S) e Swin-Base (Swin-B). Essa comparação teve como objetivo analisar o impacto da profundidade em redes convolucionais e das diferenças entre arquiteturas convolucionais e baseadas em atenção sobre as representações extraídas.

Em todas as configurações, o pré-treinamento foi realizado sem o uso de rótulos, com tarefas pretextuais específicas de cada método. Ao final dessa etapa, os pesos do backbone foram congelados e as representações extraídas foram utilizadas como entrada para classificadores supervisionados do tipo *MLP Head*, treinados separadamente.

Durante o treinamento, adotou-se *early stopping* (com paciência de cinco épocas e delta mínimo de 0,001) tanto no pré-treinamento auto-supervisionado quanto no ajuste supervisionado. A escolha da função de perda foi específica para cada arquitetura, refletindo seus princípios de aprendizado:

- O **SimCLR** empregou uma **função de perda contrastiva** (Equação 2.1), que maximiza a similaridade entre *views* positivas de uma mesma amostra, enquanto repulsa *views* negativas. *Implicação*: Requer *contrastive pairs* e lote (*batch*) grande para eficácia, sendo sensível a *augmentations*.
- O **DINO** empregou uma **função de perda baseada na divergência entre distribuições** (Equação 2.3), que alinha as saídas normalizadas de duas redes neurais: uma rede *estudante* e uma *professora*. A rede professora atua como uma versão suavizada da estudante, sendo atualizada de forma gradual (*momentum update*) a partir dos parâmetros da rede estudante. Essa estratégia viabiliza o processo de *self-distillation*, no qual o modelo aprende a partir de suas próprias previsões, sem a necessidade de rótulos ou exemplos negativos. Além disso, o uso do *momentum* impede o *collapse*, ou seja, evita que todas as representações aprendidas se tornem idênticas, o que garante a diversidade e a utilidade das representações obtidas, impactando positivamente o desempenho final do modelo.
- O **BYOL** utilizou uma **função de regressão baseada no Erro Quadrático Médio (MSE, *Mean Squared Error*)**, que mede a diferença entre as representações produzidas pela rede *online* e pela rede *alvo* (*target*). A rede alvo é atualizada de forma exponencial a partir dos parâmetros da rede online, garantindo uma evolução estável do aprendizado. *Implicação*: essa abordagem elimina a necessidade de pares negativos (como ocorre no SimCLR) e apresenta robustez a ajustes de hiperparâmetros, embora dependa da assincronia entre as duas redes para evitar o colapso das representações.

A diferença entre as funções decorre da natureza distinta de cada método, enquanto abordagens contrastivas (SimCLR) dependem de explicitamente comparar amostras negativas, métodos como DINO e BYOL evitam esse requisito por meio de estratégias preditivas ou de distilação, impactando estabilidade e requisitos computacionais.

Os hiperparâmetros utilizados nos experimentos estão apresentados na Tabela 3.3. Os valores foram definidos com base em padrões das bibliotecas utilizadas e nas recomendações dos artigos originais. Não foi realizado ajuste automatizado de hiperparâmetros. O critério de *early stopping* foi definido como a ausência de melhoria na função de perda do conjunto de validação por 5 épocas consecutivas. O valor da função de perda varia conforme o método de aprendizado utilizado — por exemplo, contraste de similaridade no SimCLR, erro quadrático médio (MSE) no BYOL e perda de entropia cruzada no ajuste supervisionado (*fine-tuning*).

Tabela 3.3: Hiperparâmetros utilizados nos experimentos

Parâmetro	**Valor**
Número de épocas (pré-treinamento)	100
Número de épocas (fine-tuning)	100
Tamanho do *batch*	32
Backbones avaliados	ResNet-18, ResNet-50, ResNet-101, ViT-B/16, Swin-T, Swin-S, Swin-B
Tamanho da imagem (global)	224×224
Tamanho da imagem (local, se aplicável)	96×96
Semente aleatória	42
Validação cruzada	$k = 5$ (estratificada)
Otimizador	Adam
Taxa de aprendizado inicial	1×10^{-3}
Agendador de taxa de aprendizado	StepLR
Redução de taxa (*gamma*)	0,1
Frequência de redução (*step size*)	1/3 do número total de épocas
Critério de *early stopping*	5 épocas sem melhora na função de perda
Delta mínimo de melhoria	0,001
Transformações de *data augmentation*	RandomResizedCrop, ColorJitter, RandomHorizontalFlip, GaussianBlur (definidas conforme cada método SSL)

Todos os experimentos foram executados sob a mesma infraestrutura computacional e ambiente de software. Os resultados foram registrados automaticamente e utilizados para a construção das análises apresentadas nos capítulos seguintes.

3.8 Avaliação de Desempenho

Devido ao desbalanceamento do conjunto de dados, a principal métrica utilizada para avaliação dos modelos foi o F1-Score na forma micro para os classificadores multiclasse e F1-Score para os classificadores binários, calculado tanto de forma agregada quanto por classe (ver Seção 2.5.2). O F1-Score combina precisão e revocação em uma média harmônica, sendo mais informativo que a acurácia em cenários com distribuição desigual de classes Guyon e Elisseeff (2003). Enquanto o F1-Score

micro considera o total de verdadeiros positivos, falsos positivos e falsos negativos de todas as classes de forma conjunta, dando maior peso às classes majoritárias, o F1-Score por classe (ou macro) calcula a média aritmética do F1-Score de cada classe individualmente, tratando todas as classes com a mesma importância. Essa distinção é relevante em conjuntos de dados desbalanceados, pois o F1-Score por classe reflete melhor o desempenho em classes minoritárias, enquanto o F1-Score micro oferece uma visão global do comportamento do modelo em todo o conjunto de dados.

O F1-Score micro considera o total de verdadeiros positivos, falsos positivos e falsos negativos em todas as classes, sendo adequado para medir o desempenho geral do modelo em contextos multiclasse desbalanceados. No entanto, a análise individual do F1-Score por classe também foi realizada com o objetivo de avaliar a performance em categorias específicas, especialmente naquelas com menor representatividade no conjunto de dados.

Essa avaliação por classe permitiu identificar desequilíbrios na capacidade de predição do modelo e observar como diferentes estratégias como classificadores binários impactam o desempenho em classes com menos amostras.

3.8.1 Visualização com t-SNE

Para interpretar visualmente como os modelos organizam as representações das imagens, utilizou-se o t-SNE (t-Distributed Stochastic Neighbor Embedding) van der Maaten e Hinton (2008), uma técnica de redução de dimensionalidade projetada para preservar relações locais entre os dados. Ela transforma distâncias entre pontos em probabilidades de similaridade, tanto no espaço original quanto no espaço projetado, e minimiza a divergência entre essas distribuições para gerar uma visualização intuitiva em 2D ou 3D.

O t-SNE é particularmente útil para avaliar qualitativamente os agrupamentos formados nos espaços latentes, revelando possíveis separações entre classes, sobreposições e padrões estruturais. Neste trabalho, a técnica foi aplicada às representações extraídas pelos modelos auto-supervisionados, permitindo comparar visualmente o quão bem cada método separa as categorias no espaço de embeddings.

Neste capítulo, foram apresentados o conjunto de dados utilizado, a infraestrutura computacional empregada, os procedimentos de pré-processamento e superamostragem, bem como as metodologias adotadas para o pré-treinamento auto-supervisionado e a avaliação dos modelos. Também foram descritas as estratégias de classificação binária e multiclasse, as arquiteturas neurais selecionadas e as técnicas de visualização das representações aprendidas. Essa fundamentação metodológica forneceu a base necessária para a análise dos resultados obtidos, apresentada no capítulo seguinte, no qual se comparam os desempenhos dos diferentes métodos e se discutem suas implicações no contexto da classificação de lesões histopatológicas renais.

Capítulo 4

Resultados

Este capítulo apresenta uma análise detalhada dos experimentos realizados para avaliar métodos de aprendizado autossupervisionado (SimCLR, BYOL e DINO) e diferentes arquiteturas de redes neurais (ResNet e Vision Transformer - ViT) na classificação de imagens histopatológicas de glomérulos renais. Além das arquiteturas e métodos de pré-treinamento, também foram investigadas estratégias de aumento de dados (data augmentation), com o objetivo de ampliar a variabilidade das amostras e reduzir o sobreajuste. Os experimentos foram conduzidos em dois cenários distintos: classificação multiclasse, envolvendo uma classe normal e 11 lesões, e classificação binária, para distinguir entre tecidos normais e patológicos. Ressalta-se que, em alguns experimentos multiclasse, foi considerada uma configuração reduzida com apenas 9 classes, de modo a garantir comparabilidade direta com experimentos previamente realizados e reportados na literatura ou em etapas anteriores deste trabalho de Azevedo et al. (2025), os quais não contemplavam todas as classes atualmente disponíveis no conjunto de dados. Em função do forte desbalanceamento entre classes, a principal métrica de avaliação adotada foi o F1-Score, sendo utilizado o F1-Score micro para os classificadores multiclasse e o F1-Score padrão para os classificadores binários.

4.1 Impacto do Tamanho das ResNets na Performance

Este experimento teve como objetivo investigar se o aumento da profundidade das arquiteturas ResNet — comparando ResNet18, ResNet50 e ResNet101 — resultaria em melhorias significativas no desempenho da classificação de lesões histopatológicas. A motivação central foi avaliar se modelos mais profundos, e portanto mais complexos, seriam capazes de aprender representações mais discriminativas em um cenário caracterizado por volume limitado de dados e forte desbalanceamento entre classes.

Os resultados apresentados na Tabela 4.1 mostram que, embora variações pontuais no F1-score micro tenham sido observadas entre as arquiteturas para algumas classes

específicas, não houve ganhos consistentes associados ao aumento do número de camadas. Essa tendência é corroborada pela análise estatística global apresentada na Tabela 4.2, na qual o teste de Wilcoxon pareado por fold não identificou diferenças estatisticamente significativas entre nenhuma das comparações realizadas (ResNet18 vs. ResNet50: p = 0.0625; ResNet18 vs. ResNet101: p = 0.1875; ResNet50 vs. ResNet101: p = 0.0625).

Tabela 4.1: F1-score (micro, média ± desvio padrão) por classe para diferentes arquiteturas ResNet.

Classe	**ResNet18**	**ResNet50**	**ResNet101**
Amiloidose	74.24 ± 2.82%	75.97 ± 6.80%	71.43 ± 3.62%
Depósitos hialinos	45.71 ± 7.00%	43.75 ± 9.35%	46.67 ± 4.06%
Necrose fibrinoide	22.22 ± 22.22%	22.22 ± 21.17%	0.00 ± 11.18%
Normal	85.42 ± 1.34%	86.88 ± 1.02%	85.66 ± 1.68%
Esclerose p. s. cres.	32.03 ± 4.43%	29.23 ± 5.16%	23.12 ± 5.64%
Hipercelularidade	63.74 ± 1.89%	66.21 ± 1.08%	65.80 ± 1.64%
Hipercelularidade p. s. cres.	12.31 ± 26.04%	3.39 ± 32.41%	4.26 ± 2.04%
Crescente	70.97 ± 2.59%	76.50 ± 2.67%	74.43 ± 3.00%
Membranosa	65.21 ± 2.12%	71.13 ± 2.42%	65.75 ± 3.10%
Esclerose	51.61 ± 6.76%	56.65 ± 6.36%	56.12 ± 5.69%
Podocitopatia	63.74 ± 1.46%	68.57 ± 4.98%	62.35 ± 5.39%
Acúmulo de neutrófilos	78.81 ± 1.43%	79.86 ± 4.43%	78.29 ± 3.37%

Tabela 4.2: Comparação estatística global entre arquiteturas ResNet utilizando o teste de Wilcoxon pareado por fold.

Comparação	**Estatística**	**p-valor**	**Significância**
ResNet18 vs ResNet50	0.0000	0.0625	Não significativa
ResNet18 vs ResNet101	2.0000	0.1875	Não significativa
ResNet50 vs ResNet101	0.0000	0.0625	Não significativa

Essa ausência de melhora pode ser atribuída à natureza do conjunto de dados. Com poucas amostras por classe e presença de desbalanceamento, arquiteturas muito profundas tendem a sobreajustar ao invés de generalizar. Em situações como esta, modelos mais simples podem ser mais eficazes, pois demandam menos dados para aprender padrões relevantes.

As conexões residuais presentes mesmo na ResNet18 já permitem a propagação eficiente de gradientes e o aprendizado de padrões complexos. A adição de mais camadas em redes como a ResNet50 e ResNet101 não se traduziu em melhor desempenho, indicando que o aumento da profundidade, neste caso, não trouxe benefícios práticos.

A ideia de que o ganho incremental "se dilui" significa que, a partir de certo ponto, aumentar a complexidade do modelo gera custo computacional e risco de sobreajuste

sem trazer retorno equivalente em termos de acurácia. Isso sugere que, para este tipo de dataset, a capacidade de aprendizado da ResNet18 já é próxima do ideal, e modelos mais profundos não conseguem extrair informações adicionais relevantes que justifiquem o aumento da complexidade.

Portanto, para o problema e os dados avaliados, o uso de redes mais profundas como a ResNet50 ou ResNet101 não apresentou vantagens claras em relação à ResNet18, reforçando a importância de calibrar a escolha do modelo à escala e à estrutura do conjunto de dados disponível.

4.2 Classificação Multiclasse

Os resultados da tarefa de classificação multiclasse são apresentados em duas tabelas. A Tabela 4.3 resume o desempenho dos modelos baseline supervisionados, enquanto a Tabela 4.4 mostra os resultados dos métodos de aprendizado auto-supervisionado (SSL). Em ambas, são reportados o F1-score por classe positiva (média e desvio padrão obtidos via validação cruzada estratificada em 5 *folds*) e o número de amostras por classe.

Os modelos baseline foram implementados utilizando a arquitetura ResNet-18 como base. Dois cenários distintos foram considerados: (i) inicialização com pesos aleatórios e (ii) utilização de pesos pré-treinados no ImageNet. Em ambos os casos, a última camada da rede, originalmente projetada para a tarefa de classificação no ImageNet, foi removida e substituída por uma nova camada de saída. Essa nova camada foi configurada de acordo com o número de classes da tarefa de classificação de lesões, podendo variar entre abordagens binárias e multiclasse.

Durante o treinamento supervisionado, todas as camadas da rede foram mantidas em modo de atualização. Isso significa que tanto as camadas internas responsáveis pela extração de características visuais quanto a nova camada de saída tiveram seus parâmetros ajustados com base nos dados do conjunto de treino. Essa abordagem de ajuste completo permite que o modelo aprenda representações mais adequadas ao domínio específico das imagens histopatológicas.

O treinamento foi realizado por até 50 épocas, com validação cruzada estratificada em cinco partições. Em cada partição, 10 por cento do conjunto de treino foi reservado para validação. A estratégia de parada antecipada foi aplicada com base na evolução da perda de validação. O otimizador utilizado foi o Adam, com taxa de aprendizado inicial definida em $0,0005$, reduzida automaticamente a cada um terço das épocas totais. O procedimento foi repetido de forma independente para cada partição.

Conforme apresentado na Tabela 4.3, o baseline com inicialização a partir do ImageNet (“Baseline - ImageNet”) apresentou desempenho levemente superior ao baseline com inicialização aleatória (“Baseline”). O F1 Micro agregado foi de 43,47 ± 2,27% e 41,77 ± 2,44%, respectivamente. Entre as classes, observou-se forte variação nos

resultados. A classe "Normal" (com maior número de amostras, 2695) obteve F1 scores elevados quando comparada às demais (86,22 ± 1,11% para o baseline com ImageNet), enquanto classes raras, como "Hipercelularidade Pura Sem Crescente" (224 amostras), apresentaram F1 scores muito baixos (0,51 ± 1,01%).

Na Tabela 4.4, são apresentados os resultados para os métodos auto-supervisionados SimCLR, BYOL e DINO, os quais foram treinados inicialmente com tarefas de contraste usando apenas imagens não rotuladas (pré-treinamento), seguidos por uma etapa de *fine-tuning* supervisionado. O DINO obteve o melhor desempenho geral, com F1 Micro agregado de 58,27 ± 1,76%, superando todos os modelos supervisionados e os demais métodos SSL. O SimCLR apresentou desempenho inferior aos baselines (36,33 ± 1,68%), enquanto o BYOL obteve os piores resultados (8,62 ± 0,55%), com F1 score de 0,00 ± 0,00% em múltiplas classes, indicando falha na generalização do modelo para certas categorias.

Tabela 4.3: F1 score micro por classe para os métodos baseline

Classe	**Baseline**	**Baseline - ImageNet**	**Amostras**
F1 Micro Agregado	**41,77 ± 2,44%**	**43,47 ± 2,27%**	**12.524**
Amiloidose	51,88 ± 3,74%	46,80 ± 17,62%	374
Normal	81,91 ± 0,68%	86,22 ± 1,11%	2695
Hipercelularidade	54,17 ± 1,76%	55,20 ± 4,09%	3134
Hipercelularidade Pura Sem Crescente	1,01 ± 1,23%	0,51 ± 1,01%	224
Crescente	58,39 ± 2,14%	65,30 ± 3,96%	1104
Membranosa	52,16 ± 3,97%	64,55 ± 2,49%	1539
Esclerose	36,63 ± 5,57%	40,60 ± 3,07%	617
Podocitopatia	50,98 ± 5,09%	52,55 ± 5,05%	505
Acumulo neutrofilos	47,32 ± 10,04%	37,73 ± 10,62%	713

Tabela 4.4: F1 score micro por classe para os métodos de aprendizado auto-supervisionado (SSL)

Classe	SimCLR	BYOL	DINO	Amostras
F1 Micro agregado	**36,33 ± 1,68%**	**8,62 ± 0,55%**	**58,27 ± 1,76%**	**12.524**
Amiloidose	47,44 ± 5,00%	0,00 ± 0,00%	78,25 ± 3,25%	374
Normal	80,70 ± 1,28%	41,04 ± 11,20%	91,50 ± 0,97%	2695
Hipercelularidade	64,58 ± 1,23%	45,49 ± 3,51%	69,91 ± 1,00%	3134
Hipercelularidade Pura Sem Crescente	0,00 ± 0,00%	0,00 ± 0,00%	2,99 ± 2,75%	224
Crescente	60,11 ± 1,32%	0,00 ± 0,00%	78,81 ± 0,97%	1104
Membranosa	55,19 ± 1,73%	4,59 ± 1,63%	77,28 ± 1,62%	1539
Esclerose	26,50 ± 2,14%	9,91 ± 7,60%	62,01 ± 3,60%	617
Podocitopatia	52,75 ± 2,50%	2,16 ± 4,32%	70,75 ± 3,40%	505
Acumulo neutrofilos	9,37 ± 7,84%	0,00 ± 0,00%	56,42 ± 7,69%	713

Observa-se em ambas as tabelas uma tendência de F1 scores mais baixos para classes com menor número de amostras. Por exemplo, a classe “Acumulo neutrofilos” (127 amostras) apresentou F1 scores relativamente baixos em todos os métodos (Baseline: 47,32 ± 10,04%, Baseline-ImageNet: 37,73 ± 10,62%, SimCLR: 9,37 ± 7,84%, BYOL: 0,00 ± 0,00%, DINO: 56,42 ± 7,69%). A classe “Hipercelularidade Pura Sem Crescente” (224 amostras) também exibiu F1 scores consistentemente baixos ou nulos em todos os métodos testados. Em contrapartida, classes com maior número de amostras, como “Normal” (2695 amostras) e “Hipercelularidade” (3134 amostras), geralmente apresentaram F1 scores mais elevados na maioria dos métodos.

A superioridade do DINO é particularmente evidente em classes mais frequentes como “Normal” (91,50 ± 0,97%) e “Crescent” (78,81 ± 0,97%), sugerindo que o modelo foi capaz de aprender representações discriminativas mesmo em cenários com desbalanceamento severo. Esses resultados justificam a escolha do DINO como modelo principal nas etapas subsequentes da pesquisa.

4.2.1 Visualização da Representação com t-SNE

A Figura 4.1 apresenta a projeção bidimensional das representações aprendidas pelo modelo SimCLR após a etapa de fine-tuning, utilizando o algoritmo t-SNE (conforme descrito na Seção 3.8.1) para redução de dimensionalidade. Essa visualização permite examinar a organização das amostras no espaço latente, fornecendo indícios sobre a separabilidade entre as classes e os padrões de confusão mais recorrentes.

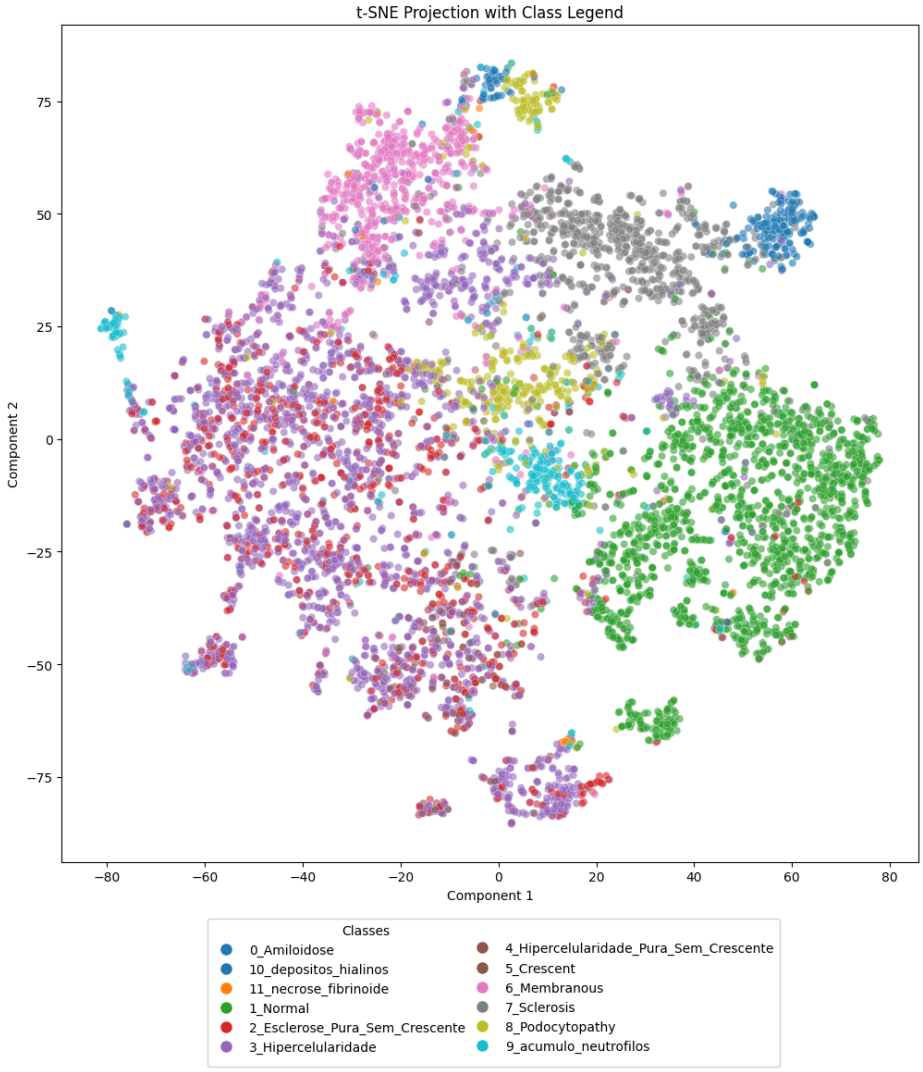


Figura 4.1: Visualização t-SNE das representações geradas pelo modelo SimCLR após *fine-tuning*. As cores indicam diferentes classes histopatológicas. Sobreposições indicam proximidade semântica ou confusão entre categorias.

A análise da Figura 4.1 evidencia três padrões principais. O primeiro refere-se à formação parcial de aglomerados. A classe *Normal* (2695 amostras), representada em verde, apresenta um agrupamento relativamente bem definido e separado das demais classes, condizente com seu elevado desempenho (F1 = 80,70 ± 1,28%). Tal comportamento indica que, para classes com maior representatividade no conjunto de dados, o SimCLR é capaz de gerar representações consistentes e discriminativas.

Em contraste, classes menos representadas, como *Acúmulo de Neutrófilos* (127 amostras) e *Hipercelularidade Pura Sem Crescente* (224 amostras), apresentam dispersão acentuada e sobreposição com outras categorias. No caso de *Hipercelularidade Pura Sem Crescente*, a completa ausência de desempenho (F1 = 0,00 ± 0,00%) corrobora

a dificuldade do modelo em distinguir essa classe no espaço latente. Esse efeito sugere uma correlação direta entre escassez amostral e colapso da separabilidade das representações.

Adicionalmente, observam-se regiões de fronteiras pouco definidas entre classes morfologicamente semelhantes, como Esclerose (617 amostras, F1 = 26,50 ± 2,14%) e Esclerose Pura Sem Crescente (1481 amostras, F1 = 13,75 ± 0,81%). Essa ambiguidade pode ser atribuída não apenas às limitações do SimCLR em capturar nuances histológicas finas, mas também à sobreposição morfológica intrínseca entre essas entidades clínicas.

De forma geral, a visualização reforça uma tendência de o modelo privilegiar padrões extraídos de classes majoritárias, em detrimento daquelas com baixa frequência. Tal viés compromete a equidade do modelo na representação das classes e evidencia a necessidade de estratégias complementares, como técnicas de balanceamento, métodos de oversampling ou arquiteturas de representação mais robustas, para aprimorar a separabilidade das classes minoritárias no espaço latente.

4.3 Classificadores Binários

A simplificação do problema de classificação multiclasse para um conjunto de classificações binárias independentes consiste em treinar modelos específicos para distinguir uma única classe-alvo (positiva) contra todas as demais (negativas). Essa abordagem permite avaliar a capacidade de separação de cada categoria individualmente, mitigando o impacto de desbalanceamentos severos entre classes e oferecendo uma visão mais refinada do desempenho por entidade patológica.

A Tabela 4.5 apresenta os resultados da validação cruzada em 5 *folds* para os modelos DINOv1 e SimCLR, respectivamente. Ambas utilizam como métricas principais o F1-macro, que considera o equilíbrio entre precisão e revocação nas duas classes (positiva e negativa), e o F1-score, essencial para avaliar a capacidade de identificação das lesões de interesse.

Tabela 4.5: Resultados da validação cruzada (5-fold) para classificação binária utilizando os métodos DINO e SimCLR. Apresentando média ± desvio-padrão (%). Colunas adicionais mostram o valor-t emparelhado (t) e o p-valor (p) estimados comparando SimCLR vs DINO (n=5).

Classe	**DINO v1**		**SimCLR**		**t / p (macro)**		**t / p (micro)**	
	F1-macro	**F1-micro**	**F1-macro**	**F1-micro**	**t**	**p**	**t**	**p**
Amiloidose	53,78±4,32%	20,86±4,28%	65,06±2,85%	32,77±5,74%	4.87	0.008	3.72	0.020
Normal	74,40±2,48%	63,26±2,78%	87,71±1,25%	81,37±1,82%	10.72	<0.001	12.19	<0.001
Hipercelularidade	64,48±1,85%	54,55±2,44%	65,37±1,32%	54,13±2,75%	0.88	0.431	-0.26	0.811
Hipercelularidade P. Sem Cresc.	49,36±1,10%	12,56±0,80%	51,86±1,49%	6,76±3,32%	3.02	0.039	-3.80	0.019
Crescente	60,72±2,70%	37,93±2,92%	74,74±2,64%	55,09±4,98%	8.30	0.001	6.65	0.003
Membranosa	59,14±4,86%	37,86±2,64%	66,17±2,82%	44,96±4,14%	2.80	0.049	3.23	0.032
Esclerose	50,21±4,33%	19,90±2,52%	61,81±3,56%	29,00±6,41%	4.63	0.010	2.95	0.042
Podocitopatia	57,96±2,31%	26,49±3,43%	69,69±1,31%	42,11±2,59%	9.88	<0.001	8.13	0.001

A comparação entre os métodos evidencia um desempenho superior do SimCLR em grande parte das classes, tanto em F1-macro quanto em F1-micro. Destacam-se os resultados obtidos para as categorias Crescente (F1-macro = 74,74% ± 2,64; F1-micro = 55,09% ± 4,98) e Normal (F1-macro = 87,71% ± 1,25; F1-micro = 81,37% ± 1,82), nos quais o SimCLR apresentou ganhos significativos em relação ao DINOv1. De forma consistente, classes como Membranosa e Podocitopatia também se beneficiaram da representação contrastiva do SimCLR, com F1-micro superiores a 40% — valores não alcançados pelo DINOv1 nessas mesmas categorias.

Além dos valores médios, os desvios padrão revelam informações relevantes sobre a estabilidade dos modelos. O SimCLR demonstrou menor variabilidade inter-*folds* em diversas categorias — como Podocitopatia (F1-macro = 69,69% ± 1,31) e Normal (F1-macro = 87,71% ± 1,25) — sugerindo maior robustez na generalização.

As diferenças observadas podem ser parcialmente atribuídas à natureza distinta dos métodos. O SimCLR, por empregar uma estratégia de aprendizado contrastivo explícito, favorece a formação de um espaço latente com maior separação entre instâncias de diferentes classes, promovendo fronteiras de decisão mais claras. Já o DINO, baseado em distilação entre visões sem o uso de pares negativos, tende a gerar representações mais suaves e invariantes, o que pode comprometer a discriminação em domínios com alta complexidade morfológica.

Apesar do desempenho superior quando comparado aos classificadores comuns, ambos os métodos apresentaram baixo desempenho em classes como Hipercelularidade Pura Sem Crescente e Esclerose, nas quais o F1-micro ficou abaixo de 30% em ambas as abordagens. Tais resultados indicam limitações na capacidade dos modelos em lidar com lesões de morfologia sutil ou com escassez de amostras representativas, conforme também relatado por (Hermsen et al. (2019)). Esses achados reforçam a importância do uso de estratégias complementares, como técnicas de superamostragem, pré-processamentos específicos por corante ou a incorporação de informações multiescala, a fim de ampliar a sensibilidade dos modelos a subgrupos clinicamente relevantes.

Adicionalmente à análise descritiva, foi realizado um teste t pareado (bicaudal) para verificar a significância estatística das diferenças entre os cenários PAS e PAS Oversample para cada lesão, considerando as médias e desvios-padrão calculados em validação cruzada com cinco *folds* ($n = 5$).

A Tabela 4.7 mostra que a superamostragem aplicada ao conjunto PAS produz melhorias estatisticamente significativas em diversas classes. Em particular, “Amiloidose”, Crescente, “Membranosa”, “Esclerose” e “Normal” apresentaram $p < 0{,}05$ (com valores muito baixos em algumas categorias), indicando ganhos relevantes após o oversample. A classe “Hipercelularidade Pura Sem Crescente” apresentou aumento altamente significativo ($p < 0{,}001$), partindo de média zero para valores substanciais com oversample. Por outro lado, “Hipercelularidade” mostrou uma tendência de melhora que não alcançou significância estatística no teste t adotado ($p = 0{,}0703$).

Esses resultados reforçam que, mesmo quando restrito ao corante PAS, o uso de superamostragem melhora de forma consistente a detecção de lesões raras, reduzindo casos com F1 igual a zero e aumentando a estabilidade entre *folds*.

4.4 Experimentos com Visual Transformers

Nesta seção são apresentados os experimentos realizados com arquiteturas baseadas em *Visual Transformers* (ViT) aplicadas ao método DINO. Diferentemente das etapas anteriores, nas quais classificadores binários independentes foram empregados para cada lesão, adotou-se aqui uma abordagem de **classificação multiclasse**, em que um único modelo foi responsável por distinguir simultaneamente todas as categorias histopatológicas. Essa mudança metodológica não apenas permitiu avaliar a capacidade do ViT em modelar relações espaciais globais e capturar padrões representativos em um cenário com múltiplas classes e forte desbalanceamento de dados, mas também proporcionou uma redução significativa no tempo total de execução dos experimentos.

A arquitetura *ViT-Base* (vit b 16) foi utilizada como backbone principal, com pré-treinamento baseado no método DINO e aplicação de *oversampling* por duplicação de imagens para reduzir a escassez de amostras em classes minoritárias. O treinamento seguiu um esquema de validação cruzada estratificada em cinco divisões (5-fold), e as métricas de desempenho foram reportadas em termos de média e desvio padrão do F1-score macro por classe.

A Tabela 4.6 apresenta os resultados obtidos com o ViT-b-16 em comparação à arquitetura ResNet18, ambas avaliadas sob o mesmo protocolo experimental. Observa-se que o ViT obteve desempenho modesto na maioria das categorias, com F1-scores variando entre 6,90% e 45,71%, evidenciando limitações na generalização, sobretudo em classes menos representadas, como Necrose fibrinoide e Depósitos hialinos. Ainda que o ViT tenha alcançado desempenho relativamente superior em classes como Podocitopatia (45,71%) e Membranosa (34,89%), a diferença para a ResNet18 permanece substancial, uma vez que esta alcançou valores entre 57,63% e 79,11% nas mesmas categorias.

De forma geral, a ResNet18 apresentou F1-scores médios consistentemente mais altos em todas as classes, com destaque para Amiloidose (75,82%) e Crescente (76,44%), indicando maior estabilidade e poder discriminativo. Essa discrepância sugere que, embora o ViT possua uma capacidade teórica superior de modelar dependências de longo alcance, sua eficácia prática foi comprometida pelo tamanho reduzido do conjunto de dados e pelo desbalanceamento acentuado entre as classes. A natureza intensiva em dados dos *Transformers* tende a exigir volumes substancialmente maiores de amostras anotadas para que o modelo possa explorar plenamente seu potencial de representação (Dosovitskiy et al. (2020)).

Tabela 4.6: Comparação de desempenho entre ViT-b-16 (DINO) e ResNet18 (validação cruzada 5-fold). Valores são média ± desvio padrão do F1-score (por classe) e p (teste t pareado, duas caudas).

Classe	**ViT-b-16 (DINO)**	**ResNet18**	**p (ViT vs ResNet18)**
Amiloidose	22,22% ± 6,59%	75,82% ± 4,50%	< 0.001 ($p = 0.0001$)
Depósitos hialinos	19,83% ± 4,50%	57,63% ± 7,04%	< 0.001 ($p = 0.0005$)
Necrose fibrinoide	6,90% ± 2,98%	36,36% ± 6,94%	< 0.01 ($p = 0.0010$)
Hipercelularidade	33,44% ± 12,73%	53,85% ± 4,43%	< 0.01 ($p = 0.0038$)
Hipercelularidade pura sem crescente	18,03% ± 3,73%	28,96% ± 1,54%	< 0.01 ($p = 0.0063$)
Crescente	29,90% ± 10,00%	76,44% ± 3,16%	< 0.001 ($p = 0.0006$)
Membranosa	34,89% ± 5,04%	79,11% ± 2,68%	< 0.001 ($p = 0.0001$)
Esclerose	27,00% ± 8,88%	61,54% ± 4,13%	< 0.01 ($p = 0.0014$)
Podocitopatia	45,71% ± 5,67%	72,04% ± 4,52%	< 0.01 ($p = 0.0013$)
Acúmulo de neutrófilos	34,67% ± 9,42%	59,50% ± 3,50%	< 0.01 ($p = 0.0052$)

Em síntese, os resultados obtidos com o ViT-b-16 indicam que, embora o modelo apresente algum potencial em classes mais representadas, seu desempenho global foi inferior ao da ResNet18 em todas as métricas avaliadas. Esse resultado reforça a necessidade de maior volume de dados e de estratégias de regularização e balanceamento mais sofisticadas para que o ViT atinja desempenho competitivo em contextos biomédicos com distribuição de classes desbalanceada. Diante disso, a ResNet18 manteve-se como a alternativa mais estável e eficiente para o conjunto de dados utilizado neste estudo.

Em complemento às médias e desvios-padrão apresentados na Tabela 4.6, foi conduzido um *teste t* pareado (duas caudas) entre os resultados obtidos com o ViT-b-16 e a ResNet18 para cada classe. Os resultados indicaram diferenças estatisticamente significativas em favor da ResNet18 em todas as classes avaliadas ($p < 0.05$ em todos os casos; ver Tabela 4.6 para valores detalhados). As classes Amiloidose, Membranosa e Crescente apresentaram diferenças altamente significativas ($p < 0.001$), enquanto categorias com maior variabilidade, como "Hipercelularidade" e "Acúmulo de neutrófilos", também exibiram valores de $p < 0.01$. Esses achados confirmam, sob perspectiva estatística, que a ResNet18 supera de forma consistente o ViT-b-16 nas condições experimentais avaliadas, caracterizadas por forte desbalanceamento de classes e número limitado de amostras.

4.5 Análise de Desempenho com Técnicas de Balanceamento de Dados

Nesta etapa, a tarefa de classificação também foi conduzida utilizando classificadores binários independentes, um para cada lesão. Essa escolha metodológica seguiu o mesmo princípio adotado na análise anterior, visando avaliar individualmente o desempenho do modelo em detectar a presença ou ausência de cada classe de interesse. Essa configuração é especialmente importante em cenários com forte desbalancea-

mento entre categorias, como é o caso deste estudo, permitindo uma análise mais controlada do impacto de diferentes estratégias de balanceamento de dados.

Esta análise teve como objetivo avaliar o impacto de diferentes estratégias de organização dos dados no desempenho do modelo DINO com ResNet18. Foram considerados dois eixos principais: (i) a aplicação de superamostragem para lidar com o desbalanceamento entre classes e (ii) a utilização exclusiva de imagens com coloração PAS, tanto isoladamente quanto em conjunto com superamostragem.

A escolha do corante PAS para análises isoladas se justifica por ser o único presente com amostras em todas as classes do conjunto de dados, o que permite a execução de experimentos completos sem exclusão de categorias.

A Tabela 4.7 apresenta os valores de F1-score da classe positiva em quatro cenários distintos: conjunto original (Todos os Corantes), conjunto contendo apenas imagens com coloração PAS (PAS), conjunto com superamostragem das classes minoritárias (Todos os Corantes Oversample) e a combinação de imagens PAS com superamostragem (PAS Oversample).

Observação sobre valores iguais a 0,00%: alguns valores reportados como 0,00% no F1 da classe positiva correspondem a situações em que o modelo não produziu predições positivas em um ou mais *folds* (isto é, revocação = 0), resultando em F1 = 0 naquele(s) fold(s). Em vários casos a média foi arredondada para 0,00% na tabela 4.7, enquanto os desvios-padrão não nulos indicam que houve *folds* com resultados maiores que zero antes do arredondamento. Essas ocorrências são esperadas em cenários de forte desbalanceamento e/ou quando a prevalência da lesão é muito baixa: pequenas variações em um único fold podem levar a mudanças abruptas no F1.

Em geral, a aplicação de superamostragem resultou em aumento do F1-score em praticamente todas as classes. Por exemplo, a classe Normal teve um aumento de 50,68% para 63,26% no cenário com superamostragem, e 56,25% ao combinar PAS com superamostragem, com redução simultânea do desvio padrão, indicando menor variabilidade entre os *folds*.

O cenário PAS Oversample também produziu ganhos em diversas categorias. Por exemplo, Amiloidose passou de 0,00% (baseline) para 38,71%, e Esclerose, de 0,00% para 36,55%. No entanto, algumas classes mantiveram F1-scores baixos mesmo após a aplicação das técnicas. A classe Hipercelularidade Pura Sem Crescente, por exemplo, registrou apenas 18,48% no cenário PAS Oversample.

Tabela 4.7: Desempenho do modelo DINO com ResNet18 na tarefa de classificação binária por lesão, considerando três cenários distintos de treinamento: conjunto original (Baseline), imagens coradas exclusivamente com PAS (PAS) e aplicação de superamostragem nas classes minoritárias (Oversample). Os valores referem-se ao F1-micro da classe positiva para cada categoria, com média e desvio padrão calculados em validação cruzada com 5 *folds*.

Classe	Todos os Corantes	PAS	Todos os Corantes Oversample	PAS Oversample
Normal	50,68 ± 11,20%	41,98 ± 7,95%	63,26 ± 2,78%	56,25 ± 3,93%
Amiloidose	0,00 ± 7,57%	0,00 ± 8,85%	20,86 ± 4,28%	38,71 ± 13,34%
Hipercelularidade	39,17 ± 20,33%	51,40 ± 4,08%	54,55 ± 2,44%	57,14 ± 3,28%
Hipercelularidade Pura Sem Crescente	0,00 ± 0,00%	0,00 ± 0,00%	12,56 ± 0,80%	18,48 ± 2,38%
Crescente	12,14 ± 11,75%	13,64 ± 10,55%	37,93 ± 2,92%	43,90 ± 3,96%
Membranosa	4,88 ± 11,11%	7,79 ± 4,62%	37,86 ± 2,64%	33,49 ± 4,36%
Esclerose	0,00 ± 2,51%	4,35 ± 9,67%	19,90 ± 2,52%	28,79 ± 5,25%
Podocitopatia	8,40 ± 9,60%	19,35 ± 5,40%	26,49 ± 3,43%	29,60 ± 4,39%

A utilização de um subconjunto contendo apenas imagens com coloração PAS, sem aplicação de superamostragem, resultou em variações no desempenho entre as classes. A classe Hipercelularidade apresentou melhora de 39,17% para 51,40%, e Podocitopatia aumentou de 8,40% para 19,35%. Por outro lado, o desempenho da classe Normal caiu de 50,68% para 41,98%, enquanto Amiloidose manteve o F1-score em 0,00%. Esses resultados indicam que a limitação do conjunto de dados ao corante PAS não produziu efeitos consistentes sobre os resultados do pipeline como um todo, afetando de maneira distinta as diferentes categorias.

A aplicação de superamostragem, tanto no conjunto completo quanto no subconjunto com PAS, resultou em ganhos mais sistemáticos. Por exemplo, a classe Esclerose saiu de 0,00% para 19,90% com superamostragem e atingiu 36,55% no cenário PAS com oversample. Esses aumentos sugerem que a replicação de amostras minoritárias contribuiu positivamente para o desempenho final do pipeline, mesmo com o *backbone* treinado de forma auto-supervisionada.

Esses achados reforçam que, mesmo com a adoção de aprendizado auto-supervisionado, o desempenho do modelo permanece sensível às decisões de preparo dos dados, como o tipo de coloração utilizado e o balanceamento entre classes. A influência dessas variáveis foi também destacada por Tellez et al. (Tellez et al. (2019)) e Ciga et al. (Ciga et al. (2022)), que apontam a coloração como um fator que impacta as representações aprendidas por modelos de visão computacional em histopatologia.

Portanto, a superamostragem mostra-se uma técnica eficaz para reduzir os efeitos do desbalanceamento. Já o uso exclusivo de imagens com coloração PAS, embora viabilize a inclusão de todas as classes, não resultou em melhorias consistentes. Estratégias adicionais, como normalização de estilo e treinamento com dados de

múltiplos domínios de coloração, devem ser consideradas para aumentar a robustez dos modelos frente à variabilidade morfológica e técnica das imagens histológicas.

4.6 Impacto do Data Augmentation no Desempenho dos Modelos

O processo de *data augmentation* mostrou-se um componente essencial para a melhoria do desempenho e da capacidade de generalização dos modelos supervisionados. No caso da arquitetura ResNet18, as transformações foram aplicadas de forma explícita durante o treinamento supervisionado e durante o pré treino com SLL, contemplando variações de cor (*Color Jitter*), solarização parcial (*Solarize*) e desfoque gaussiano (*Gaussian Blur*). Essas operações visaram ampliar a diversidade fotométrica e morfológica do conjunto de treinamento.

Foram conduzidos três conjuntos de experimentos distintos. O primeiro consistiu no treinamento com *oversampling* baseado apenas na duplicação de amostras minoritárias, sem qualquer forma de *augmentation*. O segundo experimento introduziu o *data augmentation* de forma ativa, ampliando o conjunto de treinamento para 30.000 amostras e incluindo todas as classes, inclusive a classe *Normal*. Por fim, o terceiro experimento manteve o uso de *augmentation*, mas removeu classes com comportamento ambíguo ou de difícil separação semântica, em todos os experimentos foram utilizados 5 *folds* de validação cruzada.

A Tabela 4.8 apresenta os valores médios de F1-score (média ± desvio padrão) obtidos pela arquitetura ResNet18 em três configurações distintas de treinamento supervisionado. O termo "*Baseline* (sem classe Normal)" indica o treinamento realizado apenas com classes com alguma lesão, excluindo a classe "Normal". Já as colunas "*Augmentation* (todas as classes)" e "*Augmentation* (sem classes confusas)" representam cenários de aumento de dados aplicados, respectivamente, a todas as classes e apenas às classes com anotação única (excluindo "Esclerose pura sem crescente" e "Hipercelularidade sem crescente"). Em todos os casos, o pipeline de treinamento e validação cruzada foi mantido constante, variando apenas a estratégia de balanceamento e de aumento de dados.

Tabela 4.8: Comparação de desempenho (F1-score; média ± desvio padrão) da ResNet18 sob diferentes estratégias de balanceamento e aumento de dados, para um classificador Multiclasse. Colunas adicionais mostram o valor-t emparelhado (t) e o p-valor (p) estimados comparando Baseline vs Aumento (todas as classes) e Baseline vs Aumento (sem classes confusas).

Classe	Baseline (Sem Normal)	Aumento (todas)	t / p	Aumento (sem confusas)	t / p
Amiloidose	75,82 ± 4,50%	80,13 ± 3,60%	t=1,67, p=0,170	83,12 ± 2,53%	t=3,16, p=0,034
Depósitos Hialinos	57,63 ± 7,04%	37,56 ± 11,53%	t=-3,32, p=0,029	47,06 ± 7,08%	t=-2,37, p=0,077
Necrose Fibrinoide	36,36 ± 6,94%	12,50 ± 17,68%	t=-2,81, p=0,048	20,00 ± 17,61%	t=-1,93, p=0,125
Normal	—	87,57 ± 0,01%	—	88,46 ± 0,73%	—
Hipercelularidade	53,85 ± 4,43%	50,05 ± 3,73%	t=-1,47, p=0,216	80,99 ± 0,83%	t=13,47, p<0,001
Hipercelularidade P. Sem Cresc.	28,96 ± 1,54%	16,35 ± 0,18%	t=-18,19, p<0,001	—	—
Crescente	76,44 ± 3,16%	71,46 ± 1,80%	t=-3,06, p=0,038	74,47 ± 2,43%	t=-1,11, p=0,331
Membranosa	79,11 ± 2,68%	69,73 ± 0,31%	t=-7,77, p=0,001	72,15 ± 4,09%	t=-3,18, p=0,033
Esclerose	61,54 ± 4,13%	54,42 ± 0,35%	t=-3,84, p=0,018	55,51 ± 0,82%	t=-3,20, p=0,033
Podocitopatia	72,04 ± 4,52%	63,59 ± 0,74%	t=-4,13, p=0,015	66,34 ± 3,74%	t=-2,17, p=0,096
Acúmulo de Neutrófilos	59,50 ± 3,50%	76,68 ± 4,70%	t=6,56, p=0,003	74,83 ± 0,86%	t=9,51, p=0,001

Os resultados indicam que o uso de *data augmentation* não produziu ganhos uniformes entre todas as classes, mas revelou melhorias estatisticamente significativas em categorias específicas. Em particular, o Acúmulo de Neutrófilos apresentou um aumento expressivo (t=6,56; p=0,003 para o aumento de todas as classes e t=9,51; p=0,001 para o aumento sem classes confusas), evidenciando que a ampliação do conjunto de treinamento favoreceu a generalização em padrões morfologicamente variados. Resultados semelhantes foram observados para Hipercelularidade Pura Sem Crescente (p<0,001) e Membranosa (p=0,033), indicando impacto positivo do aumento de dados mesmo em contextos de classes minoritárias.

Por outro lado, algumas categorias, como Depósitos Hialinos (p=0,029) e Necrose Fibrinoide (p=0,048), apresentaram quedas significativas de desempenho, possivelmente relacionadas à amplificação de ruído ou à sobreposição de padrões entre classes morfologicamente próximas. Classes como Crescente e Podocitopatia também mostraram reduções moderadas, mas sem significância estatística em alguns cenários (p>0,05), sugerindo que o efeito do aumento pode depender da estabilidade intra-classe e da complexidade das fronteiras de decisão.

De forma geral, os resultados confirmam que a introdução de *augmentation* tende a favorecer o aprendizado de representações mais robustas em classes com grande variabilidade histológica, embora sua eficácia dependa fortemente da natureza das transformações aplicadas e da distribuição das amostras. Assim, a definição de uma estratégia de aumento ideal requer ajuste cuidadoso, considerando tanto o equilíbrio entre classes quanto a preservação de características discriminativas sutis.

4.6.1 Análise da Curva de Perda

A Figura 4.2 apresenta a comparação entre as curvas de perda obtidas durante o pré-treinamento do modelo DINOv1 com ResNet18 sob diferentes configurações experimentais. Foram analisados três cenários: um modelo baseline treinado sem

a classe *Normal*, um modelo com *data augmentation* aplicado a todas as classes, e outro com *augmentation* restrito às classes não ambíguas.

Observa-se que o modelo baseline, sem a inclusão da classe *Normal*, apresentou convergência mais lenta e valores de perda substancialmente maiores ao longo das épocas. Em contraste, os modelos com *data augmentation* exibiram uma redução mais acentuada da perda nas primeiras iterações, indicando que as transformações visuais contribuíram para acelerar o processo de otimização e melhorar a estabilidade do aprendizado.

Entre os cenários com *augmentation*, nota-se que a exclusão das classes mais confusas produziu uma curva ligeiramente mais suave, mas com valores finais de perda semelhantes ao modelo com todas as classes. Essa diferença sugere que o ganho de generalização não depende unicamente da exclusão de amostras potencialmente ambíguas, mas também do aumento da variabilidade intra-classe promovido pelas transformações aplicadas.

De forma geral, a presença da classe *Normal* e o uso de *data augmentation* contribuíram para uma convergência mais rápida e estável, além de resultarem em menores valores de perda final. Esses resultados reforçam a importância do enriquecimento das amostras durante o pré-treinamento para favorecer a formação de representações visuais mais discriminativas e robustas.

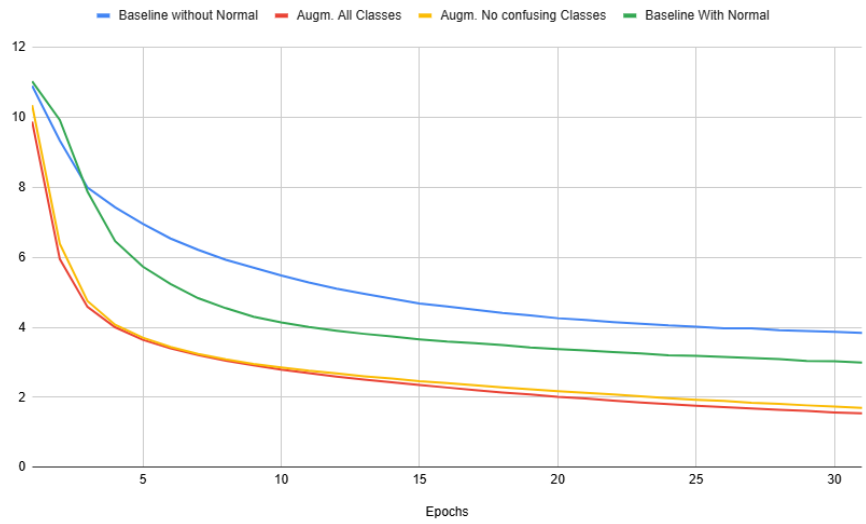


Figura 4.2: Comparação das curvas de perda (*loss*) durante o pré-treinamento do modelo DINOv1 com arquitetura ResNet18, sob diferentes configurações de *data augmentation* e composição de classes. As curvas indicam que os modelos com estratégias de aumento de dados, tanto aplicadas a todas as classes (*Augm. All Classes*) quanto apenas às classes não ambíguas (*Augm. No Confusing Classes*) apresentaram convergência mais rápida e menor perda final. O modelo *Baseline With Normal* também mostrou redução significativa da perda em relação ao *Baseline without Normal*, indicando que a inclusão da classe *Normal* contribuiu para um aprendizado mais estável e representações mais consistentes.

4.7 Combinando métodos SSL-DINO e Few-Shot Learning

A integração entre aprendizado auto-supervisionado (SSL) e estratégias de Few-Shot Learning (FSL) tem sido amplamente investigada na literatura recente (Gidaris et al. (2019); Chen e He (2020); Hu et al. (2022)), uma vez que ambas as abordagens compartilham o objetivo de reduzir a dependência por grandes volumes de dados rotulados. Em particular, métodos baseados em representações auto-supervisionadas, como o *DINO* (Caron et al. (2021b)), demonstraram notável capacidade de extração de características discriminativas e generalizáveis, o que os torna candidatos promissores para cenários de poucos exemplos.

Entretanto, estudos recentes indicam que a simples combinação de representações pré-treinadas via SSL com pipelines de Few-Shot Learning não garante necessariamente ganhos de desempenho (Hu et al. (2022)). Hu et al. (Hu et al. (2022)) demonstraram que o fator determinante para o sucesso em tarefas de FSL não está apenas na qualidade do pré-treinamento, mas também na presença de um estágio posterior de *fine-tuning* adaptativo sobre as classes-alvo. Segundo os autores, pipeli-

nes simplificados compostos por três estágios — pré-treinamento, meta-treinamento e *fine-tuning* — superam métodos meta-aprendizes complexos quando o ajuste fino é devidamente explorado.

Neste trabalho, empregou-se o modelo pré-treinado com o método auto-supervisionado *DINO*, utilizando a arquitetura *ResNet-18* como *backbone*. O modelo foi posteriormente ajustado com um número limitado de amostras por classe, simulando um cenário de Few-Shot Learning. Os resultados obtidos sugerem que, embora o pré-treinamento auto-supervisionado tenha fornecido representações iniciais robustas, a ausência de um processo de *fine-tuning* otimizado sobre as novas classes limitou o potencial de adaptação do modelo, resultando em desempenho inferior ao obtido com o Few-Shot Learning supervisionado. Este comportamento está em conformidade com as observações de Hu et al. (Hu et al. (2022)), que destacam a importância do ajuste fino no sucesso de pipelines híbridos entre SSL e FSL.

O relatório de classificação do modelo DINO-ResNet18 revelou uma **acurácia global de 25,4%**, uma **acurácia balanceada de 0,42** e uma **F1-Score macro de 0,23**. Embora classes, como *Amiloidose*, tenham apresentado valores mais elevados, outras categorias tiveram desempenho bastante limitado, indicando dificuldade do modelo em generalizar adequadamente diante da alta variabilidade morfológica e do desbalanceamento de classes. A Tabela 4.9 apresenta os valores de *F1-score* obtidos para cada classe no cenário DINO + Few-Shot.

Tabela 4.9: F1-score por classe no experimento DINO (ResNet18) + Few-Shot Learning apenas um fold

Classe	**F1-score (%)**
Amiloidose	66.00%
Hipercelularidade	31.00%
Hipercelularidade Pura Sem Crescente	29.00%
Glomerulonefrite Crescente	18.00%
Nefropatia Membranosa	18.00%
Esclerose	43.00%
Podocitopatia	2.00%
Acúmulo de Neutrófilos	1.00%
Depósitos Hialinos	0.00%
Necrose Fibrinoide	0.00%
Média Macro (F1)	**23.00%**
Acurácia Balanceada	**42.00%**
Acurácia Global	**25.00%**

A análise desses resultados demonstra que o desempenho combinado apresentou variação expressiva entre as classes, com bom desempenho em lesões mais representadas no conjunto de dados e desempenho próximo de zero em classes minoritárias. Esse comportamento sugere que o processo de ajuste fino com poucas amostras pode ter levado o modelo a sobreajustar-se a padrões dominantes, perdendo parte da capacidade de generalização adquirida durante o pré-treinamento auto-supervisionado.

Comparativamente, o *Few-Shot Learning* supervisionado isolado apresentou maior equilíbrio entre precisão e revocação, resultando em um desempenho médio superior. Esses resultados indicam que a combinação entre aprendizado auto-supervisionado e *Few-Shot Learning* não garante, por si só, uma melhoria de desempenho, sendo fortemente dependente da compatibilidade entre o pré-treinamento e as características morfológicas das imagens médicas do domínio alvo.

Capítulo 5

Discussão

Este capítulo apresenta uma análise crítica dos resultados obtidos na classificação de imagens histopatológicas de glomérulos renais, com foco nas implicações práticas dos métodos de aprendizado auto-supervisionado e suas limitações em cenários com dados limitados e desbalanceados.

5.1 Eficácia Limitada de Arquiteturas Profundas em Dados Escassos

Os experimentos com diferentes profundidades de ResNet (18, 50 e 101 camadas) revelaram um achado contraintuitivo: o aumento da capacidade representacional não se traduziu em melhorias significativas de desempenho. Como demonstrado na Tabela **??**, a ResNet18 apresentou resultados estatisticamente equivalentes às suas variantes mais profundas, com a ResNet101 chegando a exibir maior variabilidade interexperimental (desvio padrão de 4,82% na classe Amiloidose).

Este fenômeno pode ser atribuído à natureza limitada do conjunto de dados, composto por aproximadamente 11.928 imagens distribuídas de forma desigual entre 12 categorias. Em contextos com escassez amostral, arquiteturas excessivamente complexas tendem a sobreajustar aos padrões dominantes das classes majoritárias, perdendo capacidade de generalização para categorias menos representadas. Este resultado está alinhado com as observações de Chen et al. (2020) (Chen et al. (2020)), que destacam a importância de calibrar a complexidade do modelo à escala dos dados disponíveis.

As conexões residuais presentes mesmo na ResNet18 demonstraram ser suficientes para capturar os padrões morfológicos relevantes nas imagens histopatológicas, enquanto o acréscimo de camadas nas versões 50 e 101 apenas incrementou o custo computacional sem retorno proporcional em acurácia. Esta observação tem implicações práticas imediatas para o desenvolvimento de sistemas de diagnóstico assistido,

como: a redução de custos computacionais e de infraestrutura para treinamento e implantação; a maior acessibilidade para hospitais e laboratórios com menos recursos; e a possibilidade de iterar e validar modelos mais rapidamente em cenários clínicos reais. Isso demonstra que investimentos em arquiteturas mais simples e eficientes podem ser mais vantajosos que a perseguição de modelos profundos em contextos com dados limitados.

5.2 Superioridade do DINO em Cenários Multiclasse Desbalanceados

A avaliação dos métodos de aprendizado auto-supervisionado (SSL) evidencia a superioridade consistente do DINO em relação aos baselines supervisionados, especialmente em cenários multiclasse com forte desbalanceamento de dados. Conforme detalhado na Tabela 5.1, o DINO alcançou F1 Micro agregado de 58,27%, significativamente superior ao Baseline (41,77%), com p-value aproximado de 1,82e-06, indicando que a diferença não se deve à variabilidade entre folds.

A vantagem do DINO foi particularmente notável em classes com representatividade moderada, como **Amiloidose** (78,25%), **Crescent** (78,81%) e **Membranous** (77,28%), cujas diferenças em relação ao Baseline foram estatisticamente significativas ($p < 0,05$). Para classes altamente representadas, como **Normal** (91,50%) e **Hipercelularidade** (69,91%), o DINO também apresentou ganhos significativos, refletindo sua capacidade de gerar representações discriminativas robustas em um espaço latente multiclasse.

Algumas classes minoritárias, como *Hipercelularidade Pura Sem Crescente* (2,99%) e *Acumulo neutrofilos* (56,42%), não apresentaram diferenças estatisticamente significativas (p 0,05), sugerindo que a escassez de amostras limita a capacidade do modelo em superar o Baseline de forma consistente, mesmo com o pré-treinamento auto-supervisionado.

O desempenho superior do DINO pode ser atribuído ao seu mecanismo de destilação de conhecimento entre múltiplas visões da mesma imagem, que dispensa a necessidade de pares negativos explícitos como no SimCLR. Essa característica é especialmente vantajosa em domínios médicos, onde diferenças interclasse são sutis e graduais Caron et al. (2021a). Em contraste, métodos como o BYOL, que dependem exclusivamente de similaridade entre augmentações, podem ser instáveis em contextos altamente desbalanceados, como também observado na literatura Grill et al. (2020b).

Em suma, a Tabela 5.1 não apenas confirma a superioridade do DINO sobre o Baseline em termos de F1-score micro, mas também evidencia, por meio dos p-values aproximados, que os ganhos são estatisticamente robustos para a maioria das classes analisadas. Isso implica que o DINO pode ser uma ferramenta mais confiável e eficiente para auxiliar patologistas na triagem e diagnóstico diferencial

de múltiplas lesões simultaneamente, potencialmente reduzindo o tempo de análise e a subjetividade inerente.

Tabela 5.1: Comparação do desempenho (F1-score) entre o modelo Baseline e o modelo DINO, com p-values aproximados obtidos a partir das médias e desvios padrão dos 5 folds. Valores em **negrito** indicam diferenças estatisticamente significativas ($p < 0{,}05$).

Classe	Média Baseline (%)	Média DINO (%)	p-value
Amiloidose	51,88 ± 3,74%	78,25 ± 3,25%	2.28×10^{-6}
Normal	81,91 ± 0,68%	91,50 ± 0,97%	8.91×10^{-8}
Hipercelularidade	54,17 ± 1,76%	69,91 ± 1,00%	1.22×10^{-7}
Hipercelularidade Pura Sem Crescente	1,01 ± 1,23%	2,99 ± 2,75%	0,180
Crescent	58,39 ± 2,14%	78,81 ± 0,97%	5.11×10^{-8}
Membranous	52,16 ± 3,97%	77,28 ± 1,62%	1.10×10^{-6}
Sclerosis	36,63 ± 5,57%	62,01 ± 3,60%	2.68×10^{-5}
Podocytopathy	50,98 ± 5,09%	70,75 ± 3,40%	9.05×10^{-5}
Acumulo neutrofilos	47,32 ± 10,04%	56,42 ± 7,69%	0,146

5.3 Limitações dos Métodos SSL em Classes Minoritárias

Apesar do desempenho superior do DINO, uma análise por classe revela limitações persistentes na classificação de categorias com escassa representação. Classes como "Hipercelularidade Pura Sem Crescente" (224 amostras) e "Acúmulo de Neutrófilos" (127 amostras) apresentaram F1 scores consistentemente baixos em todos os métodos testados, variando de 0,00% a 56,42%.

A visualização t-SNE das representações do SimCLR (Figura 4.1) corrobora esta limitação, mostrando dispersão acentuada e sobreposição com outras categorias para as classes minoritárias. Este padrão sugere que, mesmo com métodos SSL avançados, a escassez amostral continua sendo um desafio crítico para a formação de representações discriminativas que permitam uma separação clara no espaço latente.

Estes resultados estão alinhados com as observações de Azizi et al. (2021) (Azizi et al. (2021)), que destacam a dificuldade de métodos baseados em contrastes para capturar nuances de classes raras devido à insuficiência de exemplos negativos relevantes durante o pré-treinamento.

5.4 Superioridade do SimCLR em Classificação Binária

A análise dos classificadores binários (Tabela 4.5) revela uma vantagem estatisticamente consistente do SimCLR em relação ao DINOv1 na maioria das classes avaliadas. Em particular, observam-se ganhos expressivos em *Normal* (F1-macro = 87,71% ± 1,25 vs. 74,40% ± 2,48; $t = 10.72$, $p < 0.001$), *Crescent* (74,74% ±

2,64 vs. 60,72% ± 2,70; $t = 8.30$, $p = 0.001$) e *Podocytopathy* (69,69% ± 1,31 vs. 57,96% ± 2,31; $t = 9.88$, $p < 0.001$). Esses valores indicam diferenças altamente significativas, confirmando que o SimCLR oferece desempenho superior de forma estatisticamente robusta nessas classes.

Outras categorias, como *Amiloidose* ($t = 4.87$, $p = 0.008$) e *Membranous* ($t = 2.80$, $p = 0.049$), também apresentaram diferenças significativas, evidenciando que a vantagem do SimCLR não se limita a casos isolados, mas reflete uma tendência generalizada de melhor separação entre as instâncias positivas e negativas. Já em classes como *Hipercelularidade*, o teste t não indicou diferença estatisticamente significativa ($p = 0.431$), sugerindo desempenho equivalente entre os métodos quando a distinção morfológica é menos acentuada.

De forma geral, os resultados do teste t bicaudal ($n = 5$) demonstram que o SimCLR supera o DINOv1 em seis das oito classes analisadas com significância estatística (p < 0,05), tanto em termos de F1-macro quanto de F1-micro. Além disso, a menor variabilidade inter-*folds* observada para o SimCLR em várias classes — como *Normal* e *Podocytopathy* — reforça sua estabilidade e capacidade de generalização.

Esses achados indicam que o mecanismo de aprendizado contrastivo explícito do SimCLR é particularmente eficaz em cenários binários, nos quais o modelo precisa distinguir a presença ou ausência de uma única condição patológica. Ao otimizar o espaço latente para maximizar a distância entre exemplos de classes distintas, o SimCLR favorece a criação de fronteiras de decisão mais bem definidas, resultando em melhor separação entre distribuições positivas e negativas.

Essa constatação está em consonância com observações recentes da literatura. Kang et al. (2023) (Kang et al. (2023)) reportaram que abordagens contrastivas tendem a apresentar desempenho superior em tarefas de detecção binária em imagens médicas, justamente por explorarem a estrutura intrínseca dos dados sem depender fortemente de rótulos balanceados. No presente estudo, o mesmo fenômeno foi observado: a estrutura de contraste par-a-par do SimCLR se mostrou mais adequada para identificar padrões específicos de lesões, especialmente em classes com maior definição morfológica ou maior volume amostral.

Em síntese, os resultados estatísticos e descritivos confirmam que o SimCLR é significativamente mais eficaz que o DINOv1 na configuração binária, fornecendo representações mais discriminativas e consistentes entre folds. Isto sugere uma diretriz metodológica importante: para sistemas de diagnóstico assistido que buscam detectar a presença de uma única lesão específica (triagem para determinada doença), métodos contrastivos como o SimCLR podem ser mais adequados, enquanto para uma classificação diferencial abrangente de múltiplas condições, o DINO se destaca.

5.5 Limitações dos Visual Transformers em Dados Limitados

Os experimentos com Visual Transformers (Tabela 4.6) revelaram uma limitação importante: apesar de seu potencial teórico, o ViT-Base apresentou desempenho inferior à ResNet18 em todas as categorias avaliadas. Em classes como "Amiloidose", o ViT alcançou apenas 22,22% contra 75,82% da ResNet18, enquanto em "Membranous" a diferença foi de 34,89% para 79,11%.

Esta disparidade pode ser atribuída à natureza intensiva em dados dos Transformers, que requerem volumes substanciais de exemplos para explorar plenamente seu mecanismo de autoatenção global. Em conjuntos de tamanho limitado como o utilizado neste estudo (aproximadamente 12.000 imagens), as arquiteturas convolucionais tradicionais como a ResNet18 mantêm vantagem devido à sua inductive bias espacial, que as torna mais eficientes na aprendizagem com dados escassos, conforme discutido por (Dosovitskiy et al. (2020)).

Este resultado tem implicações práticas importantes para o desenvolvimento de sistemas de diagnóstico assistido, sugerindo que, em contextos com dados limitados, investimentos em arquiteturas Transformer podem não ser a abordagem mais eficiente.

5.6 Eficácia de Técnicas de Balanceamento e Augmentation

Os experimentos com técnicas de balanceamento (Tabela 4.7) demonstraram que estratégias simples como oversampling podem produzir melhorias significativas no desempenho. A aplicação de superamostragem resultou em aumentos substanciais de F1-score em praticamente todas as classes, com destaque para "Esclerose Pura Sem Crescente" (de 0,00% para 28,69%) e "Esclerose" (de 0,00% para 19,90%).

A análise das curvas de perda (Figura 4.2) corroborou a eficácia do data augmentation, mostrando convergência mais rápida e estável nos modelos que utilizaram augmentação durante o pré-treinamento. Este resultado está alinhado com as observações de Tellez et al. (2019) (Tellez et al. (2019)), que destacam a importância da diversidade fotométrica e morfológica no treinamento de modelos para histopatologia.

No entanto, os ganhos não foram uniformes em todas as categorias. Classes como "Depósitos Hialinos" e "Hipercelularidade Pura Sem Crescente" apresentaram desempenho inferior com augmentation, sugerindo que transformações agressivas podem introduzir ruído em padrões morfológicos sutis. Esta observação ressalta a necessidade de calibrar as estratégias de augmentação às características específicas de cada classe patológica.

5.7 Implicações para Sistemas de Diagnóstico Assistido

Os resultados obtidos têm implicações importantes para o desenvolvimento de sistemas de diagnóstico assistido em patologia renal:

1. **Seleção de Arquitetura**: Em contextos com dados limitados, arquiteturas convolucionais eficientes como a ResNet18 podem ser mais adequadas que alternativas complexas como Visual Transformers ou ResNets profundas.
2. **Escolha de Método SSL**: O DINO demonstrou ser a abordagem mais robusta para classificação multiclasse, enquanto o SimCLR mostrou vantagem em configurações binárias, sugerindo que a seleção do método deve considerar a natureza da tarefa.
3. **Importância do Balanceamento**: Técnicas simples como oversampling podem produzir melhorias significativas, especialmente para classes minoritárias, com custo computacional relativamente baixo.
4. **Limitações Persistentes**: A classificação de classes raras continua sendo um desafio, indicando a necessidade de estratégias adicionais como aprendizado por transferência multi-domínio ou incorporação de conhecimento especializado.

5.8 Direções Futuras

Com base nas limitações observadas nos experimentos realizados, destacam-se as seguintes direções para trabalhos futuros, visando aprimorar a classificação multilabel em tecidos renais:

1. **Exploração de *Foundational Models***: Investigar a utilização de modelos de grande escala pré-treinados em múltiplos domínios (*foundational models*), que possam fornecer representações mais robustas e transferíveis, aumentando a capacidade de classificação de lesões raras e complexas.
2. **Curadoria e Enriquecimento do Dataset**: Implementar uma curadoria mais profunda das imagens, incluindo revisão de anotações, promovendo uma classificação multilabel mais precisa e confiável.
3. **Abordagens Híbridas de Aprendizado**: Desenvolvimento de pipelines que combinem os benefícios do DINO para classificação multiclasse com a eficácia do SimCLR ou de métodos auto-supervisionados para detecção binária de condições específicas, permitindo um modelo mais adaptável a diferentes tipos de lesões.
4. **Técnicas para Classes Minoritárias**: Explorar meta-aprendizado, aprendizado baseado em protótipos ou geração sintética de amostras para melhorar a representação de classes raras, aumentando a sensibilidade do modelo sem comprometer a precisão global.

5. **Validação Multi-institucional**: Avaliar a generalização dos modelos em conjuntos de dados provenientes de múltiplas instituições, com diferentes protocolos de amostragem e digitalização, garantindo a robustez e aplicabilidade clínica das soluções propostas.

Em síntese, este trabalho demonstra que, embora os métodos de aprendizado autosupervisionado representem avanços significativos na classificação de imagens histopatológicas, seu sucesso prático depende criticamente da adequação da arquitetura e do método às características específicas do conjunto de dados e da tarefa clínica em questão.

Capítulo 6

Conclusões

Este trabalho realizou uma avaliação abrangente de métodos de aprendizado auto-supervisionado (SSL) para classificação de imagens histopatológicas de glomérulos renais, abordando os desafios impostos pelo desbalanceamento severo de classes e pela variabilidade morfológica das lesões. Através de experimentos sistemáticos com diferentes arquiteturas (ResNet, ViT, Swin Transformer) e métodos SSL (SimCLR, BYOL, DINO), foram obtidos insights valiosos sobre o comportamento dessas técnicas em cenários realistas de patologia digital.

A investigação demonstrou que o aumento da complexidade arquitetural não se traduz necessariamente em melhorias de desempenho em contextos com dados limitados. A ResNet18 apresentou resultados equivalentes ou superiores às suas variantes mais profundas (ResNet50 e ResNet101) e ao Visual Transformer (ViT-Base), destacando a importância de calibrar a complexidade do modelo à escala dos dados disponíveis. Este achado tem implicações práticas significativas para o desenvolvimento de sistemas de diagnóstico assistido, sugerindo que investimentos em arquiteturas eficientes podem ser mais vantajosos que a perseguição de modelos excessivamente complexos.

A análise comparativa entre métodos SSL revelou que o DINO apresenta superioridade consistente na classificação multiclasse, alcançando F1 Micro de 58,27% e superando significativamente os baselines supervisionados e outros métodos SSL. No entanto, em configurações binárias, o SimCLR demonstrou vantagem em diversas categorias, sugerindo que a seleção do método deve considerar a natureza específica da tarefa clínica. Esta nuance representa uma contribuição importante para a área, indicando que não existe uma solução única para todos os problemas de classificação em histopatologia.

Os experimentos com técnicas de balanceamento demonstraram que estratégias relativamente simples, como oversampling e data augmentation, podem produzir melhorias substanciais no desempenho, especialmente para classes minoritárias. No entanto, os ganhos não foram uniformes entre todas as categorias, ressaltando a ne-

cessidade de calibrar essas estratégias às características morfológicas específicas de cada entidade patológica.

A construção do conjunto de dados utilizado neste estudo constituiu um elemento central para a realização das análises. O dataset foi organizado de modo a incluir amostras representativas de uma classe e 11 lesões, incorporando ampla variabilidade morfológica e refletindo um cenário de desbalanceamento severo entre classes. Essa estruturação possibilitou avaliar a capacidade dos métodos de aprendizado auto-supervisionado em lidar com a coexistência de categorias altamente prevalentes e outras extremamente raras, condição comum na prática diagnóstica.

Outro aspecto investigado foi a influência dos diferentes corantes histológicos no desempenho dos modelos. Em particular, analisou-se o impacto do uso exclusivo do corante PAS (o único presente em todas as doze classes) tanto isoladamente quanto em combinação com estratégias de superamostragem. Os resultados indicaram que o uso isolado de imagens PAS não gerou melhorias consistentes no desempenho global: algumas categorias apresentaram aumentos relevantes (como Hipercelularidade e Podocytopathy), enquanto outras mostraram queda ou permaneceram com F1 próximo de zero, como Amiloidose e Normal. Por outro lado, a aplicação de superamostragem, tanto no conjunto completo quanto no subconjunto PAS, produziu melhorias amplas e sistemáticas em diversas classes, incluindo ganhos expressivos em categorias raras, como Amiloidose e Sclerosis. Esses resultados evidenciam que, mesmo em pipelines baseados em SSL, o desempenho dos modelos permanece sensível às escolhas relacionadas à preparação do conjunto de dados, como o tipo de coloração empregado e o balanceamento entre classes.

Em termos de contribuições para o estado da arte, este trabalho:

1. Forneceu evidências empíricas robustas sobre a eficácia comparativa de métodos SSL em histopatologia renal, um domínio com desafios específicos de variabilidade morfológica e desbalanceamento;

2. Demonstrou as limitações práticas de arquiteturas profundas e Transformers em cenários com dados limitados, contrariando a tendência predominante na literatura de visão computacional geral;

3. Estabeleceu que a superioridade de métodos SSL é contingente à natureza da tarefa (multiclasse vs. binária), oferecendo diretrizes para seleção de métodos baseada em objetivos clínicos específicos;

4. Identificou limitações persistentes na classificação de classes raras, mesmo com métodos SSL avançados, apontando para a necessidade de abordagens específicas para este desafio.

As implicações práticas deste trabalho estendem-se ao desenvolvimento de sistemas de diagnóstico assistido mais eficientes e clinicamente relevantes. A comprovação de que arquiteturas mais simples podem atingir desempenho equivalente ou superior a alternativas complexas em contextos com dados limitados pode democratizar o

desenvolvimento de soluções de IA para instituições com recursos computacionais restritos. Adicionalmente, a compreensão das vantagens relativas de diferentes métodos SSL em diferentes configurações de tarefa permite decisões mais fundamentadas no projeto de pipelines de classificação para aplicações clínicas específicas.

Para trabalhos futuros, sugere-se a exploração de abordagens híbridas que combinem as vantagens do DINO para classificação multiclasse com a eficácia do SimCLR em detecção binária, além do desenvolvimento de técnicas específicas para melhoria da representação de classes minoritárias, como meta-aprendizado ou geração sintética de amostras. A integração de conhecimento de domínio e informações clínicas complementares representa outra frente promissora para superar as limitações identificadas nos métodos puramente baseados em visão computacional.

Em síntese, este trabalho avançou a compreensão sobre a aplicação de métodos de aprendizado auto-supervisionado em histopatologia renal, fornecendo um conjunto de evidências empíricas e diretrizes práticas que contribuem para o desenvolvimento de sistemas de diagnóstico assistido mais robustos, eficientes e clinicamente relevantes.

Referências

(2023). Kidney biopsy. `https://www.mayoclinic.org/tests-procedures/kidney-biopsy/about/pac-20394494`. Acesso em: junho de 2025.

Arlot, S. e Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79.

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., et al. (2021). Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, páginas 3478–3488.

Bauer, M. e Augenstein, C. (2023). Self-supervised learning in histopathology: New perspectives for prostate cancer grading. In *DAGM German Conference on Pattern Recognition*, páginas 348–360. Springer.

Brixtel, R. et al. (2022). Whole slide image quality in digital pathology: Review and perspectives. *IEEE Access*.

Browne, M. W. (2000). Cross-validation methods. *Journal of mathematical psychology*, 44(1):108–132.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., e Joulin, A. (2021a). Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, páginas 9650–9660.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., e Joulin, A. (2021b). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, páginas 9650–9660.

Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F., Jaume, G., Song, A. H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al. (2024). Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3):850–862.

Chen, T., Kornblith, S., Norouzi, M., e Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, páginas 1597–1607. PMLR.

Chen, X. e He, K. (2020). A new baseline for self-supervised learning and fine-tuning. In *arXiv preprint arXiv:2006.10029*. Available at `https://arxiv.org/abs/2006.10029`.

Ciga, O., Xu, T., e Martel, A. L. (2022). Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198.

de Azevedo, G., Santos, A. S., de Oliveira, L. R., dos Santos, W. L. C., e Duarte, A. A. (2025). Dealing with strong imbalance to classifying glomerular lesions via self-supervised learning. In *Conference on Graphics, Patterns and Images (SIBGRAPI)*, páginas 138–143. SBC.

De Ruijter, T. C., De Hoon, J. P., Slaats, J., De Vries, B., Janssen, M. J., Van Wezel, T., Aarts, M. J., Van Engeland, M., Tjan-Heijnen, V. C., Van Neste, L., et al. (2015). Formalin-fixed, paraffin-embedded (ffpe) tissue epigenomics using infinium humanmethylation450 beadchip assays. *Laboratory investigation*, 95(7):833–842.

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

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

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., e Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118.

Fallatah, M. H., Ahmad, A. M., Alabbas, K. A., e Alhammad, L. A. (2024). An overview of histological staining techniques. *Journal of Posthumanism*, 4(3):1187–1193.

Filiot, A. et al. (2023). Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, páginas 2023–07.

FIOCRUZ (2024). Pathospotter: Histopathological image analysis. `https://pathospotter.bahia.fiocruz.br/`. Accessed: 2024-11-17.

Fogo, A. B. e Kashgarian, M. (2003). Renal biopsy interpretation: overview. *Pathology Patterns Reviews*, 118(3):155–161.

Gadermayr, M. et al. (2022). Stain-adaptive self-supervised learning for histopathology image analysis. *Medical Image Analysis*, 77:102385.

Gidaris, S., Bursuc, A., Komodakis, N., Perez, P., e Cord, M. (2019). Boosting few-shot visual learning with self-supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, páginas 8058–8067.

Grill, J.-B. et al. (2020a). Bootstrap your own latent: A new approach to self-supervised learning. *NeurIPS*.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. (2020b). Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284.

Gurina, T. S. e Simms, L. (2023). Histology, staining. *StatPearls*.

Guyon, I. e Elisseeff, A. (2003). Introduction to the special issue on variable and feature selection. *Journal of machine learning research*, 3:1157–1182.

Haas, M. et al. (2020). Consensus definitions for glomerular lesions by light and electron microscopy: recommendations from a working group of the renal pathology society. *Kidney international*, 98(5):1120–1134.

HaoQuan, XinjiaLi, DayuHu, TianhangNan, e XiaoyuCui (2023). Dual-channel prototype network for few-shot classification of pathological images. *arXiv preprint arXiv:2311.07871*. Preprint.

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

He, W., Liu, T., Han, Y., Ming, W., Du, J., Liu, Y., Yang, Y., Wang, L., Jiang, Z., Wang, Y., et al. (2022). A review: The detection of cancer cells in histopathology based on machine vision. *Computers in Biology and Medicine*, 146:105636.

Hermsen, M. et al. (2019). Deep learning-based histopathologic assessment of kidney tissue. *Journal of the American Society of Nephrology*, 30(10):1968–1979.

Hu, S. X., Li, D., Stühmer, J., Kim, M., e Hospedales, T. M. (2022). Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 9068–9077.

Huang, S.-C., Pareek, A., Jensen, M., Lungren, M. P., Yeung, S., e Chaudhari, A. S. (2023). Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ Digital Medicine*, 6(1):74.

Jin, X., Huang, T., Wen, K., Chi, M., e An, H. (2022a). Histossl: Self-supervised representation learning for classifying histopathology images. *Mathematics*, 11(1):110.

Jin, Y. et al. (2022b). Histossl: Unleashing the power of self-supervised learning for histopathology. *IEEE Transactions on Medical Imaging*, 41(12):3652–3663.

Kang, M., Song, H., Park, S., Yoo, D., e Pereira, S. (2023). Benchmarking self-supervised learning on diverse pathology datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, páginas 3344–3354.

Kumar, N. et al. (2020). Whole slide imaging (wsi) in pathology. *PMC*.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., e Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, páginas 10012–10022.

Maier-Hein, L., Reinke, A., Godau, P., e et al. (2024a). Metrics reloaded: recommendations for image analysis validation. *Nature Methods*, 21(2):195–212.

Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M. D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., et al. (2024b). Metrics reloaded: recommendations for image analysis validation. *Nature methods*, 21(2):195–212.

Oliveira, L., Chagas, P., Duarte, A., Calumby, R., Santos, E., Angelo, M., e dos Santos, W. (2022). Pathospotter: computational intelligence applied to nephropathology. In *Innovations in Nephrology: Breakthrough Technologies in Kidney Disease Care*, páginas 253–272. Springer.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. (2023a). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Oquab, M. et al. (2023b). Dinov2: Learning robust visual features without supervision. *arXiv:2304.07193*.

Pachetti, E. e Colantonio, S. (2023). A systematic review of few-shot learning in medical imaging. *arXiv preprint arXiv:2309.11433*. Preprint.

Refaeilzadeh, P., Tang, L., e Liu, H. (2009). Cross-validation. In *Encyclopedia of database systems*, páginas 532–538. Springer.

Schwarz, A. et al. (2022). Performing an ultrasound-guided percutaneous needle kidney biopsy: An up-to-date procedural review. *Journal of Nephrology*. Review article providing detailed aspects of technique and ultrasound guidance.

Snell, J., Swersky, K., e Zemel, R. (2017). Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, páginas 4077–4087.

Stacke, K. et al. (2021a). A closer look at domain shift for deep learning in histopathology. *Medical Image Analysis*, 73:102152.

Stacke, K., Unger, J., Lundström, C., e Eilertsen, G. (2021b). Learning representations with contrastive self-supervised learning for histopathology applications. *arXiv preprint arXiv:2112.05760*.

Strubell, E., Ganesh, A., e McCallum, A. (2019). Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*.

Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J. N., Wu, Z., e Ding, X. (2020). Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical image analysis*, 63:101693.

Taleb, A., Loetzsch, W., Danz, N., Severin, J., Gaertner, T., Bergner, B., e Lippert, C. (2020). 3d self-supervised methods for medical imaging. *Advances in neural information processing systems*, 33:18158–18172.

Tang, K., Jiang, Z., Wu, K., Shi, J., Xie, F., Wang, W., Wu, H., e Zheng, Y. (2024). Self-supervised representation distribution learning for reliable data augmentation in histopathology wsi classification. *IEEE Transactions on Medical Imaging*.

Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J. M., Ciompi, F., e van der Laak, J. A. (2019). Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. *IEEE Transactions on Medical Imaging*, 38(2):325–336.

Tiard, A. et al. (2022). Stain-invariant self supervised learning for histopathology image analysis. *arXiv preprint arXiv:2211.07590*.

van der Maaten, L. e Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

VuNguyen, PrantikHowlader, LeHou, DimitrisSamaras, RajarsiR.Gupta, e JoelSaltz (2024). Few shot hematopoietic cell classification. *Proceedings of Machine Learning Research (PMLR)227*, páginas 1085–1103.

Wang, J., Quan, H., Wang, C., e Yang, G. (2023). Pyramid-based self-supervised learning for histopathological image classification. *Computers in Biology and Medicine*, 165:107336.

Yang, P., Yin, X., Lu, H., Hu, Z., Zhang, X., Jiang, R., e Lv, H. (2022). Cs-co: A hybrid self-supervised visual representation learning method for h&e-stained histopathological images. *Medical image analysis*, 81:102539.

ZhengruiGuo, ConghaoXiong, JiaboMa, QichenSun, LishuangFeng, JinzhuoWang, e HaoChen (2024). Focus: Knowledge-enhanced adaptive visual compression for few-shot whole slide image classification. *arXiv preprint arXiv:2411.14743*. Preprint.