PETRONIO AUGUSTO DE SOUZA MELO

Concordância entre inteligência artificial e uropatologista expert na detecção e graduação histológica do câncer de próstata em espécimes de prostatectomia radical

São Paulo

2023

PETRONIO AUGUSTO DE SOUZA MELO

Concordância entre inteligência artificial e uropatologista expert na detecção e graduação histológica do câncer de próstata em espécimes de prostatectomia radical

Tese apresentada à Faculdade de Medicina da Universidade de São Paulo para obtenção do título de Doutor em Ciências

Programa de Urologia

Orientadora: Profª. Drª. Katia Ramos Moreira Leite

São Paulo

2023

Preparada pela Biblioteca da
Faculdade de Medicina da Universidade de São Paulo

Melo, Petronio Augusto de Souza
Concordância entre inteligência artificial e uropatologista expert na detecção e graduação histológica do câncer de próstata em espécimes de prostatectomia radical / Petronio Augusto de Souza Melo. -- São Paulo, 2023.
Tese(doutorado)--Faculdade de Medicina da Universidade de São Paulo.
Programa de Urologia.
Orientadora: Katia Ramos Moreira Leite.

Descritores: 1.Neoplasias da próstata 2.Inteligência artificial 3.Aprendizado profundo 4.Aprendizado de máquina 5.Processamento de imagem assistida por computador 6.Redes neurais

USP/FM/DBD-388/23

Responsável: Erinalva da Conceição Batista, CRB-8 6755

DEDICATÓRIA

Antes de tudo, quero dedicar essa tese ao Senhor Deus, o Deus de nossos antepassados, o Todo Poderoso que era, que é e que há de vir. Aquele que tem me ajudado desde o meu nascimento. O Senhor Deus me dá forças para prosseguir no caminho, mesmo quando tudo parece obscuro e tortuoso, quando as coisas não fazem sentido, quando o desejo de desistir é intenso. Mas o caminho é o caminho, e o Senhor Deus permanece ao meu lado, guiando meus passos e carregando-me nos momentos mais difíceis.

Ao meu filho, Isaac Pisicchio Melo, que transformou minha vida por completo após sua concepção e que me deu a honra de ser um pai. Agradeço por ele me inspirar a ser um homem melhor a cada dia. Que eu possa dar o exemplo para o seu crescimento e para as próximas gerações.

Ao meu amado pai, Hamilton Antonio de Melo, agradeço por todo o suporte, em todos os sentidos, que me permitiu chegar até aqui. Com seu amor incondicional, fez tudo o que estava ao seu alcance para que eu me tornasse uma pessoa honesta e trabalhadora. Devo tudo ao senhor. Espero poder inspirar meu filho da mesma forma que o senhor continua a me inspirar.

À minha finada mãe, Maria de Fatima Souza Melo, que foi embora desse mundo há muito tempo, mas que em sua breve jornada conseguiu transmitir valores essenciais. Creio que ela nos observa de onde estiver, e espero estar honrando sua memória com quem me tornei.

À minha esposa, Mariana Pisicchio Melo, que tem sido minha companheira nos últimos anos, e que estará comigo até o fim dos meus dias. Sou grato por sua presença constante, nos momentos alegres e desafiadores, pela constituição de nossa família e pelo presente que é nosso filho, Isaac. Agradeço por ser meu porto seguro, por seu amor e cuidado e por compartilhar sua vida comigo.

AGRADECIMENTOS

Agradeço a minha orientadora, a Prof. Dra. Katia Ramos Moreira Leite, que mais do que uma orientadora, foi uma mãe e amiga para mim nos últimos anos, sendo gentil e paciente em cada etapa deste desafio. Seu profundo conhecimento e paixão pelo assunto foram fontes inesgotáveis de inspiração e motivação. Agradeço a confiança depositada em mim, pelas horas dedicadas à leitura e revisão de cada página, pelas incontáveis e trabalhosas anotações de lâminas histopatológicas digitalizadas e pelas valiosas críticas e sugestões que enriqueceram imensamente o meu trabalho. Sua capacidade de me guiar com sabedoria, corrigindo-me quando necessário, mas também celebrando cada conquista, fez toda a diferença em minha jornada acadêmica. Agradeço por ter me acolhido em momentos de dúvida, incentivando-me a seguir em frente mesmo nos momentos mais difíceis. O caminho do doutorado é repleto de altos e baixos, mas sob sua orientação, fui capaz de perseverar e alcançar o objetivo final. Por tudo isso e muito mais, minha eterna gratidão.

Agradeço também a Dra. Carmen Liane Neubarth Estivallet, patologista, que me auxiliou muito em nosso projeto, pela sua dedicação, contribuindo com sua expertise para que esse trabalho fosse executado.

Aos senhores Dr. Matheus Cardoso Moraes, Dr. Vinicius Meneguette Gomes de Souza e Maira Suzuka Kudo, meu reconhecimento pelo apoio no início dessa jornada. Seus conselhos e suporte foram cruciais para alicerçar meu caminho no doutorado.

Ao amigo Fabio Leme Ortega, que inspirou a execução desse projeto com suas ideias visionárias, minha gratidão. Igualmente, ao sempre prestativo Iran Silva do LIM 55, que sempre me ajudou nas mais diversas situações, sou grato.

Minha gratidão também se estende aos amigos virtuais ao redor do mundo que compartilham seu conhecimento na internet, em especial aos que contribuem em plataformas como GitHub, Kaggle, Coursera e YouTube. Seus ensinamentos foram essenciais para a realização deste projeto.

À Universidade de São Paulo, a todos os seus colaboradores e, em especial, ao LIM 55 e sua dedicada equipe, pela oportunidade de realizar o curso de doutorado.

EPÍGRAFE

"Não sei como posso parecer ao mundo, mas para mim mesmo, sinto como se fosse apenas um menino brincando à beira-mar, entretendo-me de vez em quando ao encontrar uma pedrinha mais lisa ou uma concha mais bonita que o comum, enquanto o vasto oceano da verdade permanece todo inexplorado diante de mim."

Isaac Newton

Esta dissertação ou tese está de acordo com as seguintes normas, em vigor no momento desta publicação:

Referências: adaptado de *International Committee of Medical Journals Editors* (Vancouver).

Universidade de São Paulo. Faculdade de Medicina. Divisão de Biblioteca e Documentação. *Guia de apresentação de dissertações, teses e monografias*. Elaborado por Anneliese Carneiro da Cunha, Maria Julia de A. L. Freddi, Maria F. Crestana, Marinalva de Souza Aragão, Suely Campos Cardoso, Valéria Vilhena. 3a ed. São Paulo: Divisão de Biblioteca e Documentação; 2011.

Abreviaturas dos títulos dos periódicos de acordo com *List of Journals Indexed in Index Medicus*.

SUMÁRIO

LISTA DE ABREVIATURAS

ASAP	Proliferação Atípica de Pequenos Ácidos
CAAE	Certificado de Apresentação de Apreciação Ética
CaP	Câncer de próstata
CNN	Rede neural convolucional
DHT	Diidrotestosterona
EUA	Estados Unidos da América
FCN	Rede neural completamente convolucional
GPU	Unidade de processamento gráfico
H&E	Hematoxilina e eosina
HIFU	Ultrassom focalizado de alta intensidade
IA	Inteligência artificial
INCA	Instituto Nacional de Câncer
ISUP	International Society of Urological Pathology
LIME	Decomposição em camadas
LPCM	Laboratório de Patologia Cirúrgica e Molecular de São Paulo
LRP	Propagação de relevância
R-CNN	Region-based convolutional neural network
ROC-AUC	Característica de Operação do Receptor - Área Sob a Curva
RoI	Região de interesse
PSA	Antígeno específico da próstata

ReLU	Ativação linear retificada
RNAs	Redes neurais artificiais
RPN	Region Proposal Network
RNMmp	Ressonância magnética multiparamétrica
SNPs	Polimorfismos de nucleotídeo único
TDA	Terapia de deprivação androgênica
WHO	World Health Organization

LISTA DE TABELAS

LISTA DE FIGURAS

RESUMO

Melo PAS. Concordância entre inteligência artificial e uropatologista expert na detecção e graduação histológica do câncer de próstata em espécimes de prostatectomia radical [tese]. São Paulo: Faculdade de Medicina, Universidade de São Paulo; 2023.

OBJETIVOS: Desenvolver um modelo de inteligência artificial (IA) para identificação e graduação histológica do câncer de próstata (CaP) e comparar sua precisão com a avaliação de um uropatologista especializado. MATERIAL E MÉTODOS: 36 lâminas histológicas provenientes de espécimes de prostatectomia radical foram escaneadas, segmentadas, analisadas e anotadas por um uropatologista experiente. As seguintes áreas foram caracterizadas: estroma, glândulas normais, Gleason 3, 4 e 5. Dois métodos foram utilizados para classificar as imagens: um método de classificação categórica que consistia em atribuir um único rótulo a cada imagem analisada e um método de segmentação de instâncias que envolvia o delineamento específico de cada área de interesse em cada imagem, em vez de atribuir um único rótulo a toda a imagem. As arquiteturas utilizadas para o treinamento dos métodos foram a VGGNet, um tipo de rede neural convolucional, e Mask R-CNN (*region-based convolutional neural network*), respectivamente. Após o treinamento, imagens que não foram utilizadas no treinamento foram avaliadas pelo modelo. O modelo que apresentou melhor performance também foi aplicado para avaliar lâminas de biópsia de próstata que não foram usadas no treinamento. RESULTADOS: O modelo de classificação categórico utilizou em seu treinamento 2.583 imagens, sendo 662 imagens com tecido benigno, 640, 640 e 641 imagens, com câncer de próstata padrão 3, 4 e 5 predominante, respectivamente. Esse modelo apresentou uma acurácia de treinamento de 97,57% e uma acurácia de validação de 95,74%. No entanto, ao utilizar o modelo treinado no conjunto de teste com 100 novas imagens nunca usadas no treinamento, a acurácia obtida foi de 45% na determinação de tecido benigno e maligno, e quando maligno, em definir corretamente sua graduação de Gleason. No método de segmentação por instâncias, foram utilizadas 6.160 imagens e 8.367 anotações realizadas nessas imagens (glândulas normais: 3.982, estroma: 3.049, Gleason 3: 858, Gleason 4:

2.321, Gleason 5: 1.361). O modelo apresentou uma acurácia de treinamento de 95,1% e acurácia de validação de 93,2%. Após o término do treinamento, o modelo foi exposto a imagens nunca vistas, separadas antes do treinamento e não usadas para esse fim. A concordância entre o modelo de aprendizado profundo e a anotação feita pelo patologista na detecção correta de tecido benigno e maligno, incluindo seus padrões, foi de 89%. Por fim, aplicamos o modelo de Mask R-CNN em 292 lâminas de biópsia de próstata. O coeficiente de Dice entre o modelo de IA e o especialista em uropatologia foi de 0,7962. CONCLUSÕES: Algoritmos de aprendizagem profunda tem um alto potencial para uso no diagnóstico e graduação do câncer de próstata. O método de segmentação de instâncias foi superior ao método de classificação categórico.

Palavras-chave: Neoplasias da próstata. Inteligência artificial. Aprendizado profundo. Aprendizado de máquina. Processamento de imagem assistida por computador. Redes neurais.

ABSTRACT

Melo PAS. Agreement between artificial intelligence and expert uropathologist in the detection and histological grading of prostate cancer in radical prostatectomy specimens [thesis]. São Paulo: “Faculdade de Medicina, Universidade de São Paulo”; 2023.

OBJECTIVES: To develop an artificial intelligence (AI) model for identification and histological grading of prostate cancer (PCa) and to compare its accuracy with the assessment of a specialized uropathologist. MATERIALS AND METHODS: 36 histological slides from radical prostatectomy specimens were scanned, segmented, analyzed, and annotated by an experienced uropathologist. The following areas were characterized: stroma, normal glands, Gleason 3, 4, and 5. Two methods were employed to classify the images: a categorical classification method, which involved assigning a single label to each analyzed image, and an instance segmentation method that entailed the specific delineation of each area of interest in each image, rather than assigning a single label to the whole image. The architectures used for training the methods were VGGNet, a type of convolutional neural network, and Mask R-CNN (region-based convolutional neural network), respectively. After training, images not used in the training were assessed by the model. The top-performing model was also applied to evaluate prostate biopsy slides not utilized in the training. RESULTS: The categorical classification model used 2,583 images for training, with 662 images of benign tissue, 640, 640, and 641 images of prostate cancer with predominant grades 3, 4, and 5, respectively. This model exhibited a training accuracy of 97.57% and a validation accuracy of 95.74%. However, when applying the trained model to a test set of 100 new images never used in training, the accuracy was 45% in determining benign and malignant tissues, and when malignant, in correctly defining its Gleason grade. In the instance segmentation method, 6,160 images and 8,367 annotations on these images were used (normal glands: 3,982, stroma: 3,049, Gleason 3: 858, Gleason 4: 2,321, Gleason 5: 1,361). The model achieved a training accuracy of 95.1% and a validation accuracy of 93.2%. After training, the model was exposed to previously unseen images set aside before training and not used for that purpose. The agreement between the deep learning

model and the pathologist's annotation in correctly detecting benign and malignant tissues, including their patterns, was 89%. Finally, we applied the Mask R-CNN model to 292 prostate biopsy slides. The Dice coefficient between the AI model and the uropathology specialist was 0.7962. CONCLUSIONS: Deep learning algorithms possess high potential for use in the diagnosis and grading of prostate cancer. The instance segmentation method proved superior to the categorical classification method.

Keywords: Prostatic neoplasms. Artificial intelligence. Deep learning. Machine learning. Computer-assisted image processing. Neural network computer.

1. INTRODUÇÃO

1.1 Epidemiologia do câncer de próstata

O câncer de próstata (CaP) é uma das neoplasias malignas mais prevalentes em homens ao redor do mundo, representando uma preocupação crescente na saúde pública global. Com uma estimativa de quase 1,4 milhão de novos casos e 375.000 mortes em todo o mundo, o CaP é o segundo câncer mais frequente e a quinta principal causa de morte por câncer entre os homens em 2020. As taxas de incidência são três vezes maiores em países desenvolvidos do que em países em desenvolvimento (37,5 e 11,3 por 100.000, respectivamente); enquanto isso, as taxas de mortalidade são menos variáveis, com 8,1 e 5,9 por 100.000, respectivamente. É o câncer diagnosticado com mais frequência em homens em mais da metade (112 de 185) dos países do mundo (1).

Fatores de risco conhecidos para o desenvolvimento desta neoplasia incluem idade avançada, histórico familiar de CaP, predisposição genética e fatores ambientais e comportamentais (2, 3).

A incidência do CaP varia muito entre as diferentes regiões geográficas do mundo. No Brasil, o CaP é o tipo de câncer mais frequente em homens, excluindo-se os tumores de pele não melanoma (4). A estimativa do Instituto Nacional de Câncer (INCA) aponta que, em 2020, ocorreram cerca de 65.840 novos casos da doença no país, correspondendo a uma taxa de incidência bruta de 62,95 casos por 100 mil homens (4). A mortalidade por CaP no Brasil tem aumentado nas últimas décadas, embora a taxa de sobrevida geral também tenha apresentado melhora, especialmente devido aos avanços no diagnóstico e tratamento (5).

Os gastos com o diagnóstico e tratamento do CaP são exorbitantes. Nos Estados Unidos, estima-se que as despesas com o tratamento do CaP em 2006 foram de 9,86 bilhões de dólares. O custo do tratamento no primeiro ano após o diagnóstico foi calculado em torno de 21.040 dólares por paciente, aumentando em relação ao maior estádio tumoral (6).

1.2 Etiopatogenia do câncer de próstata

A etiopatogenia do CaP é complexa e multifatorial, envolvendo interações entre predisposição genética, fatores hormonais, ambientais e comportamentais.

1.2.1 Predisposição genética

Estima-se que cerca de 5% a 10% dos casos de CaP sejam hereditários, sendo a predisposição genética um importante fator de risco (7). Vários genes e variantes genéticas foram associados ao aumento do risco de CaP, incluindo mutações em genes de alta penetrância, como BRCA1, BRCA2 e HOXB13, e polimorfismos de nucleotídeo único (SNPs) em regiões do genoma associadas ao CaP (8, 9).

1.2.2 Fatores hormonais

Os hormônios androgênicos, como a testosterona e seu metabólito ativo diidrotestosterona (DHT), têm sido implicados na progressão do CaP. No entanto, a relação entre os níveis de testosterona e o risco de desenvolvimento de CaP é complexa e ainda objeto de debate na literatura científica (10, 11). Estudos mostram que a inibição da via de sinalização dos androgênios pode reduzir a incidência de CaP e retardar sua progressão (12).

1.2.3 Fatores ambientais e comportamentais

Fatores ambientais e comportamentais têm sido investigados como possíveis contribuintes para o risco de CaP, mas os dados disponíveis ainda não são conclusivos. A dieta tem sido estudada como um possível fator de risco modificável. Alguns estudos sugerem que dietas ricas em gorduras saturadas e carnes vermelhas podem estar associadas a um maior risco de CaP, enquanto dietas ricas em frutas, vegetais e peixes podem estar relacionadas a um menor risco (13). No entanto, ainda não há dados conclusivos que possam fundamentar medidas preventivas ou dietéticas específicas para reduzir o risco de se

desenvolver CaP (14). Além disso, a relação entre a obesidade, o sedentarismo e o tabagismo com o risco de CaP e o prognóstico ainda não é totalmente estabelecida (15, 16).

1.3 Apresentação clínica do câncer de próstata

O CaP pode ser assintomático em seus estágios iniciais, sendo frequentemente descoberto durante exames de rotina, como o teste de antígeno prostático específico (PSA) e o exame de toque retal (14). À medida que a doença progride, os pacientes podem apresentar uma variedade de sinais e sintomas. Os sinais e sintomas locais incluem disúria, polaciúria, noctúria, hesitação urinária, retenção urinária e hematúria (17).

Além dos sintomas locais, os pacientes com CaP avançado ou metastático podem apresentar sinais e sintomas sistêmicos, como dor óssea (resultante de metástases ósseas), fraqueza (devido à anemia ou comprometimento geral do estado de saúde) e perda de peso involuntária (18).

1.4 Diagnóstico do câncer de próstata

O diagnóstico do CaP envolve uma combinação de métodos clínicos, laboratoriais e de imagem. Inicialmente, o exame de toque retal e a dosagem sérica do PSA são utilizados como ferramentas de rastreamento para identificar homens com maior risco de desenvolver a doença (14). Quando há suspeita clínica de CaP, geralmente decorrente de um PSA elevado ou alterações no exame de toque retal, é indicada a realização da biópsia prostática. A biópsia é frequentemente guiada por ultrassonografia transretal ou transperineal e pode ser feita através de diferentes abordagens, como a sistemática ou a dirigida por fusão de imagens (19).

A análise histopatológica confirmará a presença de neoplasia e determinará a agressividade do tumor. O escore de Gleason, baseado na arquitetura das glândulas neoplásicas, é uma ferramenta amplamente utilizada para a classificação da agressividade do CaP e é o dado mais importante para a tomada de decisões clínicas (20).

O diagnóstico preciso do CaP é fundamental para o manejo adequado da doença e a determinação do prognóstico. No entanto, a avaliação histopatológica é um processo subjetivo e sujeito a variações inter e mesmo intraobservador (21). Portanto, a busca de metodologias para o auxílio do patologista na identificação e graduação do CaP é de fundamental importância.

1.5 Escore de Gleason

O Escore de Gleason é o principal fator prognóstico do CaP, utilizado para a previsão da agressividade do tumor e decisão terapêutica (20). Baseado na arquitetura das glândulas neoplásicas (22), o sistema de graduação de Gleason classifica o tumor em padrões que recebem uma pontuação de 1 a 5, considerando a heterogeneidade e multicentricidade características da neoplasia. Devem ser considerados os dois padrões predominantes que comporão o escore, variável de 2 a 10. Se somente um padrão estiver presente, ele deve ser dobrado para a nota final do escore. Quando há três padrões, o escore utiliza os dois padrões mais comuns, se maiores que 5%. Se houver um padrão de maior agressividade ≤5% será considerado como padrão terciário ou de menor proporção.

Originalmente, o escore de Gleason variava de 2 (1+1) a 10 (5+5). Porém, após o consenso da *International Society of Urological Pathology* (ISUP) de 2005 mostrou-se que os padrões 1 e 2 provavelmente correspondiam a lesões benignas da zona de transição (23). A soma das pontuações dos dois padrões dominantes resulta no escore de Gleason. Tumores com escores mais altos indicam maior agressividade e pior prognóstico (20). Assim, o escore de Gleason atualmente é variável de 6 a 10. Em 2014, a ISUP propôs uma nova graduação, variável de 1 a 5 que tem sido utilizada ao lado da graduação de Gleason (Figura 1).

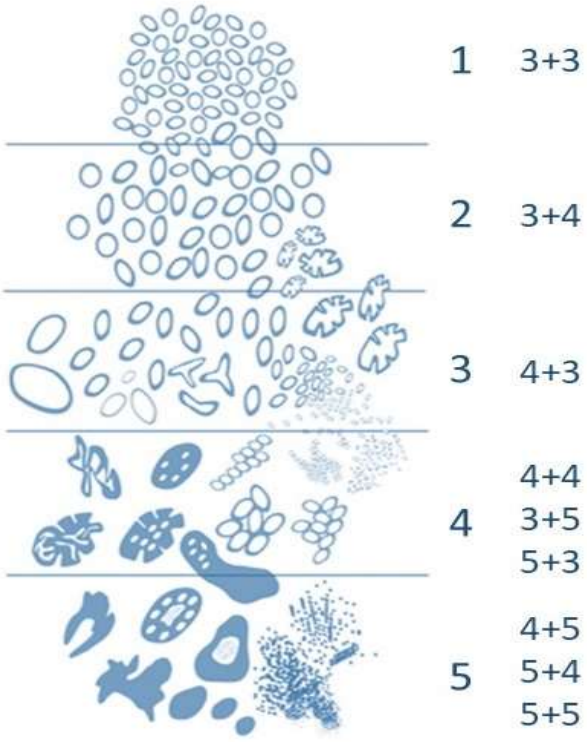

Figura 1. Graduação WHO / ISUP e a correspondência com a graduação de Gleason.

A graduação histológica é o fator prognóstico mais importante do CaP. Os tumores com escore de Gleason 6 / ISUP 1 apresentam excelente prognóstico. Nessa condição, progressão local ou desenvolvimento de metástases após a prostatectomia radical são extremamente raras (24). Pierorazio et al (25), utilizando o escore de Gleason, demonstraram que 95% e 97% dos pacientes com escore de Gleason 6 na biópsia e na prostatectomia radical não apresentaram recidiva da doença em 5 anos após a cirurgia, respectivamente. Em pacientes com Gleason 7 na biópsia ou na análise da peça cirúrgica, não houve recorrência da doença em 83% e 88% dos pacientes, respectivamente. Pacientes com Gleason 9 e 10 tiveram quase o dobro do risco de recorrência comparados com o Gleason 8.

Apesar de muito importante, a graduação de Gleason apresenta algumas limitações, como a subjetividade inerente à avaliação histopatológica e a

variabilidade inter e até mesmo intraobservador, que podem levar a discrepâncias no diagnóstico e no escore atribuído (21).

Além disso, o escore de Gleason obtido a partir da biópsia prostática pode não corresponder exatamente ao escore encontrado no espécime da prostatectomia radical, devido à heterogeneidade tumoral e à representatividade limitada da biópsia (26). Essa discordância entre os escores de Gleason pode influenciar a decisão terapêutica e o prognóstico do paciente.

A avaliação do escore de Gleason é um processo laborioso e demanda o desenvolvimento de métodos automáticos e eficientes para auxiliar no diagnóstico e na avaliação prognóstica do CaP. A introdução de abordagens relacionadas à inteligência artificial, como o aprendizado profundo (*deep learning*) no campo da patologia pode oferecer soluções inovadoras e precisas, reduzindo a variabilidade inter e intraobservador, melhorando a acurácia do diagnóstico (27).

1.6 Biópsia de próstata

A indicação da biópsia de próstata é baseada no nível sérico de PSA, nas características do exame clínico (toque retal) ou exame de imagem suspeito. A biópsia guiada por ultrassonografia é atualmente o método padrão. Pode ser realizada por via transretal ou transperineal (14, 19).

Segundo as recomendações da ISUP, os fragmentos biopsiados devem ser processados separadamente, sua localização deve ser registrada e deve ser atribuído um escore de Gleason a cada fragmento individual. Um escore global compreendendo todos os fragmentos também deve ser reportado. O escore de Gleason global leva em consideração todos os fragmentos positivos para adenocarcinoma, estimando a extensão total de cada grau presente (20). Para otimizar a detecção de pequenas lesões, os blocos de parafina devem ser cortados em pelo menos 3 níveis. Carcinoma intraductal, invasão perineural e extensão extraprostática também devem ser reportados, se identificados (28). O padrão cribriforme expansivo e carcinoma intraductal em biópsias são fatores prognósticos independentes de doença avançada e de pior comportamento (29).

Alguns fatores da biópsia de próstata que tem relação com o prognóstico são o número de fragmentos positivos para câncer, porcentagem de acometimento do fragmento, total em milímetros de câncer em todos os fragmentos, porcentagem de câncer nos fragmentos somados, e invasão perineural. Entretanto, o preditor mais poderoso do prognóstico na biópsia de próstata é realmente o escore de Gleason (30).

1.7 Espécime da prostatectomia radical

O exame histopatológico do espécime da prostatectomia radical descreve o estádio patológico, tipo histopatológico, grau e margens cirúrgicas. Recomenda-se que as peças sejam totalmente incluídas para permitir avaliação da localização do tumor, multifocalidade e heterogeneidade. Entretanto, para obter uma melhor custo-efetividade, geralmente realiza-se uma inclusão parcial, principalmente para próstatas com mais de 60 cm^3. O método mais aceito inclui inclusão completa da próstata posterior e um corte único médio-anterior direito e esquerdo. O exame parcial detecta 98% dos cânceres de próstata com escore de Gleason ≥ 7 com acurácia de estadiamento de 96% (31).

O escore de Gleason é também o fator prognóstico mais importante no espécime da prostatectomia radical e a recomendação é semelhante com a informação dos dois padrões prevalentes e, se presente e <5%, o padrão terciário de maior graduação (25, 32).

1.8 Tratamento do câncer de próstata

O tratamento do CaP é baseado no estágio, na agressividade e na extensão do tumor, bem como na idade, na saúde geral do paciente e nas preferências individuais. Um diagnóstico histopatológico preciso é fundamental para a seleção do tratamento adequado, e novas abordagens tecnológicas, como a inteligência artificial e, em especial, o aprendizado profundo, têm potencial para melhorar a precisão diagnóstica. As principais modalidades de tratamento incluem:

1.8.1 Prostatectomia radical

A prostatectomia radical é uma opção de tratamento cirúrgico para o CaP localizado ou localmente avançado, que consiste na remoção completa da próstata e das vesículas seminais (33). Este procedimento pode ser realizado através de uma abordagem aberta, laparoscópica ou robótica-assistida. A prostatectomia robótica-assistida tornou-se cada vez mais popular devido a suas vantagens, como menor perda de sangue, menor tempo de hospitalização e recuperação mais rápida em comparação com a abordagem aberta (34).

No entanto, a prostatectomia radical pode estar associada a complicações, como disfunção erétil e incontinência urinária (35). A preservação dos feixes neurovasculares responsáveis pela função erétil e o uso de técnicas cirúrgicas aprimoradas podem minimizar essas complicações em pacientes selecionados (36). A precisão do diagnóstico histopatológico é fundamental para a tomada de decisão do tratamento, incluindo a indicação de prostatectomia radical, a fim de evitar tratamentos desnecessários e suas possíveis complicações.

1.8.2 Radioterapia

A radioterapia utiliza radiação ionizante para destruir as células cancerígenas. Pode ser aplicada externamente (radioterapia externa) ou através da introdução de sementes radioativas (braquiterapia). A radioterapia é uma opção para pacientes com tumores localizados ou localmente avançados (37). Complicações podem incluir irritação da bexiga e do reto, disfunção erétil e fadiga.

1.8.3 Terapia hormonal

A terapia hormonal, também conhecida como terapia de deprivação androgênica (TDA), é um tratamento comumente utilizado para o CaP avançado, recidivado ou metastático (38). O objetivo principal da TDA é reduzir os níveis de andrógenos, como a testosterona, que estimulam o crescimento das células

cancerígenas da próstata. A terapia hormonal pode ser realizada por meio de orquiectomia bilateral ou através do uso de medicamentos que bloqueiam a produção ou ação dos andrógenos, como agonistas ou antagonistas do hormônio liberador de gonadotrofina e antiandrógenos (39).

A terapia hormonal tem demonstrado melhorar os sintomas, retardar a progressão da doença e prolongar a sobrevida em pacientes com CaP avançado (40). No entanto, a TDA também está associada a efeitos colaterais significativos, como fadiga, perda de massa óssea, disfunção sexual e aumento do risco de doenças cardiovasculares (41). Portanto, é essencial que o diagnóstico histopatológico seja preciso e que a indicação da terapia hormonal seja cuidadosamente considerada, a fim de minimizar o risco de tratamentos desnecessários e seus efeitos adversos.

1.8.4 Quimioterapia

A quimioterapia é geralmente reservada para casos de CaP avançado, metastático ou resistente à terapia hormonal (42). O agente quimioterápico mais frequentemente usado no tratamento do CaP é o docetaxel, geralmente administrado em combinação com prednisona (43). A quimioterapia tem demonstrado prolongar a sobrevida e melhorar a qualidade de vida em pacientes com CaP metastático resistente à castração (40). No entanto, a quimioterapia também pode causar efeitos colaterais significativos, como náusea, vômito, perda de cabelo e mielossupressão. Portanto, é crucial que o diagnóstico histopatológico seja preciso para garantir que a quimioterapia seja indicada apenas quando absolutamente necessário, minimizando o risco de tratamentos desnecessários e seus efeitos adversos.

1.8.5 Terapias focais

As terapias focais são tratamentos minimamente invasivos que têm como alvo apenas a área do tumor na próstata, preservando o tecido saudável circundante. Essas abordagens têm sido cada vez mais investigadas como alternativas às terapias tradicionais para pacientes com CaP localizado de baixo

e intermediário risco. Entre as terapias focais, destacam-se a crioterapia e o ultrassom focalizado de alta intensidade (HIFU).

A crioterapia consiste no congelamento e descongelamento do tecido prostático afetado, causando a destruição das células cancerígenas. Esta técnica utiliza agulhas especiais para introduzir gases criogênicos, como o argônio e o hélio, diretamente no tumor (44). A crioterapia tem demonstrado ser eficaz no tratamento do CaP localizado, com taxas de controle de câncer semelhantes às da prostatectomia radical e radioterapia em pacientes selecionados (45).

O HIFU utiliza ondas ultrassônicas de alta energia para aquecer e destruir seletivamente as células tumorais na próstata. O procedimento é realizado com a ajuda de um transdutor de ultrassom que é inserido no reto do paciente, permitindo a focalização precisa da energia ultrassônica no tumor (46). Estudos têm mostrado que o HIFU pode ser efetivo no tratamento do CaP localizado, com taxas de controle de câncer e efeitos colaterais aceitáveis (47).

Embora as terapias focais apresentem potencial para minimizar os efeitos colaterais associados aos tratamentos tradicionais, como a disfunção erétil e a incontinência urinária, ainda são necessárias mais pesquisas para determinar a eficácia a longo prazo e identificar os pacientes que se beneficiariam mais dessas abordagens (48).

1.8.6 Terapias emergentes e imunoterapia

Novos tratamentos, como terapias-alvo e imunoterapia, estão sendo desenvolvidos e estudados em ensaios clínicos. Essas terapias visam estimular o sistema imunológico do paciente a combater o câncer (49).

1.8.7 Vigilância ativa e espera vigilante

A vigilância ativa é uma abordagem conservadora para o manejo do CaP em estágio inicial. A vigilância ativa envolve o monitoramento regular do CaP com exames de imagem, testes de PSA e biópsias periódicas, com o objetivo de identificar progressão tumoral e então seguir para o tratamento curativo (50). A

espera vigilante, por outro lado, é uma abordagem mais passiva, na qual o tratamento é adiado até que os sintomas se tornem evidentes ou que haja sinais de progressão da doença. É uma conduta adotada para pacientes idosos e frágeis com expectativa de vida em geral menor que 10 anos (51).

Em conclusão, várias modalidades de tratamento estão disponíveis para o CaP, e a escolha do tratamento adequado depende de um diagnóstico histopatológico preciso. A introdução de técnicas modernas, como a inteligência artificial e o aprendizado profundo, pode contribuir para melhorar a precisão diagnóstica, auxiliando na escolha do tratamento e, consequentemente, no prognóstico dos pacientes. Ao evitar tratamentos desnecessários ou inapropriados, é possível reduzir as complicações inerentes aos tratamentos e melhorar a qualidade de vida dos pacientes com CaP.

1.9 Introdução à inteligência artificial e aprendizado de máquina na medicina

A inteligência artificial (IA) é um ramo da ciência da computação que busca desenvolver algoritmos e sistemas capazes de realizar tarefas que, até então, exigiam a inteligência humana, como o reconhecimento de padrões, tomada de decisões e resolução de problemas complexos (52). Desde suas origens na década de 1950, a IA passou por várias fases de evolução e otimização, culminando em avanços significativos nas últimas décadas, impulsionados pelo aumento do poder computacional, a disponibilidade de grandes volumes de dados e o desenvolvimento de algoritmos sofisticados.

O aprendizado de máquina (*machine learning*), um subcampo da IA, é uma abordagem que busca desenvolver algoritmos capazes de aprender e melhorar seu desempenho por meio da experiência, ou seja, a partir de dados (53). Os algoritmos de aprendizado de máquina podem ser classificados em três categorias principais: aprendizado supervisionado, não supervisionado e por reforço (54).

No aprendizado supervisionado, os algoritmos são treinados com base em conjuntos de dados rotulados, ou seja, dados em que os resultados desejados são conhecidos.

No aprendizado não supervisionado, os algoritmos são treinados para identificar padrões e agrupar dados não rotulados, enquanto no aprendizado por reforço, os algoritmos aprendem com base em recompensas e punições dadas durante a realização de tarefas.

Nos últimos anos, a medicina tem se beneficiado dos avanços no campo da IA e do aprendizado de máquina, resultando em aplicações inovadoras que abrangem várias áreas, como diagnóstico por imagem, genômica, medicina personalizada e epidemiologia, para citar algumas (55). Estudos recentes têm demonstrado o potencial do aprendizado de máquina em melhorar a precisão diagnóstica e prognóstica, auxiliando profissionais de saúde na tomada de decisões mais informadas e personalizadas (56, 57).

Em particular, o aprendizado profundo, uma subcategoria do aprendizado de máquina baseada em redes neurais artificiais profundas, tem revolucionado o campo da análise de imagens médicas (58). Redes neurais convolucionais (CNNs) têm sido amplamente aplicadas na análise de imagens médicas, como tomografia computadorizada, ressonância magnética e imagens histopatológicas, com resultados promissores em termos de desempenho e precisão (27).

Nos últimos cinco anos, a literatura tem apresentado uma série de estudos que investigam a aplicabilidade do aprendizado de máquina e do aprendizado profundo em diversos aspectos da medicina. Por exemplo, um estudo de Esteva et al. (56) demonstrou a eficácia de um algoritmo de aprendizado profundo na classificação de lesões cutâneas em imagens clínicas, atingindo uma precisão comparável à de dermatologistas especializados. Outro estudo, conduzido por Gulshan et al. (57), mostrou que um algoritmo de aprendizado profundo poderia identificar retinopatia diabética e edema macular diabético em imagens de retina com precisão comparável ou superior à de oftalmologistas especializados. Esses estudos ilustram o potencial da IA e do aprendizado de máquina para transformar o campo da medicina.

A aplicação da IA e do aprendizado de máquina também se estende a outras áreas da medicina, como a análise de sequências genômicas e a identificação de biomarcadores para diagnóstico e tratamento personalizado (59, 60). Além disso, algoritmos de aprendizado de máquina têm sido empregados

na análise de grandes conjuntos de dados epidemiológicos, auxiliando na identificação de padrões e tendências relacionadas a doenças e condições de saúde (61).

Na área da medicina personalizada, a IA e o aprendizado de máquina têm sido utilizados para identificar padrões e associações em dados clínicos e moleculares, permitindo a identificação de subgrupos de pacientes e a previsão de respostas a tratamentos específicos (62). Isso pode levar a abordagens de tratamento mais direcionadas e personalizadas, melhorando a qualidade de vida e os resultados clínicos para os pacientes.

A crescente integração da IA e do aprendizado de máquina na medicina também tem levantado questões éticas e legais importantes. Por exemplo, a proteção e a privacidade dos dados dos pacientes são de extrema importância, especialmente quando se trata de informações médicas sensíveis. Além disso, a responsabilidade e a tomada de decisões em relação ao tratamento e ao diagnóstico também precisam ser cuidadosamente consideradas à medida que a IA e o aprendizado de máquina assumem um papel cada vez maior na prática médica (63).

Em suma, a inteligência artificial e o aprendizado de máquina, especialmente o aprendizado profundo, estão trazendo mudanças significativas para a medicina e têm um potencial considerável para melhorar o diagnóstico e o tratamento de diversas condições. A investigação contínua da aplicabilidade e eficácia dessas abordagens nas várias áreas da medicina é crucial para a adoção bem-sucedida dessas tecnologias na prática clínica e para garantir que os benefícios sejam maximizados, ao mesmo tempo em que são abordados os desafios éticos e legais.

1.10 Aprendizado profundo aplicado ao diagnóstico histopatológico

O aprendizado profundo, uma subcategoria do aprendizado de máquina, baseia-se em redes neurais artificiais (RNAs) com múltiplas camadas ocultas, permitindo a extração de características de alto nível e a aprendizagem de representações hierárquicas dos dados (58). Essa abordagem tem demonstrado vantagens significativas em relação a outras técnicas de aprendizado de

máquina, particularmente em tarefas envolvendo grandes volumes de dados e alta dimensionalidade, como é o caso da análise de imagens médicas e histopatológicas (64).

Redes neurais convolucionais (CNNs) são uma classe específica de RNAs profundas que têm sido amplamente aplicadas na análise de imagens, devido à sua capacidade de aprender automaticamente características visuais relevantes a partir dos dados brutos (65). CNNs são compostas por camadas convolucionais, que realizam operações de convolução localizadas para extrair características espaciais das imagens, seguidas por camadas de agrupamento (*pooling*), que reduzem a dimensionalidade dos dados e aumentam a invariância a variações de escala e posição (58).

A aplicação de CNNs na análise de imagens médicas e histopatológicas tem resultado em avanços significativos na detecção, diagnóstico e prognóstico de diversas condições, incluindo o câncer (27). Em particular, a análise de imagens histopatológicas apresenta desafios específicos, devido à complexidade e variabilidade das estruturas celulares e teciduais, bem como às variações na coloração e preparação das lâminas (66). Nesse contexto, o aprendizado profundo e, especificamente, as CNNs têm mostrado um potencial considerável para melhorar a precisão e eficiência do diagnóstico histopatológico.

Estudos recentes têm demonstrado o sucesso das CNNs na análise de imagens histopatológicas para a detecção e classificação de lesões cancerígenas, bem como na avaliação de biomarcadores e características prognósticas (67, 68).

A aplicação de CNNs no diagnóstico histopatológico também tem o potencial de reduzir a variabilidade interobservador e melhorar a padronização das avaliações, dado que os algoritmos de aprendizado profundo são treinados para aprender características consistentes e discriminativas a partir dos dados (66). A análise automatizada de lâminas histopatológicas por meio do aprendizado profundo pode aumentar a eficiência do processo diagnóstico, permitindo uma triagem mais rápida e sistemática das amostras e auxiliando os patologistas na identificação de casos mais complexos e desafiadores (68).

O aprendizado profundo tem sido aplicado na quantificação de características histopatológicas que são de difícil avaliação, como a densidade celular, a intensidade de infiltrado inflamatório, número de mitoses e a arquitetura tecidual (69, 70). Essas informações quantitativas podem fornecer insights valiosos para a tomada de decisões clínicas e a estratificação de risco dos pacientes.

Apesar dos avanços promissores, a aplicação do aprendizado profundo no diagnóstico histopatológico ainda enfrenta desafios, como a necessidade de anotação de grandes conjuntos de dados para o treinamento de modelos precisos (62). Abordagens recentes têm buscado superar essas limitações, por exemplo, através do uso de aprendizado semi-supervisionado, transferência de aprendizado e técnicas de interpretação, como a decomposição em camadas (LIME) e a propagação de relevância (LRP) (71, 72).

Em suma, o aprendizado profundo e, especificamente, as CNNs têm mostrado um potencial significativo para melhorar o diagnóstico histopatológico e a análise de imagens médicas. A investigação contínua das abordagens baseadas em aprendizado profundo e a comparação com a performance dos patologistas especializados são etapas cruciais para a adoção bem-sucedida dessas tecnologias na prática clínica. Além disso, a exploração de métodos avançados de aprendizado de máquina e o desenvolvimento de soluções para superar os desafios atuais são fundamentais para a continuidade dos avanços no campo e para a realização do potencial transformador da inteligência artificial na medicina.

1.11 Aplicações da inteligência artificial no diagnóstico e prognóstico do câncer de próstata

Os métodos de IA tem mostrado um potencial promissor na melhoria dos processos de diagnóstico e prognóstico em diversas áreas da medicina. Um exemplo de grande sucesso tem sido a avaliação automatizada de tumores de mama (73). E essa aplicabilidade dos algoritmos de IA e aprendizado profundo também inclui o CaP (74). As aplicações da IA no diagnóstico do CaP incluem,

a análise de imagens radiológicas, como ressonância magnética multiparamétrica (RNMmp) e histopatológicas (27, 67).

A análise histopatológica do tecido prostático, obtida por meio de biópsia, é fundamental para o diagnóstico do CaP e para a determinação de seu potencial de agressividade. Recentemente, algoritmos de deep learning, como as CNNs, têm sido aplicados na análise de imagens de lâminas digitalizadas para identificar o câncer e atribuir escores de Gleason (27). Estudos como o de Bulten et al. (67) mostraram que a IA pode alcançar uma precisão diagnóstica semelhante ou superior à dos patologistas especializados e contribuir para uma maior concordância interobservador.

Em um dos estudos pioneiros sobre o uso da IA na detecção do CaP, Litjens et al. (27) analisou 225 lâminas de pacientes submetidos a biópsia de próstata: 100 no grupo de treinamento, 50 no grupo de validação e 75 no grupo teste. Todas as lâminas contendo CaP puderam ser identificadas automaticamente no grupo teste usando o modelo de aprendizado profundo treinado.

Na análise de imagens de RNMmp, a IA tem auxiliado na detecção e localização de lesões cancerígenas e na discriminação entre tumores clinicamente significativos e insignificantes

A IA também tem sido aplicada na estratificação de risco e prognóstico do CaP, auxiliando na seleção do tratamento mais adequado para cada paciente. Por exemplo, o uso de algoritmos de aprendizado de máquina tem sido investigado para prever a probabilidade de recorrência após prostatectomia radical. Esses modelos podem integrar dados clínicos, patológicos e moleculares para fornecer estimativas personalizadas de risco, melhorando a tomada de decisão clínica (75).

1.12 Implicações da inteligência artificial na prática clínica e no mundo real

A IA e o aprendizado de máquina têm o potencial de revolucionar a prática clínica em diversas áreas da medicina, incluindo o diagnóstico e prognóstico do

CaP. No entanto, a integração da IA na prática clínica também apresenta desafios e implicações que precisam ser considerados.

1.12.1 Complementando o trabalho de médicos e patologistas na tomada de decisões clínicas

Os algoritmos de aprendizado profundo podem servir como uma ferramenta complementar para médicos e patologistas, fornecendo informações adicionais e objetivas que podem melhorar a precisão e eficiência do diagnóstico e prognóstico do CaP (76). A combinação de conhecimento humano e insights gerados por algoritmos de IA pode levar a uma tomada de decisão clínica mais precisa e personalizada, beneficiando os pacientes através de tratamentos mais precisos e eficazes (55).

1.12.2 Desafios na implementação da inteligência artificial na prática clínica

A implementação da IA na prática clínica apresenta vários desafios, incluindo questões éticas, legais e regulatórias. Algumas das principais preocupações e desafios incluem:

a) Validade e confiabilidade dos algoritmos: Os algoritmos de IA devem ser validados em populações diversificadas e em diferentes contextos clínicos para garantir sua aplicabilidade e generalização no mundo real (77). A confiabilidade dos algoritmos também é crucial para a adoção bem-sucedida da IA na prática clínica, uma vez que os profissionais de saúde precisam confiar nas informações fornecidas pelos sistemas de IA (78).

b) Interpretabilidade e transparência: Muitos algoritmos de aprendizado profundo são considerados "caixas-pretas", ou seja, suas decisões podem ser difíceis de entender ou explicar (79). A interpretabilidade e a transparência são importantes para garantir que os profissionais de saúde possam compreender e confiar nos resultados gerados pela IA, bem como identificar possíveis erros ou vieses (80).

c) Questões éticas e legais: A IA levanta questões éticas e legais, como a responsabilidade em casos de erros de diagnóstico ou tratamento resultantes

do uso de algoritmos de IA. Além disso, a privacidade e a proteção dos dados dos pacientes são preocupações significativas, especialmente considerando que os algoritmos de IA geralmente exigem grandes quantidades de dados para treinamento e validação (81).

d) Regulação e aprovação: A regulamentação e aprovação de dispositivos médicos baseados em IA representam um desafio significativo, pois as agências reguladoras precisam garantir a segurança e eficácia dessas tecnologias (82). Os processos de aprovação devem ser adaptados para lidar com as especificidades dos algoritmos de IA garantindo que sejam rigorosos o suficiente para proteger os pacientes e flexíveis o suficiente para permitir a inovação (83).

e) Aceitação e adoção pelos profissionais de saúde: A aceitação e adoção da IA pelos profissionais de saúde é fundamental para o sucesso da integração da IA na prática clínica (84). A formação e a educação dos profissionais de saúde em IA e aprendizado de máquina é essencial para garantir que eles possam usar adequadamente essa tecnologia e compreender suas limitações (85).

1.13 Justificativa

A IA e os modelos de aprendizagem profunda são uma realidade e o desenvolvimento de um modelo próprio, com dados confiáveis para utilização na prática clínica é imperativo.

Considerando a alta prevalência do CaP e o valor determinante de sua caracterização anátomo-patológica, buscamos desenvolver um modelo que possa ser utilizado na rotina, auxiliando o patologista e aumentando a precisão diagnóstica.

2. OBJETIVOS

2.1 Objetivo geral

Desenvolvimento de um modelo de inteligência artificial (IA) e aprendizado profundo (*deep learning*) no diagnóstico e graduação histológica do câncer de próstata.

2.2 Objetivos específicos

Treinar um modelo de IA para o diagnóstico do CaP.

Treinar um modelo de IA para a determinação dos padrões histológicos de Gleason.

Avaliar a acurácia do modelo no diagnóstico do CaP e identificação do padrão de Gleason em relação àquele determinado pelo uropatologista.

Identificar as possíveis dificuldades do sistema no diagnóstico e graduação do CaP.

Testar e comparar diferentes métodos de aprendizado profundo para avaliar aquele com melhor acurácia na identificação e graduação do CaP.

3. MATERIAL E MÉTODOS

3.1 Tipo do estudo

Estudo transversal exploratório.

3.2 Local e época

O presente estudo foi desenvolvido e conduzido na cidade de São Paulo, SP, Brasil. O período de realização deste projeto compreendeu os anos de 2018 a 2023.

3.3 Pacientes

O objeto deste estudo consiste em lâminas histológicas provenientes de espécimes de prostatectomia radical, obtidas de pacientes submetidos à cirurgia para tratamento do CaP. As lâminas analisadas foram selecionadas a partir dos arquivos do Laboratório Genoa/LPCM, que possui uma ampla coleção de amostras histopatológicas, proporcionando uma base sólida para a realização deste estudo.

A seleção dos espécimes foi realizada por uma uropatologista expert no assunto (Prof. Dra. Katia Ramos Moreira Leite). Foram criteriosamente selecionados espécimes de prostatectomia radical representativos dos três padrões de Gleason (3, 4 e 5), buscando garantir uma representatividade equitativa de cada padrão na amostra analisada.

3.4 Critérios de inclusão e exclusão

Critérios de Inclusão:

- Lâminas histológicas obtidas de espécimes de prostatectomia radical, realizadas em pacientes submetidos à cirurgia para tratamento do CaP;

- Representação de tecido benigno e contendo câncer com os padrões 3, 4 e 5 de Gleason.

Critérios de Exclusão:

- Pacientes submetidos a radioterapia ou terapia anti-androgênica prévia.

3.5 Processamento do espécime da prostatectomia radical

Os espécimes foram fixados por imersão em formalina tamponada a 10% por período variável entre 24 e 72 horas. As margens cirúrgicas foram tingidas com tinta Nanquim no momento inicial do exame macroscópico. O ápice e a margem do colo vesical foram removidos e seccionados em cortes sagitais. O restante da peça é seccionado de forma transversal, em cortes de 3-4 mm, perpendiculares ao eixo longitudinal da uretra. Os cortes resultantes são processados em banhos de álcool, xilol e embebidos em parafina (86).

Cortes de 3,0 μm foram corados pela hematoxilina e eosina (H&E) e analisados ao microscópio óptico. Foram registrados: tipo histológico, graduação de acordo com escore de Gleason e ISUP, volume tumoral, envolvimento do tecido extraprostático, envolvimento das vesículas seminais, situação das margens cirúrgicas, multifocalidade, localização do tumor e invasão angiolinfática. Em relação a linfadenectomia pélvica, dissecou-se os linfonodos reconhecidos macroscopicamente, sendo incluído também todo o restante do tecido adiposo. Registrou-se a presença de metástases, o tamanho da maior lesão e a infiltração do tecido adiposo perinodal.

3.6 Processamento e aquisição das imagens

Lâminas representativas de tecido benigno e câncer com os padrões 3, 4 e 5 de Gleason foram selecionadas para a digitalização. Todas as lâminas foram digitalizadas em scanner de lâminas de alta qualidade (Panoramic Flash II 250 scanner – 3DHISTECH Ltd., Budapest, Hungary) (Figura 2). A magnificação das imagens utilizadas no treinamento da rede foi de 20 vezes.

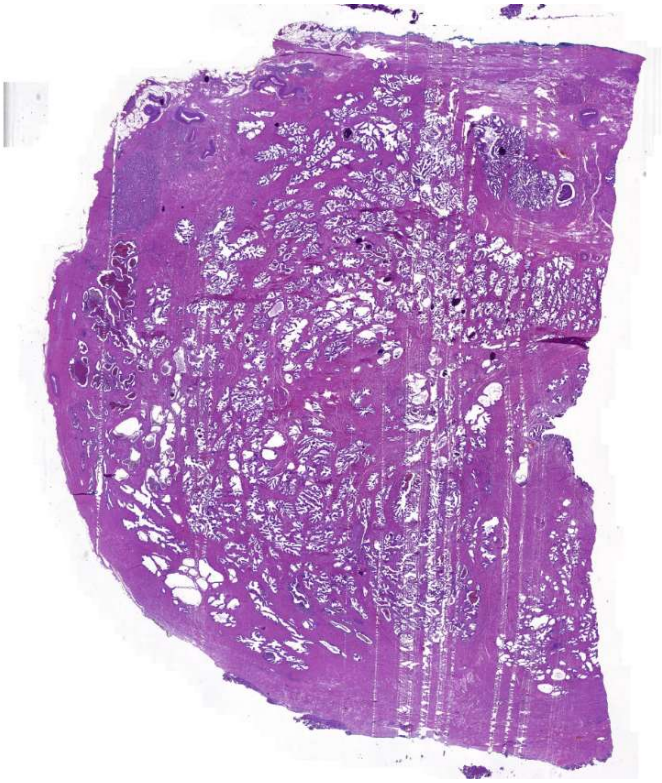

Figura 2. Imagem completa da lâmina de prostatectomia radical escaneada

Após a digitalização, cada lâmina foi subdividida em pequenas imagens quadradas de dimensões iguais, medindo 2000 x 2000 pixels (Figura 3). A segmentação das lâminas completas em imagens menores facilitou o processamento e treinamento da rede, tornando o processo viável, em comparação à utilização das imagens das lâminas de prostatectomia radical inteiras. A divisão das lâminas em imagens menores foi realizada por meio de um código desenvolvido na linguagem de programação Python 3 (87).

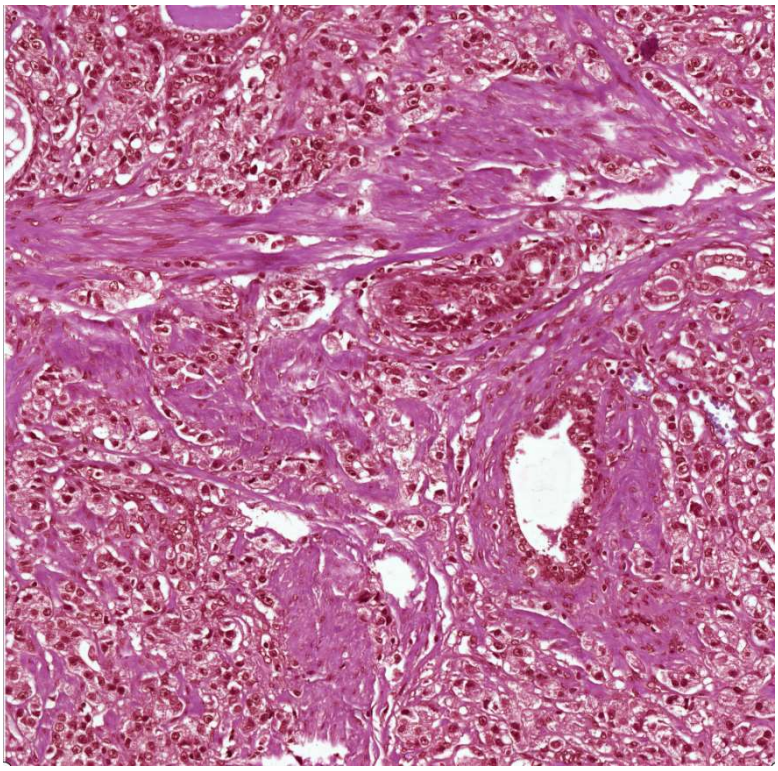

Figura 3. Imagem resultante da segmentação da lâmina completa em imagens menores de 2000 x 2000 pixels

A análise das imagens digitalizadas foi conduzida pela mesma especialista em uropatologia responsável pela seleção das lâminas (Prof. Dra. K.R.M.L.). Cada uma das pequenas imagens geradas foi examinada minuciosamente, e as regiões representativas foram delineadas, incluindo:

- Glândulas normais;
- Estroma prostático;
- Glândulas cancerígenas com padrões de Gleason 3, 4 e 5.

Para essa demarcação foi utilizada a ferramenta de código aberto Coco Annotator (88).

3.7 Tipos de classificação

Neste estudo, foram empregados dois métodos distintos para analisar e classificar as imagens histológicas: classificação categórica e segmentação de instâncias. Ambos os métodos foram escolhidos devido às suas potenciais implicações na precisão e eficácia do diagnóstico do CaP.

3.7.1 Método de classificação categórica

A classificação categórica consistiu em atribuir um único rótulo a cada imagem analisada (Figura 4). As imagens eram classificadas como tecido benigno, caso não apresentassem sinais de malignidade, ou como câncer com o respectivo padrão de Gleason mais significativo identificado na imagem. Essa análise foi realizada individualmente em cada uma das pequenas imagens obtidas a partir da lâmina inteira e, posteriormente, os resultados foram combinados para determinar a porcentagem de cada padrão presente na lâmina completa. Após a marcação, as imagens foram agrupadas em diretórios separados, conforme suas categorias.

A escolha das redes neurais convolucionais (*convolutional neural networks - CNN*) (89) para o treinamento e a classificação das imagens no modelo categórico deve-se à sua eficiência em processar imagens e identificar padrões complexos, sendo amplamente empregada no reconhecimento computacional de padrões em imagens.

.

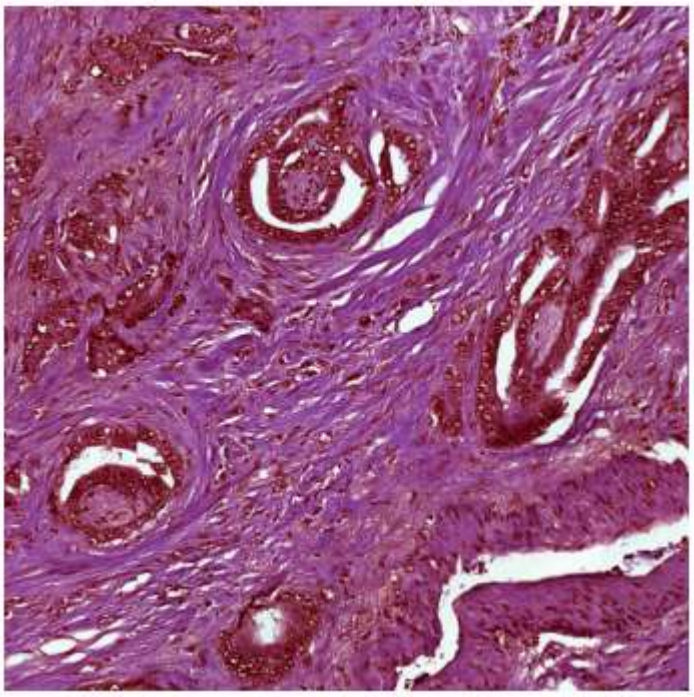

Figura 4. Método de classificação categórica - Imagem rotulada como câncer padrão 4 de Gleason

3.7.2 Método de segmentação de instâncias

A segunda abordagem para a análise das imagens envolveu o delineamento específico de cada área de interesse em cada imagem. Neste método, áreas contendo um ou mais padrões foram delineadas e anotadas nas imagens, em vez de atribuir um único rótulo a toda a imagem. Esta abordagem permite uma análise mais detalhada das áreas de interesse e possibilita a identificação de características específicas e interações entre diferentes padrões de Gleason (Figura 5).

Para o treinamento e a classificação das imagens no modelo de segmentação de instâncias, utilizou-se a técnica de aprendizado profundo conhecida como Mask R-CNN (*region-based convolutional neural network*) (90). A Mask R-CNN foi selecionada devido à sua capacidade de realizar segmentação de instâncias, permitindo a identificação e separação de objetos individuais em uma imagem e proporcionando uma análise mais precisa das regiões anotadas.

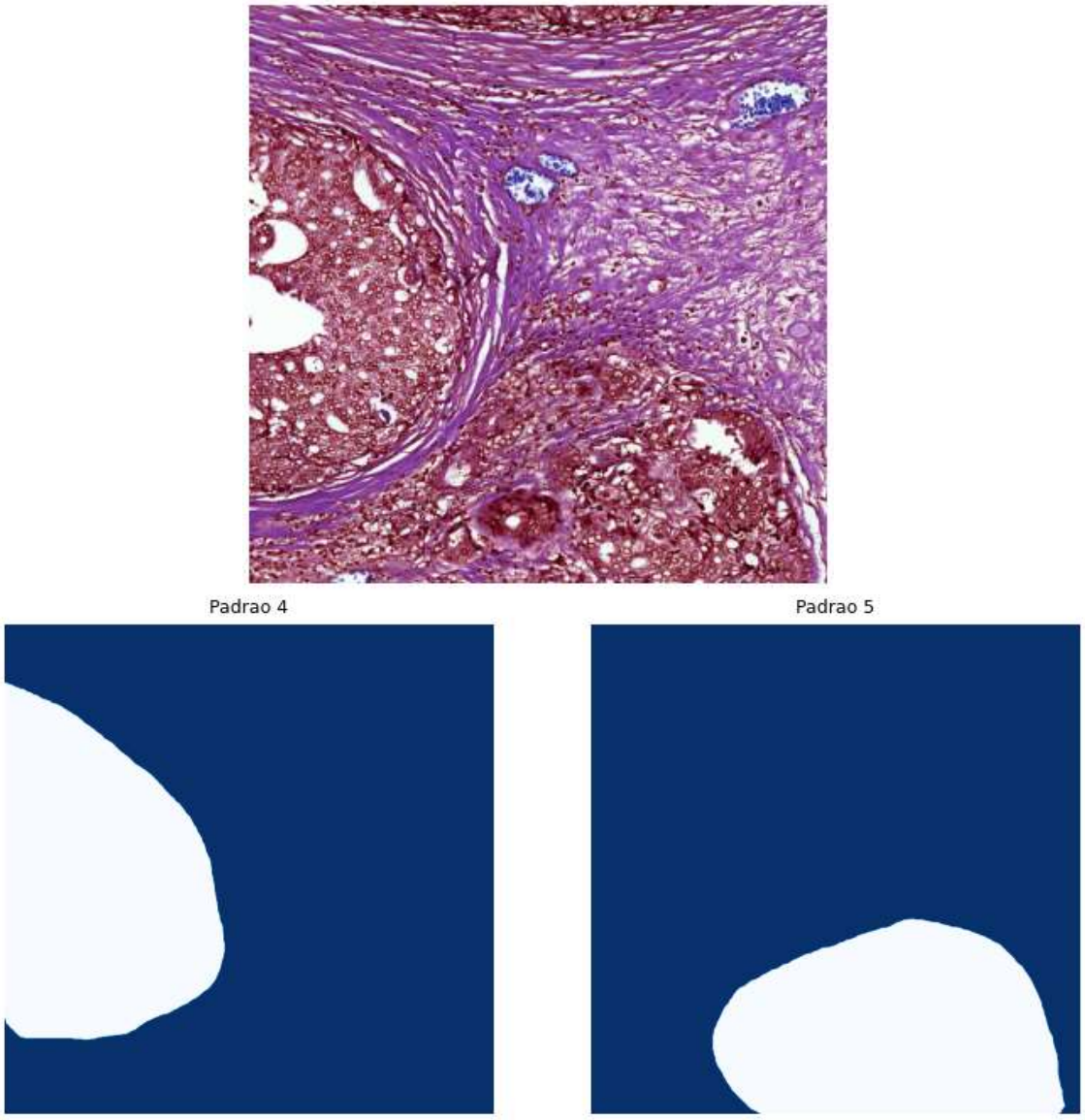

Figura 5. Método de segmentação de instâncias - Acima a imagem analisada. Abaixo, as duas marcações que foram feitas na mesma imagem representando uma área com padrão 4 à esquerda e uma área com padrão 5 à direita.

3.8 Redes neurais convolucionais

As redes neurais convolucionais (CNNs) são uma classe de redes neurais profundas que têm demonstrado um desempenho excepcional em tarefas de visão computacional, como classificação de imagens e detecção de objetos (89). Neste estudo, as CNNs foram empregadas para realizar a classificação

categórica das imagens histológicas, aproveitando sua capacidade de aprender automaticamente características relevantes das imagens e realizar análises precisas e eficientes.

A característica distintiva das CNNs é a camada convolucional, que permite a identificação de padrões locais em imagens de entrada. Diferentemente das camadas totalmente conectadas, onde cada neurônio está conectado a todos os neurônios na camada anterior, os neurônios nas camadas convolucionais estão conectados apenas a um subconjunto de neurônios na camada anterior, conhecido como campo receptivo (65). Essa arquitetura localizada permite que a CNN aprenda a detectar características específicas em diferentes partes da imagem, resultando em uma representação mais rica e discriminativa das imagens de entrada (Figura 6).

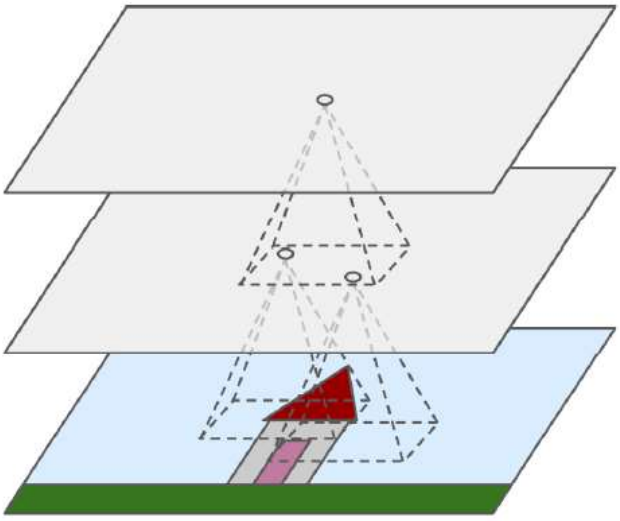

Figura 6. Camadas da CNN com os campos receptivos locais retangulares. Abaixo, a imagem de entrada. No meio, a camada convolucional 1 e acima a camada convolucional 2. Adaptado de Géron (91).

Um neurônio localizado na linha x e coluna y de uma dada camada é conectado aos neurônios da camada prévia localizados nas linhas x + h - 1 e colunas y + w - 1, onde h e w são a altura e comprimento do campo receptivo (Figura 7). Para que uma camada tenha a mesma altura e comprimento da

camada prévia, é comum se adicionar zeros ao redor dessa camada de entrada. Isso é chamado de *zero padding*.

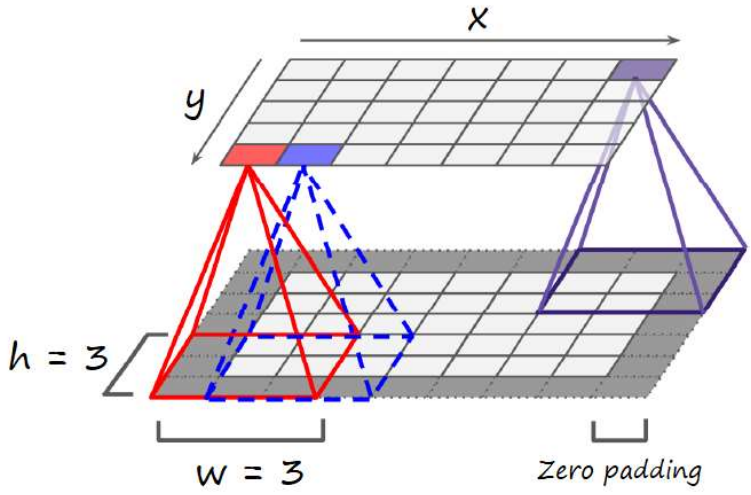

Figura 7. Conexão entre as camadas e o *zero padding.* Adaptado de Géron (91).

Também é possível conectar uma grande camada de entrada a uma camada muito menor ao espaçar os campos receptivos (Figura 8). A mudança de um campo receptivo para o outro é chamada de *stride*. Resumidamente, é a quantidade de pixels que o campo receptivo "salta" em cada direção na análise da camada de entrada (sh = stride vertical; sw = stride hotizontal).

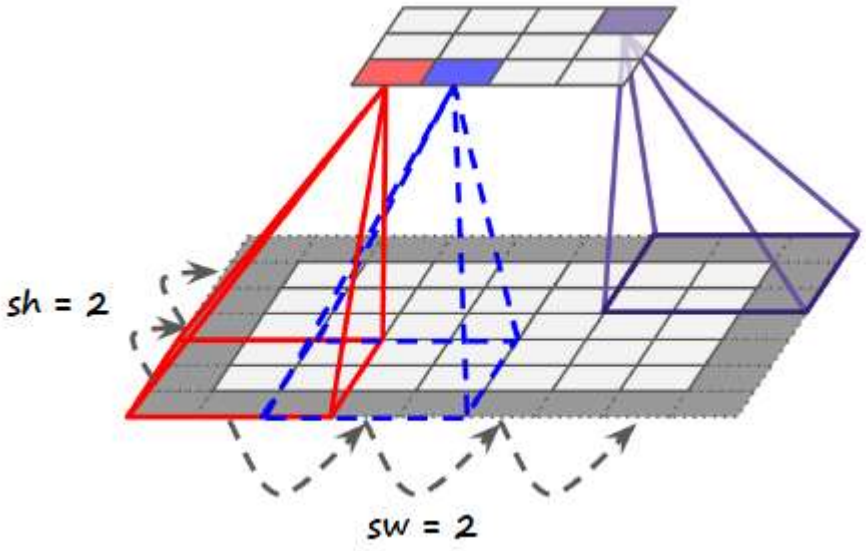

Figura 8. Reduzindo a dimensionalidade usando um *stride* de 2. Adaptado de Géron (91).

A cada uma das conexões da rede neural é atribuído um valor chamado de peso neuronal. Esses pesos neuronais podem ser representados como uma pequena imagem do tamanho do campo receptivo. Um grupo de dois pesos é chamado de filtro (ou kernel convolucional). Esses filtros percorrem toda a imagem de entrada gerando um mapa de características, que de forma geral, destaca as áreas da imagem onde o filtro foi mais ativado, com o objetivo de extrair da imagem suas características mais importantes. Esses filtros não são definidos manualmente: durante o treinamento, a CNN vai automaticamente aprendendo quais são os filtros mais úteis para sua tarefa e eles vão sendo combinados em padrões complexos em todas as camadas da rede. Cada camada convolucional tem múltiplos filtros e um mapa de características é gerado para cada filtro.

Além disso, as imagens de entrada são compostas de múltiplas subcamadas: uma por canal de cor. Há tipicamente três: vermelho, verde e azul (RGB – *red, green, blue*) (Figura 9).

Especificamente, um neurônio localizado na linha x, coluna y de um mapa de características k em uma dada camada convolucional l é conectado às saídas dos neurônios da camada prévia l - 1, localizados nas linhas x * sh até x * sh + h - 1 e colunas y * sw até y * sw + w - 1, através de todos os mapas de características (na camada l - 1). Todos os neurônios localizados na mesma linha x e coluna y mas em diferentes mapas de características são conectados às saídas dos mesmos neurônios na camada prévia.

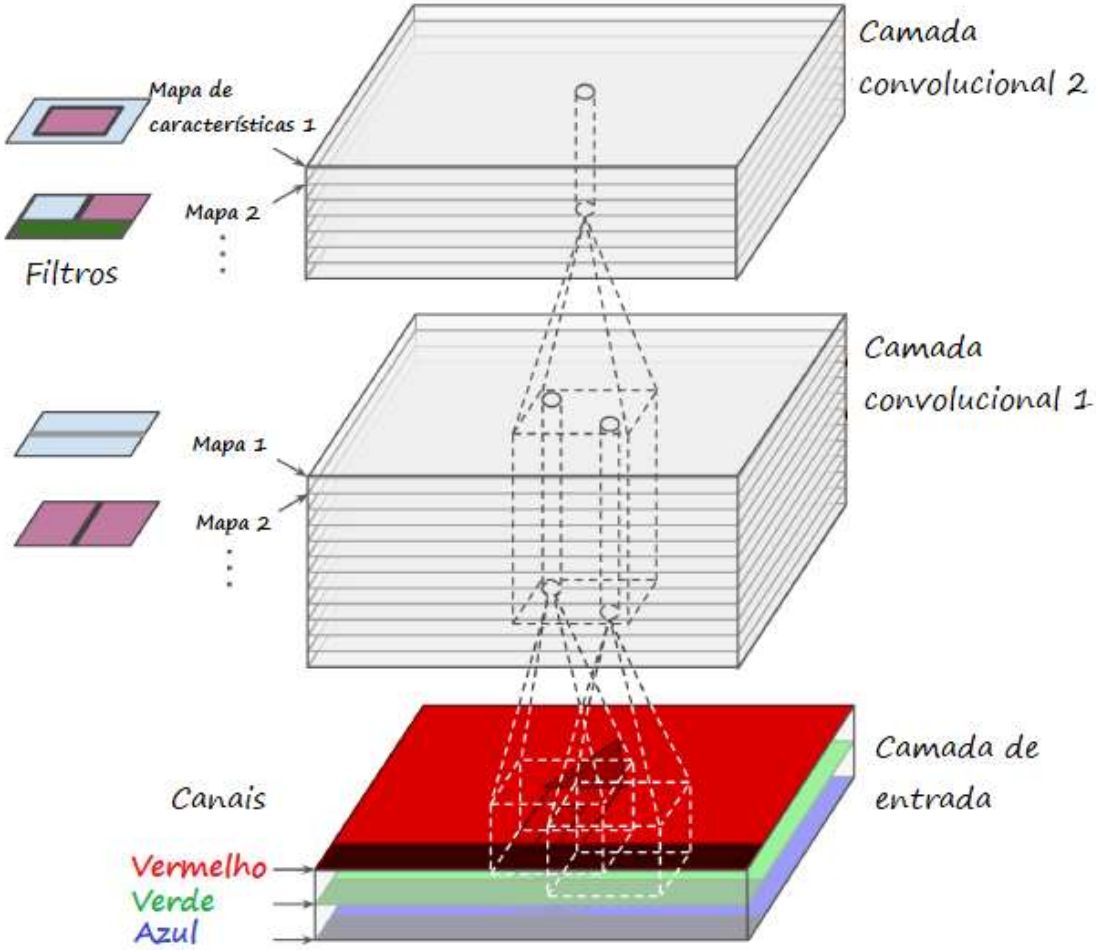

Figura 9. Camadas convolucionais com múltiplos mapas de características e imagens com três canais de cores. Adaptado de Géron (91).

Além das camadas convolucionais, as CNNs também incluem camadas de pooling, que têm como objetivo reduzir a dimensionalidade das representações intermediárias e, consequentemente, diminuir a carga computacional, o uso da memória e o número de parâmetros (92). Isso também ajuda a prevenir o chamado sobreajuste (*overfitting*), que é um problema comum em redes neurais profundas. As camadas de pooling realizam uma operação de agregação, como o máximo ou a média, em uma janela deslizante aplicada aos mapas de características.

Como nas camadas convolucionais, cada neurônio na camada de pooling é conectada a um número limitado de neurônios da camada prévia, localizado dentro de um pequeno campo receptivo retangular. Assim como nas camadas

convolucionais, também se deve definir o tamanho dos campos receptivos, do stride e o tipo de padding. Entretanto, o neurônio do pooling não tem pesos: tudo que eles fazem é agregar as entradas usando uma função de agregação como o máximo ou a média. A figura 10 mostra a camada de pooling máximo, o tipo mais comum de camada de pooling.

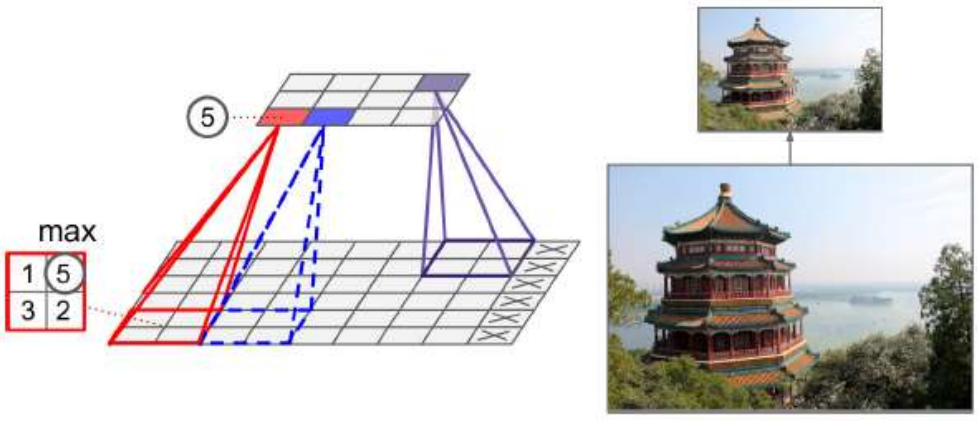

Figura 10. Camada de *pooling* máximo (filtro de *pooling* 2x2, *stride* de 2, sem *padding*). Adaptado de Géron (91).

Um outro tipo de camada de pooling é a camada de pooling médio global. Ela funciona de uma forma totalmente diferente: ela computa a média de cada mapa de características inteiro. Ela gera um único número por mapa de características. Embora isso seja extremamente destrutivo (a maioria das informações no mapa de características é perdido), isso pode ser útil na camada final de saída.

As arquiteturas típicas das CNNs consistem em várias camadas convolucionais intercaladas com camadas de pooling e camadas de ativação não linear, como a função de ativação linear retificada (ReLU) (93). No topo dessa pilha, uma rede neural feedforward é adicionada, composta por algumas camadas totalmente conectadas e uma camada de saída que realiza a predição (Figura 11).

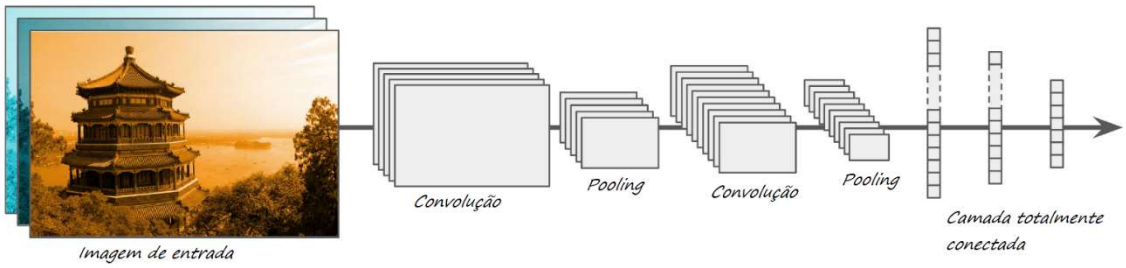

Figura 11. Arquitetura da CNN típica. Adaptado de Géron (91).

3.9 Mask R-CNN

A rede conhecida como Mask R-CNN (*region-based convolutional neural network*) (90) é uma estrutura de aprendizado profundo que tem como objetivo solucionar o problema de detecção e segmentação de instâncias de objetos em imagens (Figura 12).

Trata-se de uma abordagem que detecta de forma eficiente objetos em uma imagem enquanto simultaneamente gera uma máscara de segmentação de alta qualidade para cada exemplo de objeto. A Mask R-CNN é simples de treinar e generalizar (exemplo: pode tanto identificar um humano em uma fotografia, como também diferenciar posturas diferentes desse humano naquela foto).

Segmentação de imagens é um processo desafiador porque requer tanto a correta detecção de todos os objetos em uma imagem como também a segmentação precisa de cada objeto dentro da imagem. Portanto, esse processo combina elementos da visão computacional clássica de detecção de objetos, onde o objetivo é classificar objetos individuais e localizá-los usando uma caixa delimitadora (*"bounding box"*), e elementos da segmentação semântica, onde o objetivo é classificar cada pixel em um grupo fixo de categorias sem diferenciar cada objeto na imagem. Dado isso, se espera que um método complexo seja necessário para atingir tais objetivos, entretanto, a Mask R-CNN se demonstrou um método simples, flexível e rápido para execução da segmentação dos objetos.

Essa rede é uma extensão do modelo Faster R-CNN (94), que é uma abordagem eficiente para detecção de objetos, combinando propostas de região

e extração de características em uma única rede. A Mask R-CNN expande a funcionalidade da Faster R-CNN, adicionando a capacidade de gerar máscaras de segmentação para cada objeto identificado.

A Mask R-CNN aborda o desafio da segmentação de instâncias de objetos em duas etapas principais. A primeira etapa é a detecção de objetos, que envolve a localização e classificação de objetos na imagem. Isso é alcançado utilizando um backbone convolucional profundo, geralmente uma arquitetura como ResNet (95) ou VGG (96), para extrair características de alto nível da imagem. Em seguida, um módulo conhecido como Region Proposal Network (RPN) é usado para propor regiões candidatas de interesse (RoIs) que possam conter objetos.

Na segunda etapa, a segmentação de instâncias, cada RoI é processada por uma série de camadas para gerar máscaras de segmentação e refinamento das caixas delimitadoras. A Mask R-CNN introduz uma "head" de segmentação paralela à "head" de classificação e regressão de caixas delimitadoras existente na Faster R-CNN. Essa head de segmentação consiste em uma pequena Fully Convolutional Network (FCN) (97) aplicada a cada RoI, que produz uma máscara de segmentação pixel-a-pixel. Para lidar com as diferenças de tamanho e aspecto das RoIs, a Mask R-CNN utiliza uma camada denominada RoIAlign que interpola as características de entrada para um tamanho fixo, preservando a informação espacial e evitando distorções (Figura 13).

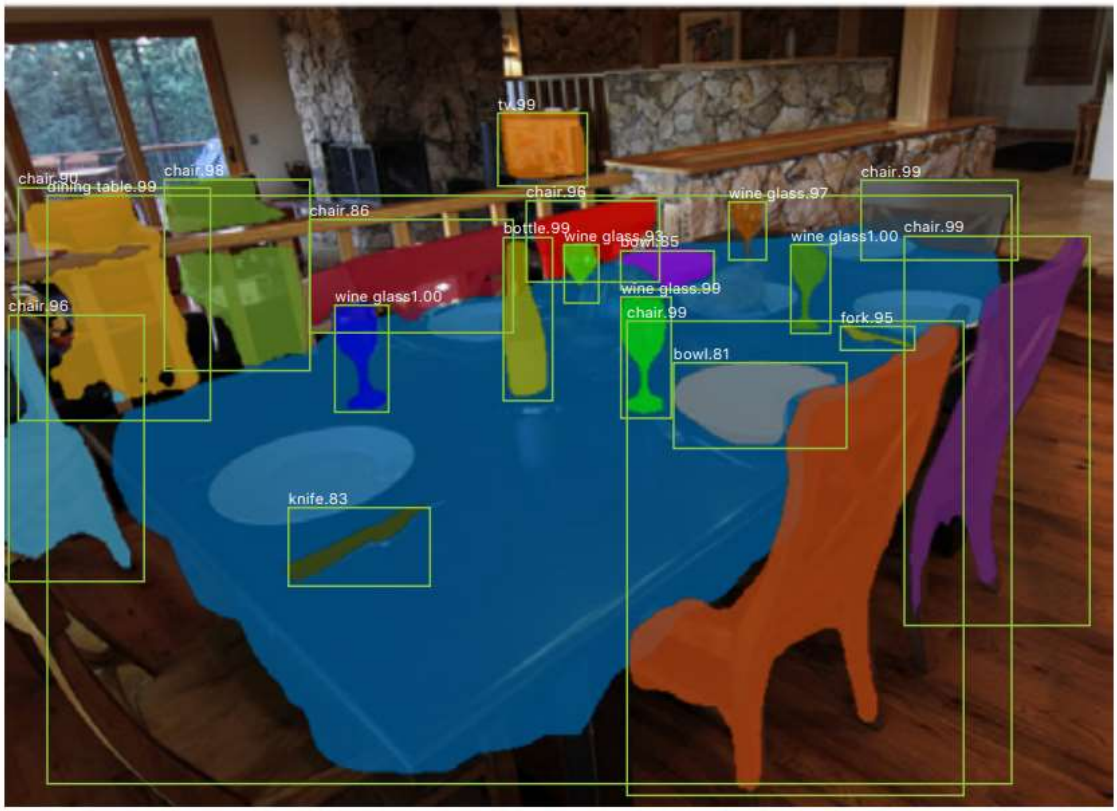

Figura 12. Exemplo de detecção de objetos em uma imagem com utilização da Mask R-CNN. Adaptado de Abdulla (98).

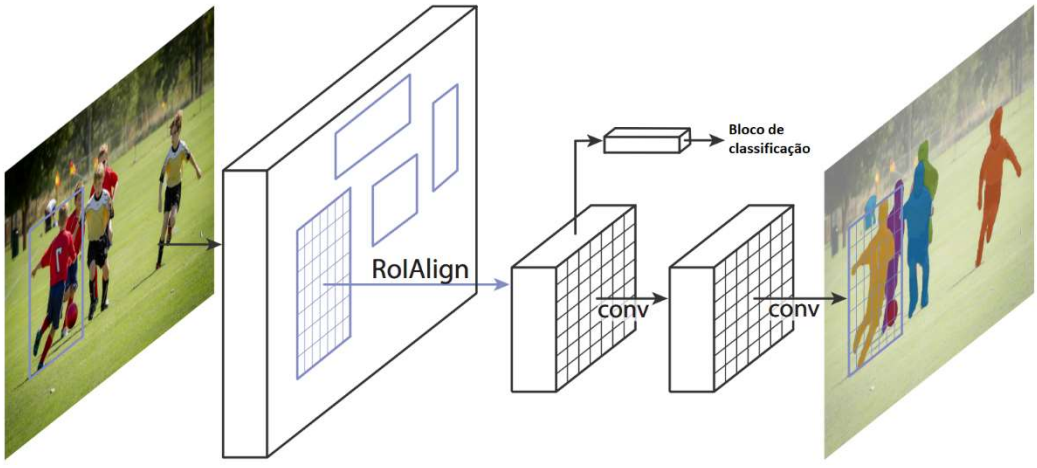

Figura 13. A estrutura da Mask R-CNN para segmentação de imagens. RoIAlign = camada da Mask R-CNN que alinha a área de interesse na imagem enquanto extrai suas características. Adaptado de Abdulla (98).

3.10 Construção e arquitetura dos modelos

3.10.1 Modelo de classificação categórico

Em nosso projeto, para a criação do modelo de classificação foi utilizado a linguagem de programação Python (87) e o TensorFlow (99), uma das principais bibliotecas de código aberto para desenvolvimento e criação de modelos de aprendizado de máquina.

No TensorFlow, cada imagem de entrada é tipicamente representada como um tensor 3D de formato [altura, largura, canais de cores]. O conjunto das imagens (*batch*) é representado como um tensor 4D de formato [tamanho do batch, altura, largura, canais de cores]. Por exemplo, se for usado um batch com 100 imagens coloridas de 256 x 256 pixels, teremos tensores 4D de formato [100, 256, 256, 3].

Os valores do canal RGB variam de 0 a 255. Isso não é o ideal para uma rede neural. Em geral, deve-se procurar deixar os valores de entrada pequenos, portanto, padroniza-se deixar esses valores variando entre 0 e 1, utilizando o reescalonamento. No caso do canal de cores, basta dividir o valor dos pixels por 255.

As imagens foram divididas em 4 diretórios: benigno, CaP Gleason 3, CaP Gleason 4 e CaP Gleason 5. O total das imagens utilizadas no treinamento foram divididas em 80% para treinamento e 20% como validação.

A arquitetura da CNN utilizada no nosso modelo foi a VGGNet (96). Essa arquitetura é simples e clássica, composta por várias camadas convolucionais seguidas de uma camada de pooling, repetindo esse padrão por diversas vezes (com um total de 16 camadas convolucionais). Ao final, uma rede densa é empregada com 2 camadas ocultas e a camada de saída (Figura 14). A VGGNet utiliza filtros 3 x 3, mas em grande quantidade.

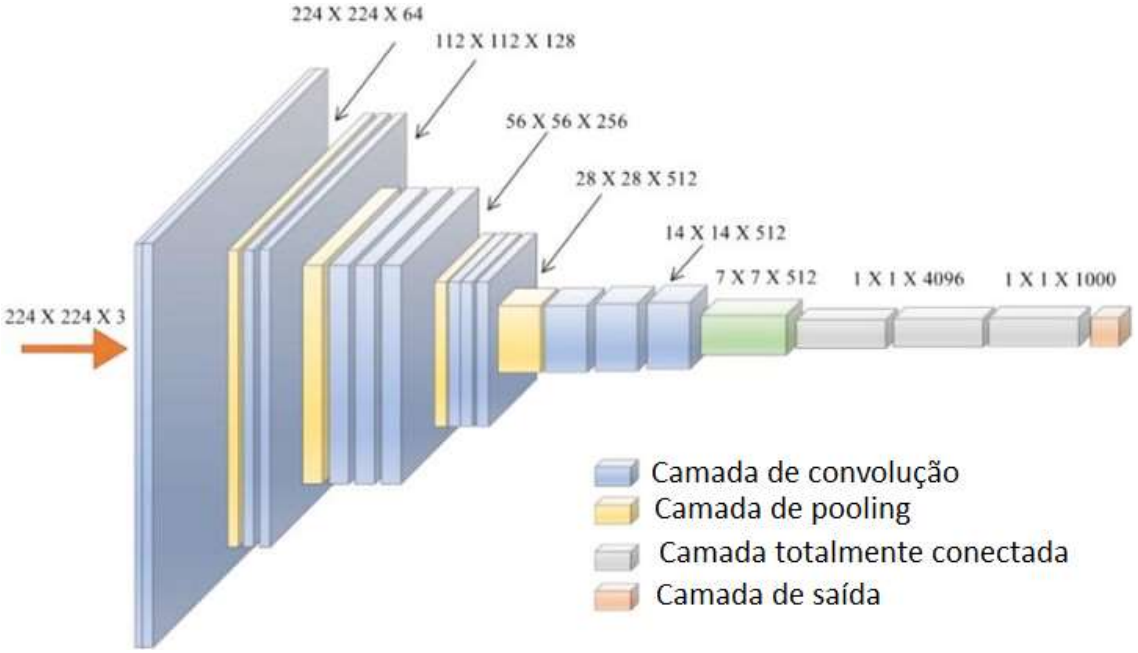

Figura 14. Arquitetura VGGNet. Adaptado de Simonyan (96).

Para treinamento das imagens foi utilizado uma GPU única NVIDIA GeForce GTX 1080 Ti (Nvidia Corporation, Santa Clara, Califórnia, EUA).

Dado que nosso número de imagens é relativamente pequeno e limitado, e o treinamento de uma CNN exige um número imenso de imagens, utilizamos a estratégia de Aumento de Dados (*Data Augmentation*) (100). O Aumento de Dados aumenta artificialmente o número de imagens ao gerar muitas variações realísticas de cada imagem de treinamento (Figura 15). Tal técnica ajuda a reduzir o *overfitting* do treinamento por gerar variações da imagem original. As imagens geradas devem ser o mais real possíveis, de forma que um humano não seria capaz de dizer se a imagem é produto de aumento ou não.

Pode-se, por exemplo, fazer rotação, desvio ou redimensionamento de cada imagem no grupo de treinamento. Essa tática força o modelo a ser mais tolerante a variações nas posições, orientações e tamanhos dos objetos nas figuras. Além disso, combinando diferentes transformações, é possível aumentar significativamente o tamanho do grupo de treinamento.

A aplicação de Aumento de Dados pode incluir operações como:

- Rotação: girar a imagem em um ângulo específico;

- Translação: deslocar a imagem horizontalmente ou verticalmente;
- Redimensionamento: aumentar ou diminuir o tamanho da imagem;
- Espelhamento: inverter a imagem horizontalmente ou verticalmente;
- Ajuste de brilho: modificar o brilho da imagem;
- Ajuste de contraste: modificar o contraste da imagem;
- Ruído: adicionar ruído aleatório à imagem.

Essas técnicas são aplicadas de maneira aleatória e combinadas, gerando um conjunto de imagens de treinamento mais diversificado e ajudando a aumentar a capacidade de generalização do modelo (101).

Durante o processo de treinamento, o modelo de CNN foi otimizado usando um algoritmo de otimização chamado Adam (102). A taxa de aprendizado, um hiperparâmetro que controla a velocidade de atualização dos pesos da rede, foi ajustada para garantir a convergência do modelo.

Além disso, foi aplicado uma técnica de regularização chamada de *Dropout* (103) para minimizar o *overfitting*. O Dropout desliga aleatoriamente uma fração dos neurônios durante o treinamento, o que força a rede a aprender representações mais robustas e generalizáveis.

Em resumo, a construção do modelo de classificação categórico envolveu o uso de uma arquitetura de CNN (VGGNet), a aplicação do Aumento de Dados para aumentar a quantidade e a diversidade das imagens de treinamento, e a utilização de técnicas de otimização e regularização para garantir a eficácia e a generalização do modelo.

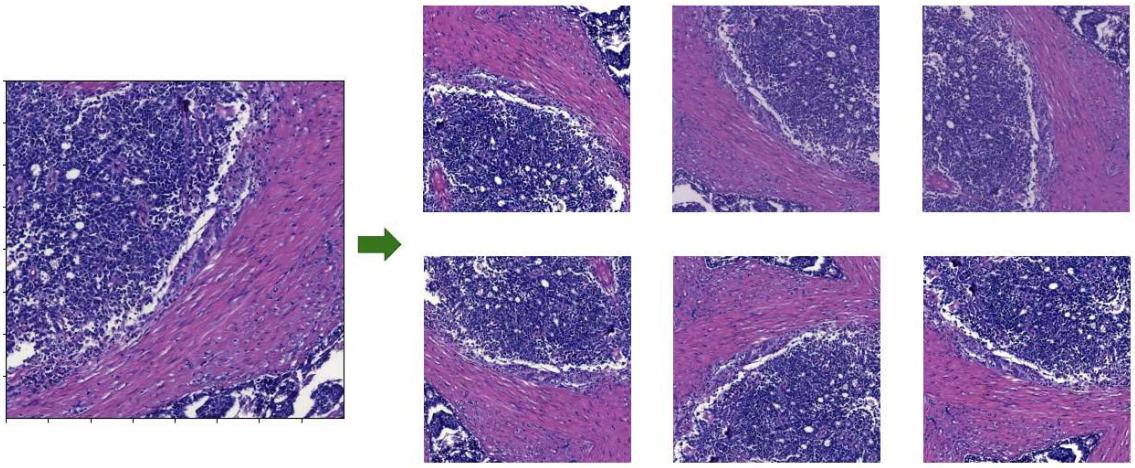

Figura 15. Gerando novas imagens a partir das imagens originais pré-existentes

3.10.2 Modelo de segmentação de instâncias

Para o modelo de segmentação de instâncias, utilizamos a linguagem de programação Python, a biblioteca TensorFlow e a plataforma Matterport para implementar a Mask R-CNN (98). A Mask R-CNN é uma extensão do modelo Faster R-CNN, desenvolvida especificamente para abordar a segmentação de instâncias em imagens.

Todas as imagens utilizadas no treinamento foram armazenadas no mesmo diretório. As anotações aplicadas manualmente a cada imagem, delimitando o padrão tumoral ou de benignidade, foram realizadas com a ferramenta CocoAnnotator (88). Essas anotações foram convertidas em um arquivo no formato JSON que pode ser interpretado pela linguagem Python.

Utilizamos uma única GPU NVIDIA GeForce GTX 1080 Ti para treinar a Mask R-CNN. O treinamento envolveu o uso de máscaras geradas a partir das anotações, que segmentam as imagens e as áreas de interesse (Figura 16). Com base nessas máscaras, a Mask R-CNN foi treinada para criar suas próprias máscaras e realizar predições em novas imagens (Figura 17).

A arquitetura da Mask R-CNN consiste em três componentes principais: a rede de backbone, a Região de Interesse (RoI) Pooling e a cabeça de predição.

A rede de backbone, geralmente uma CNN pré-treinada, como a ResNet ou a VGG, é responsável por extrair características das imagens de entrada. O RoI Pooling, por sua vez, extrai informações relevantes das características extraídas e as organiza em regiões de interesse. Finalmente, a cabeça de predição utiliza a informação das RoIs para realizar a classificação dos objetos, regressão das caixas delimitadoras e predição das máscaras de segmentação (90).

Ao longo do treinamento, a Mask R-CNN aprende a identificar e segmentar objetos nas imagens, mesmo aqueles que pertencem à mesma categoria, graças ao seu design específico para segmentação de instâncias. Essa capacidade permite que o modelo seja utilizado para segmentar e classificar padrões tumorais ou de benignidade em imagens médicas, auxiliando no diagnóstico e na análise das condições do paciente.

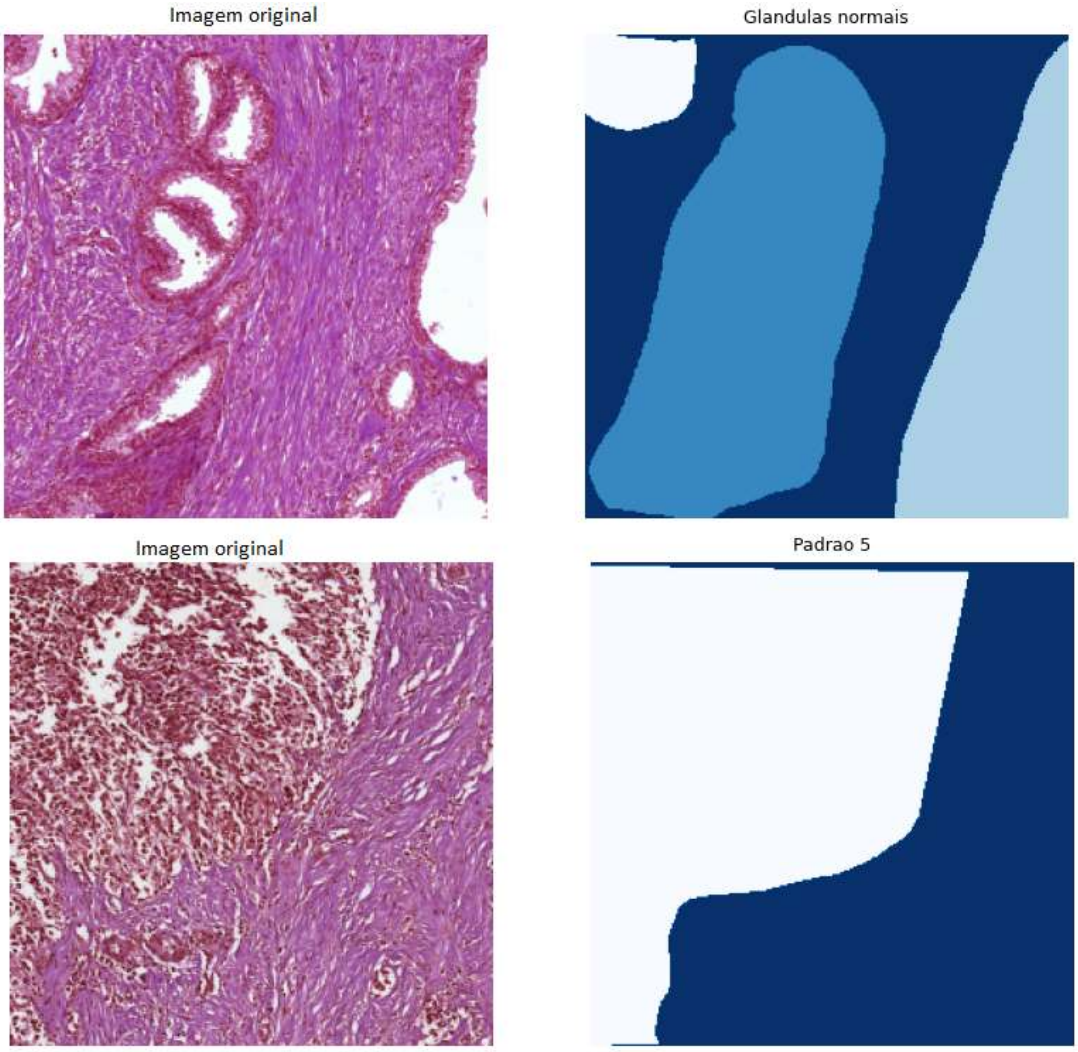

Figura 16. Imagens originais utilizadas no treinamento e suas máscaras correspondentes criadas a partir das anotações do patologista

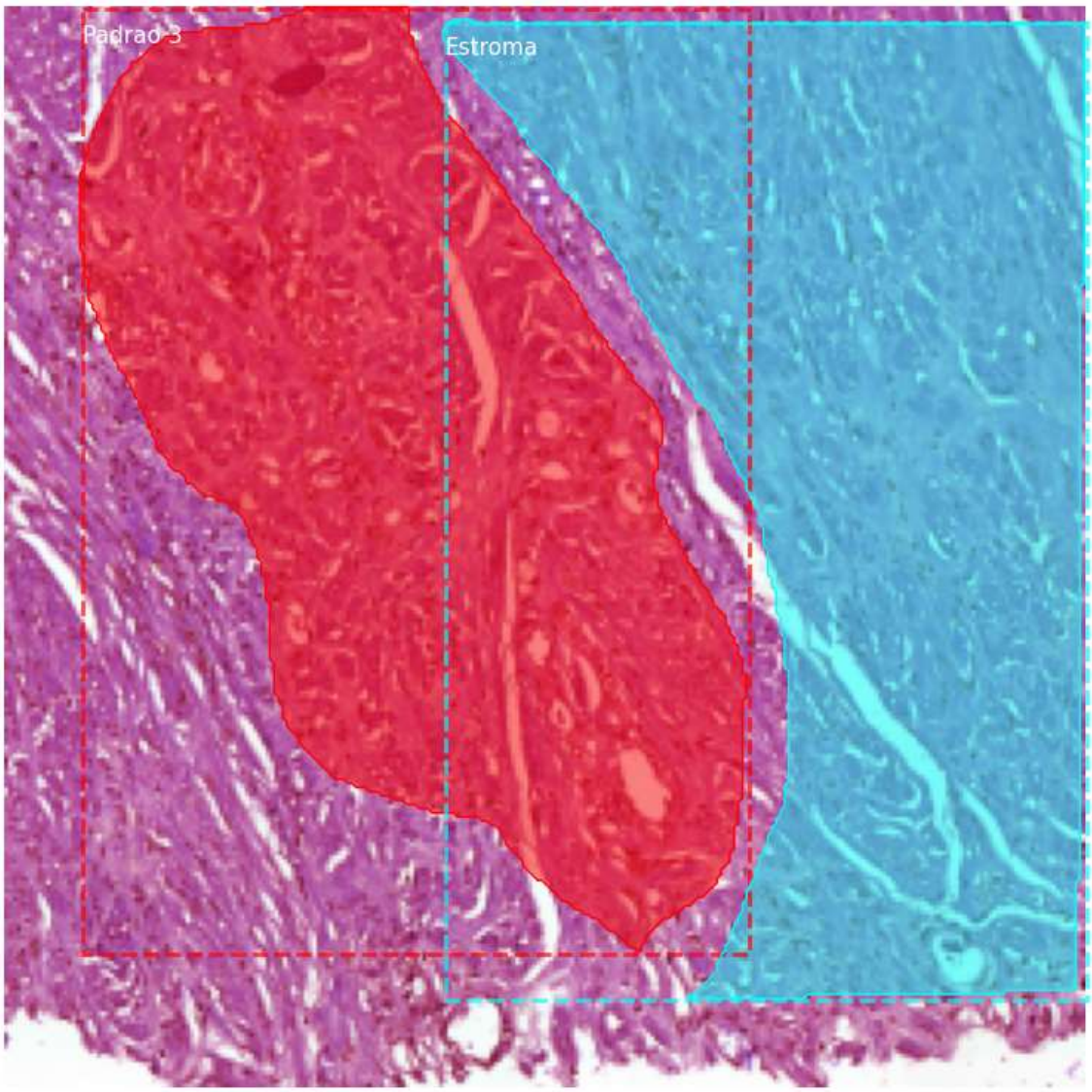

Figura 17. Máscaras e caixas delimitadoras sobrepostas à imagem original com cada padrão destacado.

3.11 Divisão em grupos de treinamento, validação e teste

A avaliação do desempenho de um modelo de aprendizado profundo em cenários reais requer o uso de exemplos não envolvidos no processo de treinamento. Para isso, é comum dividir o conjunto de dados em três grupos distintos: treinamento, validação e teste (104). Esta divisão permite uma análise mais eficiente e realista da capacidade do modelo em lidar com situações

práticas e fornece informações sobre o seu desempenho em contextos desconhecidos.

No presente estudo, o conjunto de imagens foi dividido de acordo com a seguinte estratégia: 80% das imagens foram destinadas ao treinamento e 20% à validação. Esses grupos eram intercalados em suas funções, permitindo uma avaliação mais precisa e consistente do desempenho do modelo. Além disso, uma parcela das imagens foi reservada para o grupo de teste, que não foi utilizada em nenhum momento durante o treinamento.

A divisão dos dados em grupos de treinamento e validação é crucial para o ajuste adequado do modelo e a prevenção do sobreajuste (*overfitting*) (54). O grupo de treinamento é utilizado para o ajuste dos parâmetros do modelo, enquanto o grupo de validação é empregado para avaliar o desempenho do modelo ao longo do processo de treinamento. O uso do grupo de validação também possibilita a seleção do melhor modelo com base no menor erro de generalização, que é a medida da capacidade do modelo de se adaptar a novos casos.

Ao longo do treinamento, os grupos de treinamento e validação são rotacionados, o que proporciona uma abordagem de validação cruzada mais robusta e ajuda a evitar a seleção de um modelo que seja específico para um determinado subconjunto dos dados. A validação cruzada é uma técnica amplamente utilizada em aprendizado de máquina para avaliar a capacidade de generalização de um modelo, garantindo que o modelo seja capaz de lidar com a variabilidade presente nos dados.

Após o processo de validação cruzada, o modelo final é selecionado com base em seu desempenho no grupo de validação. Este modelo é então treinado novamente utilizando todo o conjunto de treinamento, incluindo as imagens usadas na validação, para maximizar a quantidade de informações disponíveis para o ajuste dos parâmetros. O modelo final assim obtido representa a melhor combinação de parâmetros para o problema em questão e possui a maior capacidade de generalização possível.

Finalmente, o modelo é testado no grupo de teste, que consiste nas imagens reservadas no início do processo e que não foram utilizadas em nenhum momento durante o treinamento. O desempenho do modelo neste grupo

fornece uma estimativa realista do erro de generalização, ou seja, a capacidade do modelo de se adaptar a novos casos e realizar previsões corretas em situações desconhecidas.

O processo de divisão das imagens em grupos de treinamento, validação e teste é uma etapa fundamental no desenvolvimento de modelos de aprendizado profundo para a análise de imagens histológicas. Esta abordagem permite uma avaliação mais confiável e abrangente do desempenho do modelo e garante que as informações extraídas a partir dos dados sejam representativas do problema em questão. Além disso, a utilização de grupos de validação e teste permite uma melhor compreensão das limitações do modelo e das áreas em que pode haver necessidade de melhorias futuras.

3.12 Aplicação do modelo treinado em imagens de biópsia de próstata

Após o treinamento do nosso modelo de aprendizado profundo utilizando imagens de lâminas histopatológicas de prostatectomia radical, o próximo passo crucial foi testar a robustez e a aplicabilidade do modelo em um cenário clínico real. A meta desta fase é determinar se o melhor modelo treinado poderia classificar efetivamente imagens de biópsias de próstata, preparadas com hematoxilina e eosina (H&E) e fotografadas com microscópio óptico.

Para garantir a validade e a robustez do estudo, foi utilizada uma amostra de imagens de biópsias de próstata de pacientes que não estiveram envolvidos em nenhuma fase de treinamento do modelo. No treinamento, foram usadas somente lâminas de prostatectomia radical. As lâminas de biópsia de próstata foram escolhidas por um especialista em uropatologia para garantir que representassem uma variedade de estados histológicos. A seleção considerou lâminas com padrões benigno e de câncer com escore Gleason 3, 4 e 5.

Diferentemente das imagens usadas no treinamento que consistiam em imagens menores divididas (*tiling*) das lâminas originais, foram utilizadas as imagens de biópsia completas em seu tamanho original, sem nenhuma divisão e sem pré-processamento (Figura 18). A razão para isso é reproduzir as condições reais encontradas por patologistas, incluindo variações na qualidade

e no foco da imagem, e outras potenciais artefatos. As imagens foram capturadas em diferentes níveis de zoom, para melhor simular o ambiente de diagnóstico real.

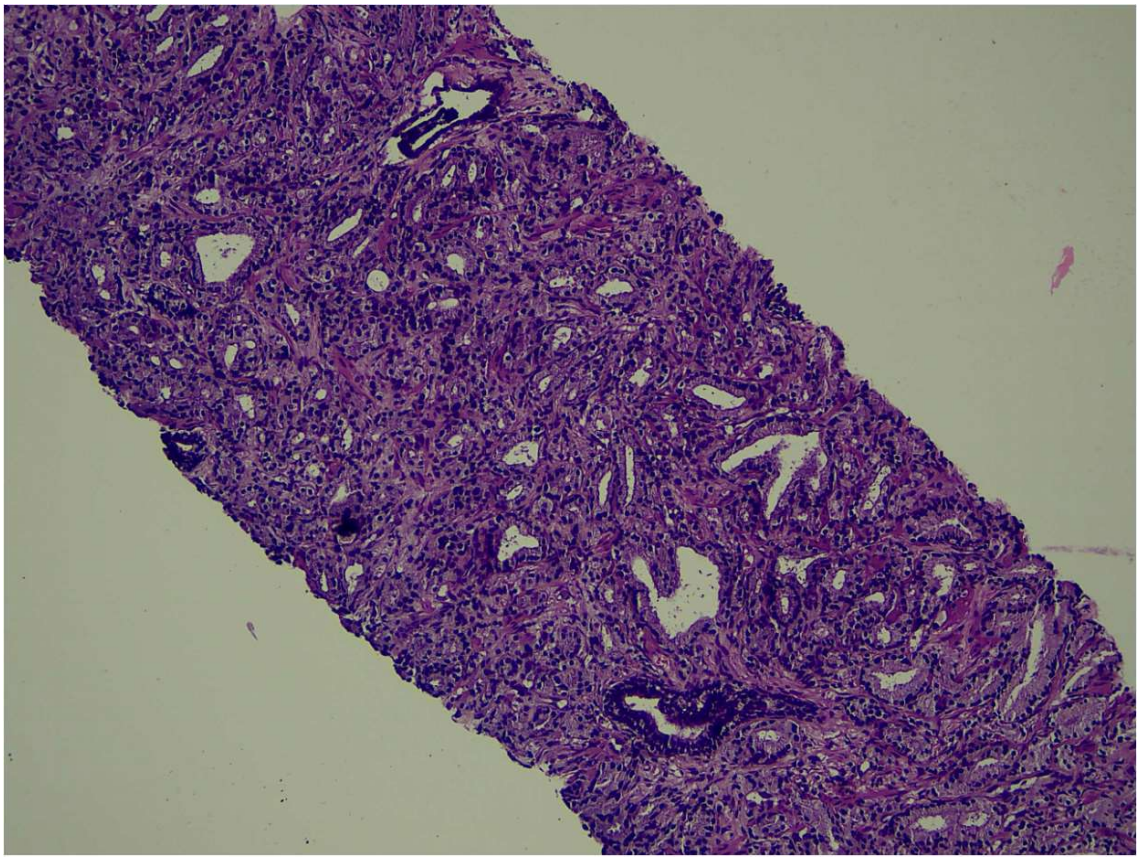

Figura 18. Exemplo de imagem de lâmina de biópsia de próstata.

Entre os métodos de classificação categórica e de segmentação por instância, foi escolhido para aplicação nas imagens de biópsia de próstata aquele modelo que apresentasse melhor performance no seu grupo teste de imagens de prostatectomia radical.

Após a aplicação do modelo, todas as imagens classificadas e segmentadas foram revisadas por um uropatologista expert. Os resultados foram comparados para determinar o grau de concordância entre a IA e o especialista.

Ambas as anotações (modelo e uropatologista) foram alinhadas e comparadas para avaliar o nível de concordância. As discrepâncias entre as duas anotações foram meticulosamente registradas e categorizadas, permitindo

uma análise subsequente para identificar quais tipos de tecido ou condições específicas apresentavam maior divergência entre as avaliações humana e automatizada.

3.13 Análise estatística

Para efetuar uma comparação metodológica abrangente e rigorosa entre as anotações do modelo e da uropatologista, foi primordial a implementação de métricas robustas, capazes de refletir não apenas a capacidade do modelo de classificar corretamente, mas também sua habilidade em discernir nuances entre tecidos com diferenças mínimas.

No contexto de classificação categórica utilizando VGGNet, destacamos a métrica ROC-AUC (Característica de Operação do Receptor - Área Sob a Curva). A ROC-AUC varia entre 0 e 1, onde um valor de 1 indica uma performance perfeita e 0,5 sugere uma capacidade de discriminação não melhor que a atribuição aleatória. Neste estudo, a ROC-AUC foi calculada individualmente para cada categoria, oferecendo insights específicos sobre o desempenho do modelo em relação a cada padrão de Gleason, bem como para o tecido benigno.

Em adição, a matriz de confusão foi empregada como uma ferramenta visual e quantitativa para identificar claramente onde o modelo concordava ou divergia das anotações da uropatologista. Este recurso é fundamental para visualizar o número real de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos para cada categoria.

Foram avaliadas, também, a acurácia, precisão e sensibilidade, sendo esta última de particular relevância em diagnósticos médicos, onde a detecção de amostras patológicas é crítica.

A acurácia é a proporção de predições corretas em relação ao total de predições. É uma métrica comum para avaliar a eficácia geral de um modelo de classificação. A precisão é a proporção de verdadeiros positivos (imagens classificadas corretamente) em relação à soma dos verdadeiros positivos e falsos positivos (imagens classificadas incorretamente como pertencentes a uma classe específica). A precisão é uma medida de quão bem o modelo identifica

corretamente uma classe sem gerar falsos positivos. A sensibilidade mede a proporção de verdadeiros positivos que são corretamente identificados pelo modelo.

Adicionalmente, a especificidade foi considerada, uma vez que mede a proporção de verdadeiros negativos que são corretamente identificados. Uma alta especificidade é crucial para garantir que o modelo não identifique incorretamente tecidos benignos como malignos, o que poderia levar a intervenções médicas desnecessárias.

Ao analisarmos a segmentação por instâncias com o modelo Mask R-CNN, a métrica de destaque foi o coeficiente de Dice. Este índice mede a similaridade entre dois conjuntos e é frequentemente usado em contextos de segmentação de imagem médica para comparar a concordância entre uma segmentação automática e uma manual.

No nosso cenário, o coeficiente de Dice serve como uma métrica adicional para avaliar a concordância entre as categorizações feitas pelo modelo e pela uropatologista. Uma alta similaridade indica eficácia na segmentação das estruturas de interesse. Um coeficiente de Dice próximo de 1 indica alta concordância entre os dois conjuntos, enquanto um valor próximo de 0 sugere baixa concordância.

Adicionalmente, utilizamos o coeficiente Kappa de Cohen para a VGGNet, avaliando a concordância entre classificações observadas e previstas. Este índice é essencial para determinar se o modelo vai além de meras predições aleatórias.

Ao utilizar este conjunto compreensivo de métricas, buscamos oferecer uma visão panorâmica e detalhada do desempenho do modelo, elucidando suas forças e áreas de melhoria. Esta abordagem metodológica se mostrou essencial para o avanço na aplicação de modelos de aprendizado profundo no campo da uropatologia, contribuindo significativamente para o objetivo maior de melhorar o diagnóstico e tratamento do CaP.

3.14 Ética

O presente estudo foi conduzido em conformidade com as diretrizes éticas estabelecidas e obteve aprovação do Comitê de Ética em Pesquisa da Faculdade de Medicina da Universidade de São Paulo. CAAE: 16159019.1.0000.0065. Número do parecer: 3.689.656.

Todas as análises realizadas neste estudo foram conduzidas de maneira anônima, garantindo a confidencialidade e a privacidade dos pacientes envolvidos. Não houve manipulação do material de arquivo ou exposição de dados pessoais dos pacientes que pudessem comprometer sua segurança ou integridade. Dado o caráter retrospectivo e anônimo do estudo, não foi necessário obter o consentimento informado dos pacientes.

4. RESULTADOS

Apresentaremos agora os nossos resultados obtidos pelos modelos de classificação categórica (VGGNet) e segmentação de instâncias (Mask R-CNN) aplicados ao diagnóstico do CaP. Todo o conjunto de dados foi dividido em subconjuntos de treinamento, validação e teste, garantindo uma avaliação justa e robusta do desempenho dos modelos.

Descreveremos o desempenho geral de cada modelo, bem como as métricas de avaliação utilizadas para quantificar sua precisão, sensibilidade, especificidade e outros indicadores relevantes.

Além disso, foi realizado a avaliação de lâminas de biópsia de próstata utilizando o modelo treinado que obteve melhor desempenho com as lâminas de prostatectomia radical.

4.1 Resultados do modelo de classificação categórica (VGGNet)

4.1.1 Desempenho geral do modelo de classificação categórica

Um total de 36 lâminas completas de diferentes pacientes submetidos à prostatectomia radical foi utilizado neste estudo. Para análise através do método categórico, cada imagem recebia um único rótulo baseado se havia somente tecido benigno amostrado ou, em caso de presença de câncer, no padrão histológico predominante naquela imagem.

O conjunto de dados foi dividido em três partes: treinamento, validação e teste. Para o treinamento, foram utilizadas 2020 imagens pertencentes a 4 classes distintas: CaP Gleason 3, CaP Gleason 4, CaP Gleason 5 e benigno, sendo 505 imagens de cada categoria. O conjunto de validação contou com 563 imagens, distribuídas entre as mesmas quatro classes, sendo 135 imagens do padrão de Gleason 3, 135 do padrão de Gleason 4, 136 do padrão de Gleason 5 e 157 imagens benignas. Todas essas imagens foram utilizadas em algum momento no processo de treinamento do modelo. A quantidade total de imagens usadas durante o processo completo de treinamento em cada grupo está disposta na Tabela 1.

Por fim, o conjunto de teste continha 100 imagens, também divididas entre as quatro classes, que não foram utilizadas em nenhum momento durante todo o processo de treinamento do modelo.

Tabela 1. Características das imagens usadas no processo de treinamento do método de classificação categórico

Número de lâminas de prostatectomia radical	**36**
Número total de imagens utilizadas	**2.583**
Imagens somente com tecido benigno	**662**
Imagens com padrão 3 predominante	**640**
Imagens com padrão 4 predominante	**640**
Imagens com padrão 5 predominante	**641**

O modelo VGGNet utilizado neste estudo foi pré-treinado com pesos do ImageNet e adaptado às necessidades específicas do problema em questão, ajustando-se a uma resolução de imagem de 300 x 300 pixels e considerando 4 classes de saída. O otimizador Adam foi empregado, com uma taxa de aprendizado de 1e-5, e o modelo foi compilado com a função de perda "*categorical_crossentropy*" e a métrica de acurácia.

O treinamento ocorreu ao longo de 100 ciclos (*epochs)*, e o Aumento de Dados foi aplicado ao conjunto de treinamento, a fim de melhorar o desempenho do modelo e evitar o *overfitting*. A estratégia de Aumento de Dados foi meticulosamente empregada para enriquecer o conjunto de treinamento e potencializar a generalização do modelo. Foram adotadas técnicas como redimensionamento, rotação, deslocamento horizontal e vertical, zoom e inversão horizontal das imagens. Essas técnicas foram escolhidas com base na natureza das lâminas histológicas, onde variações sutis, como uma rotação ou deslocamento, podem mimetizar variações reais encontradas no preparo e análise das amostras em diferentes laboratórios e contextos.

4.1.2 Métricas de avaliação do modelo de classificação categórica

Ao longo dos 100 ciclos de treinamento, o modelo de classificação categórica VGGNet apresentou um desempenho consistente, tendo uma função de perda de 0,0699 e uma acurácia de 97,57%. No conjunto de validação, a perda foi de 0,1403 e a acurácia de 95,74%.

No entanto, ao avaliar o modelo no conjunto de teste, a acurácia obtida foi de apenas 45%, com uma função de perda de 8,8852. Considerando a complexidade da tarefa e a importância do diagnóstico preciso do CaP, essa acurácia é insatisfatória e aponta para a necessidade de melhorias no modelo.

A matriz de confusão fornece uma representação visual do desempenho do modelo em cada uma das classes e permite calcular métricas importantes, como precisão, sensibilidade e especificidade (Figura 19).

- Benigno: Precisão: 0,46, Sensibilidade: 0,48, Especificidade: 0,82
- Gleason 3: Precisão: 0,60, Sensibilidade: 0,60, Especificidade: 0,87
- Gleason 4: Precisão: 0,38, Sensibilidade: 0,35, Especificidade: 0,81
- Gleason 5: Precisão: 0,36, Sensibilidade: 0,38, Especificidade: 0,79

A partir desses valores, podemos observar que o modelo apresenta melhor desempenho na classificação de CaP Gleason 3 e lesões benignas, enquanto enfrenta dificuldades para classificar corretamente as amostras de CaP Gleason 4 e 5.

Além das métricas acima mencionadas, foi calculado o coeficiente Kappa de Cohen, uma métrica estatística robusta, para avaliar a concordância entre as classificações observadas e as previstas pelo modelo. O coeficiente Kappa de Cohen obtido foi de $\kappa = 0,267$. Esse valor aponta para uma concordância moderada entre as classificações do modelo e as verdadeiras, indicando que, embora o modelo tenha conseguido algumas classificações corretas além do que seria esperado pelo acaso, ainda há um caminho considerável a ser percorrido para se atingir um nível aceitável de precisão para aplicações clínicas.

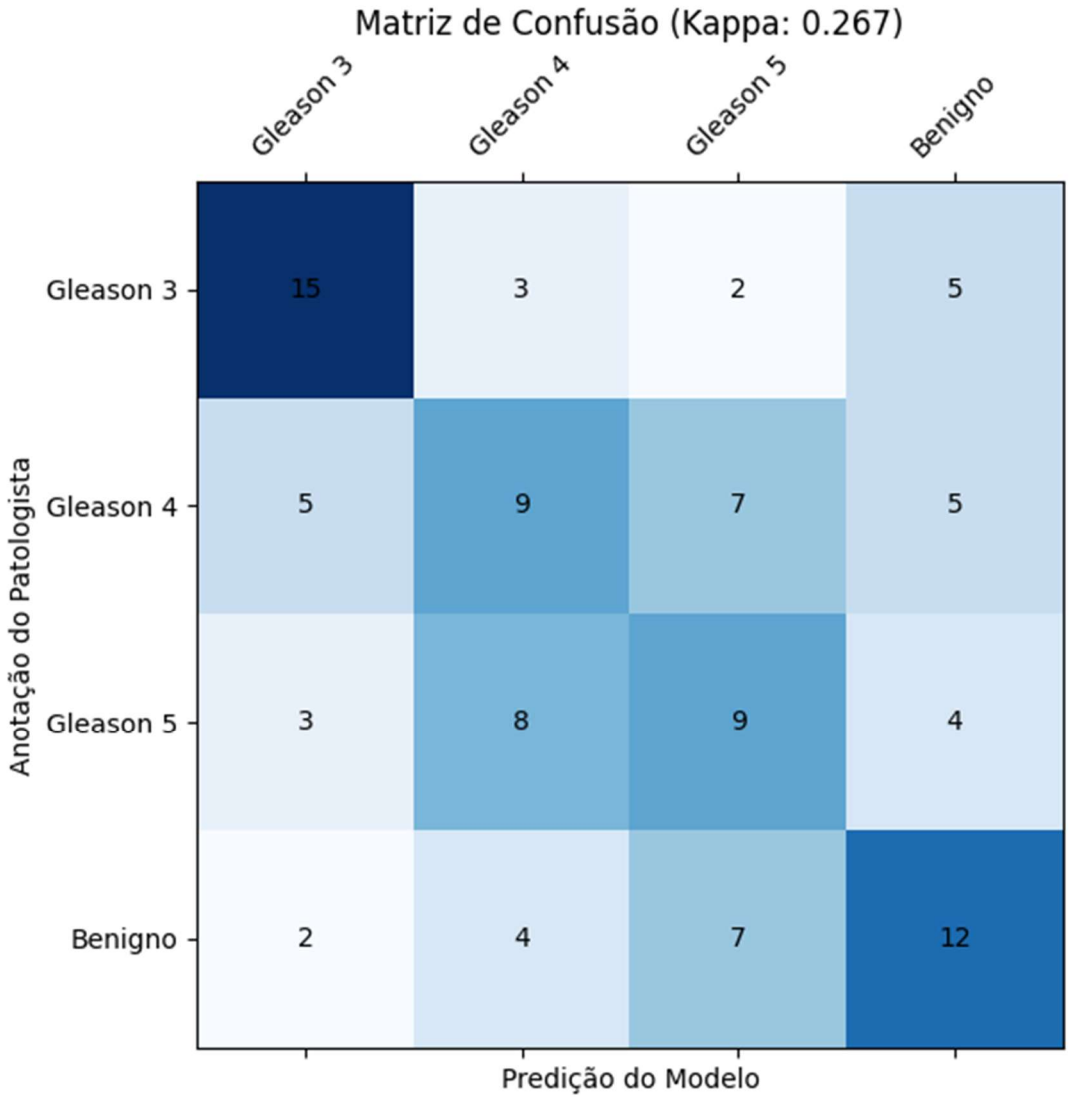

Figura 19. Matriz de confusão do modelo de classificação categoria VGGNet. Classificação feita pelo uropatologista versus predição automática do modelo

4.2 Resultados do modelo de segmentação de instâncias (Mask R-CNN)

4.2.1 Desempenho geral do modelo de segmentação de instâncias

O modelo Mask R-CNN foi treinado e validado utilizando um conjunto de dados composto por 36 lâminas de prostatectomia radical, segmentadas em cinco categorias distintas, sendo elas: glândulas normais, estroma, CaP Gleason 3, CaP Gleason 4 e CaP Gleason 5. Para o treinamento e validação do modelo, foram utilizadas imagens de tamanho 256 x 256 pixels. A validação cruzada foi realizada utilizando uma divisão de 5 *folds*.

O processo de treinamento envolveu o uso de várias técnicas de aumento de dados, como inversão horizontal e vertical, rotação e ajuste de contraste. O modelo foi treinado em três etapas, com as seguintes configurações:

- aprendizado apenas das camadas "*heads*" do modelo com uma taxa de aprendizado de 1e-4 durante 20 ciclos;

- aprendizado de todas as camadas do modelo com uma taxa de aprendizado de 1e-4 durante 60 ciclos;

- aprendizado de todas as camadas do modelo com uma taxa de aprendizado reduzida para 1e-5 durante 100 ciclos.

As anotações usadas no treinamento e validação foram distribuídas da maneira exposta na Tabela 2.

Durante o processo de treinamento, empregamos técnicas de Aumento de Dados, tais como rotações, inversões e desfoques, com o intuito de aumentar a diversidade do conjunto de treinamento e, consequentemente, melhorar a capacidade de generalização do modelo. O *backbone* da arquitetura do modelo Mask R-CNN foi a ResNet101.

Tabela 2. Quantidade de anotações feitas nas imagens usadas no treinamento do modelo de segmentação de instâncias (Mask R-CNN):

Número de lâminas de prostatectomia radical	**36**
Número total de imagens utilizadas	**6.160**
Total de anotações usadas no treinamento	**8.367**
Glândulas normais	**3.982**
Estroma	**3.049**
Gleason 3	**858**
Gleason 4	**2.321**
Gleason 5	**1.361**

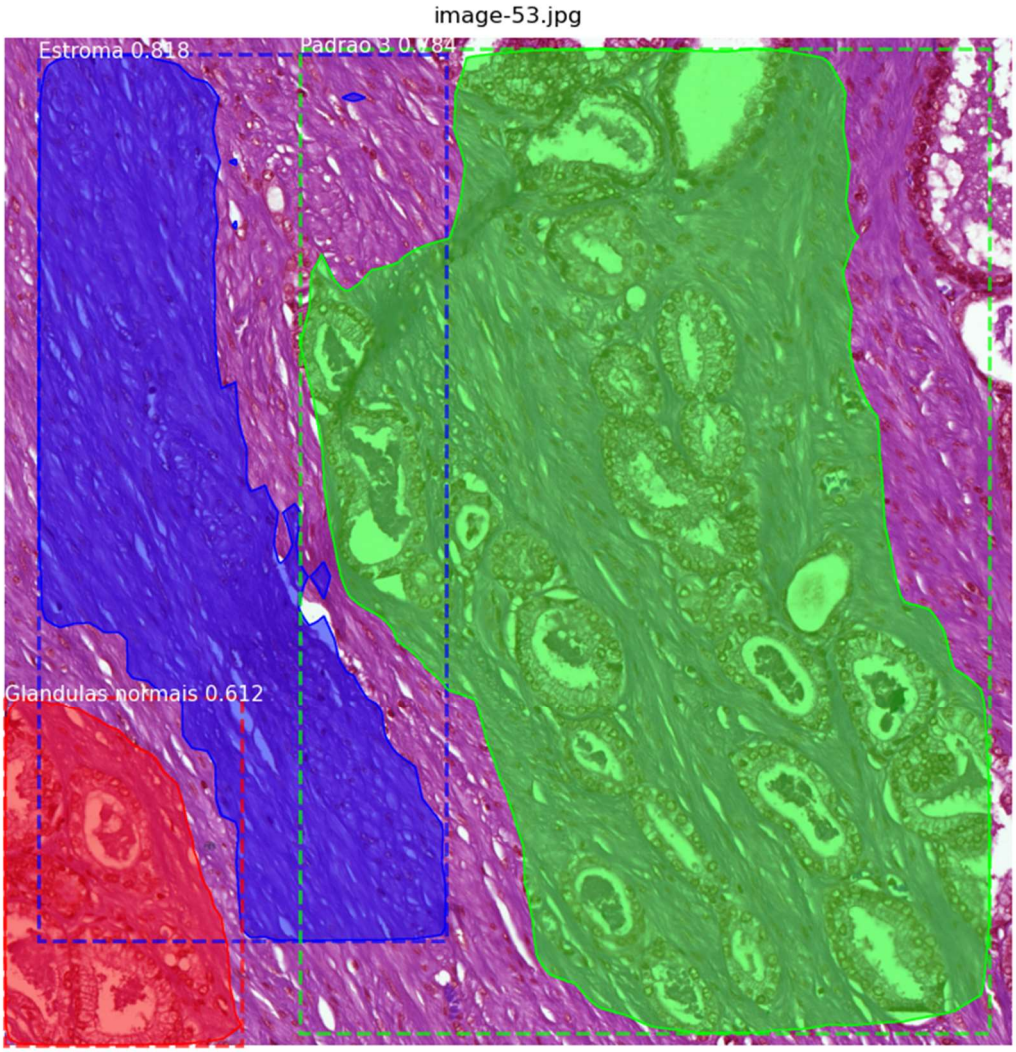

Figura 20. Na mesma imagem, áreas com glândulas normais, estroma e câncer padrão de Gleason 3, detectadas pelo modelo de Mask R-CNN automaticamente. Vermelho = glândulas normais, azul = estroma, verde = câncer de próstata padrão de Gleason 3

4.2.2 Métricas de avaliação do modelo de segmentação de instâncias

Durante o treinamento, o modelo apresentou uma acurácia de treinamento de 95,1%, indicando que ele aprendeu a distinguir efetivamente as diferentes categorias no conjunto de treinamento. Além disso, o modelo alcançou uma acurácia de validação de 93,2%.

As métricas de desempenho específicas da Mask R-CNN após a conclusão do último ciclo incluem uma perda total de 0,8704, composta pelas

seguintes contribuições: perda de classificação RPN de 0,01894, perda de caixa delimitadora RPN de 0,159, perda de classificação MRCNN de 0,1015, perda de caixa delimitadora MRCNN de 0,2324 e perda de máscara MRCNN de 0,3585.

Após a conclusão do treinamento, o modelo foi exposto a 100 novas imagens nunca vistas, separadas antes do treinamento e não usadas durante ele em nenhum momento. Para avaliar o desempenho do modelo no conjunto de teste, agrupamos as categorias "glândulas normais" e "estroma" como "benigno". Dessa forma, as métricas de avaliação foram calculadas considerando as seguintes categorias: CaP Gleason 3, CaP Gleason 4, CaP Gleason 5 e benigno.

O modelo demonstrou um coeficiente de Dice de 0,89, indicando uma alta concordância entre a segmentação predita e a verdadeira. Em adição, alcançou uma acurácia geral de 0,89 no conjunto de teste.

No que tange à área sob a curva ROC-AUC para as diferentes categorias, foram obtidos os seguintes valores (Figura 21):

- Benigno (composto pelas categorias "glândulas normais" e "estroma"): 0,8949
- Câncer de próstata Gleason 3: 0,9493
- Câncer de próstata Gleason 4: 0,9595
- Câncer de próstata Gleason 5: 0,9545

A matriz de confusão obtida no conjunto de teste é apresentada na Figura 22. A partir dessa matriz, podemos calcular as seguintes métricas de avaliação:

- Benigno: Precisão: 0,94, Sensibilidade: 0,68, Especificidade: 0,99
- Gleason 3: Precisão: 0,93, Sensibilidade: 0,93, Especificidade: 0,97
- Gleason 4: Precisão: 0,96, Sensibilidade: 0,93, Especificidade: 0,99
- Gleason 5: Precisão: 0,77, Sensibilidade: 1,00, Especificidade: 0,91

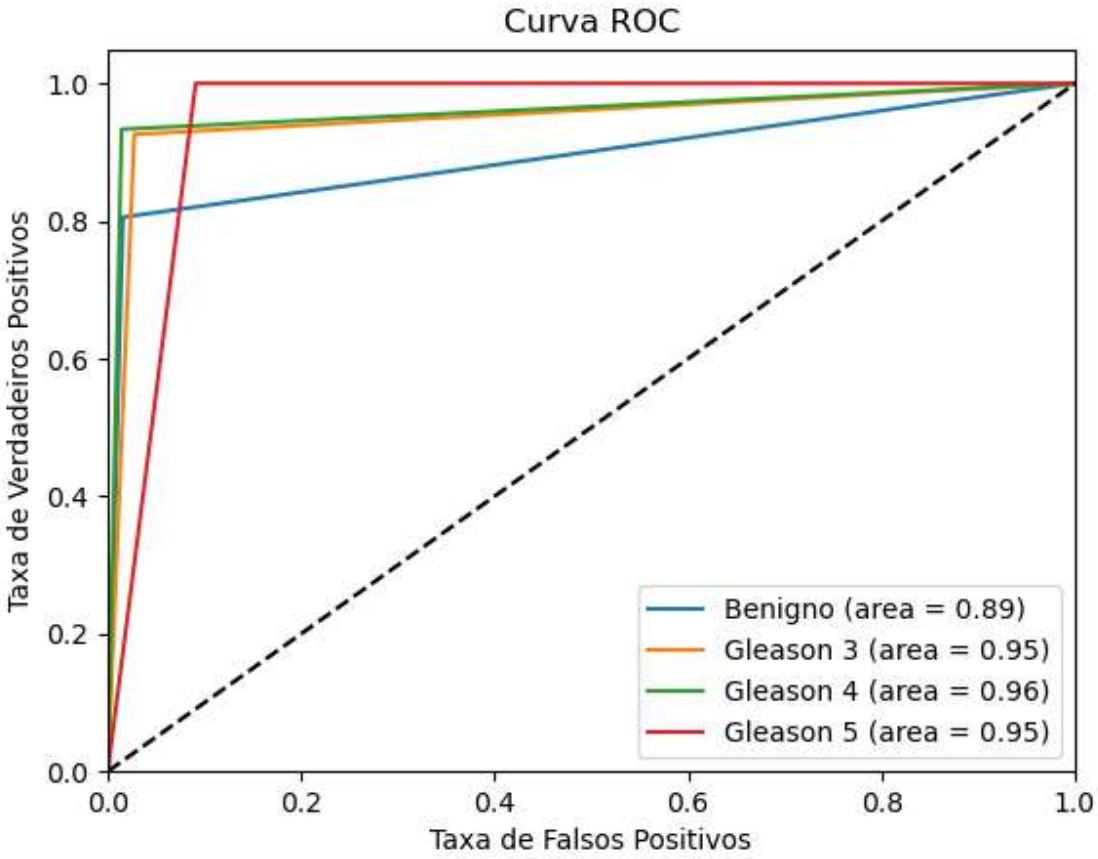

Figura 21. Análise da curva ROC nas lâminas teste do modelo treinado de Mask R-CNN.

O modelo Mask R-CNN demonstrou um desempenho consideravelmente bom no conjunto de teste. A precisão e a sensibilidade indicam que o modelo foi capaz de identificar corretamente as diferentes categorias de tecidos e lesões presentes nas lâminas de prostatectomia radical. A especificidade também mostrou resultados satisfatórios, sugerindo que o modelo teve um bom desempenho na identificação de verdadeiros negativos.

Entretanto, ainda há espaço para melhorias. Por exemplo, a sensibilidade na identificação do CaP Gleason 5 em comparação com as outras categorias indicam que o modelo pode ter tido mais dificuldades com essa categoria em específico. Isso pode ser devido à complexidade das características morfológicas associadas ou à quantidade de dados disponíveis para treinamento.

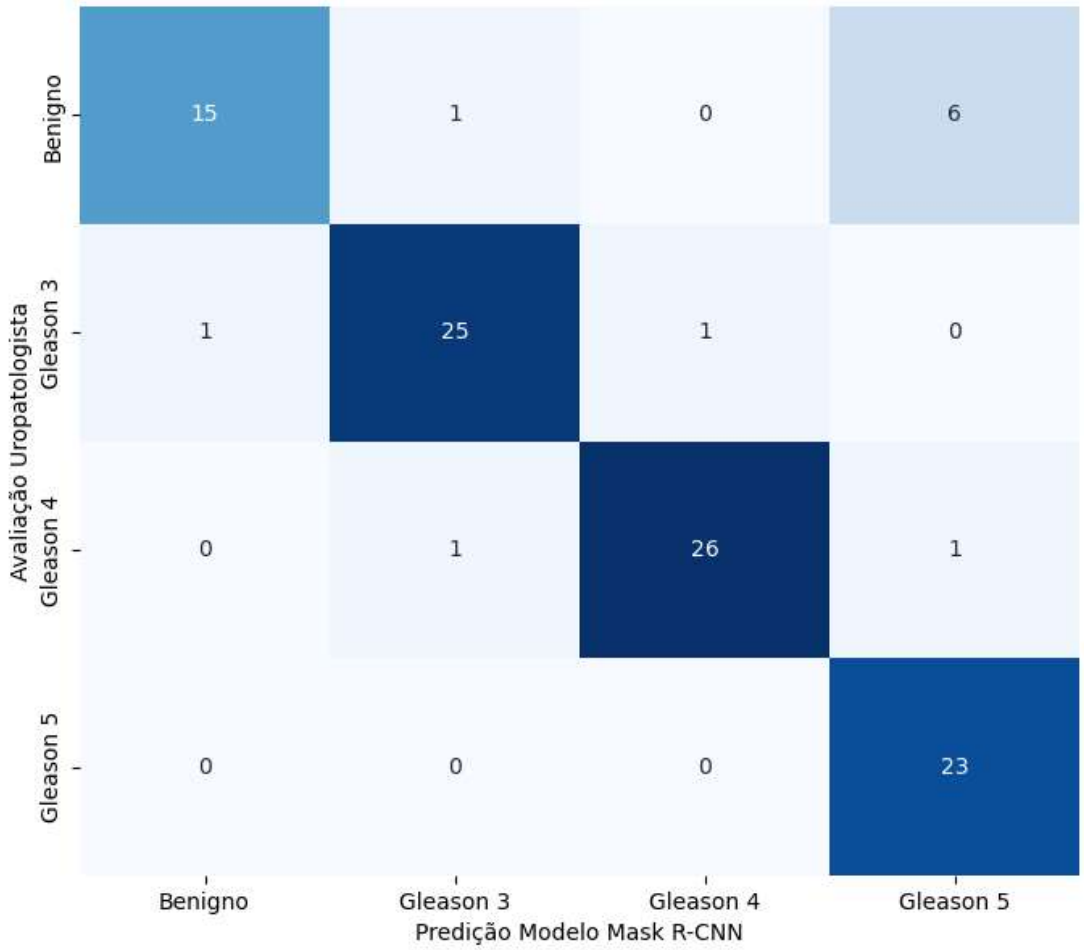

Figura 22. Matriz de confusão do modelo de segmentação por instâncias. Anotação feita pelo uropatologista versus predição automática do modelo

4.3 Aplicação do modelo de melhor desempenho nas lâminas de biópsia de próstata

4.3.1 Seleção do modelo de aprendizado profundo e dos casos de biópsia de próstata

O modelo de segmentação de instâncias utilizando a Mask R-CNN foi selecionado para análise das lâminas de biópsia de próstata devido ao seu desempenho superior. Quando comparado à classificação categórica usando a VGGNet, o modelo Mask R-CNN mostrou-se mais eficaz em classificar corretamente as amostras de tecido.

Para a análise em questão, foram selecionadas 172 imagens de lâminas de biópsias de próstata, provenientes de um conjunto de 21 casos clínicos distintos aleatórios de pacientes diagnosticados com CaP.

Adicionalmente, um conjunto controle de 120 biópsias benignas foi também incluído na análise. Esta seleção foi feita para fornecer um contraste necessário e permitir a avaliação do poder discriminativo do modelo em relação a tecidos normais e patológicos.

Utilizando o modelo de Mask R-CNN, foram automaticamente geradas máscaras que identificavam áreas de interesse nas lâminas de biópsia de próstata. Estas máscaras foram utilizadas para classificar os diferentes tecidos nas amostras em 4 categorias específicas: tecido benigno, câncer com padrão de Gleason 3, câncer com padrão de Gleason 4 e câncer com padrão de Gleason 5.

Paralelamente, uma uropatologista expert (K.R.M.L), avaliou as mesmas lâminas de biópsia e fez suas anotações de forma manual, também classificando os tecidos nas mesmas quatro categorias: benigno, Gleason 3, Gleason 4, e Gleason 5. Ambas as anotações (modelo e uropatologista) foram alinhadas e comparadas para avaliar o nível de concordância.

4.3.2 Métricas de avaliação do modelo Mask R-CNN nas lâminas individuais de biópsia de próstata com câncer confirmado pelo patologista

Inicialmente, avaliamos as lâminas individuais dos 21 casos clínicos de pacientes com diagnóstico confirmado de CaP.

A primeira métrica em foco foi o coeficiente de Dice, que atingiu um valor de 0,683. No contexto das lâminas histológicas, este índice indica uma concordância moderada entre as segmentações realizadas pelo modelo e as anotações da uropatologista. Tal valor sugere que, embora o modelo tenha demonstrado uma capacidade razoável de replicar o discernimento humano, ainda há margem para aprimoramento, principalmente quando se trata de discernir nuances nos padrões histológicos.

A análise ROC-AUC revelou uma performance distinta para cada categoria (Figura 23). Para os padrões de Gleason, os valores foram de 0,7629, 0,6408 e 0,7383 para Gleason 3, 4 e 5, respectivamente. Embora esses valores não sejam perfeitos, eles demonstram uma competência satisfatória do modelo em identificar e discriminar os padrões de Gleason, sobretudo quando confrontados com a complexidade histológica do CaP.

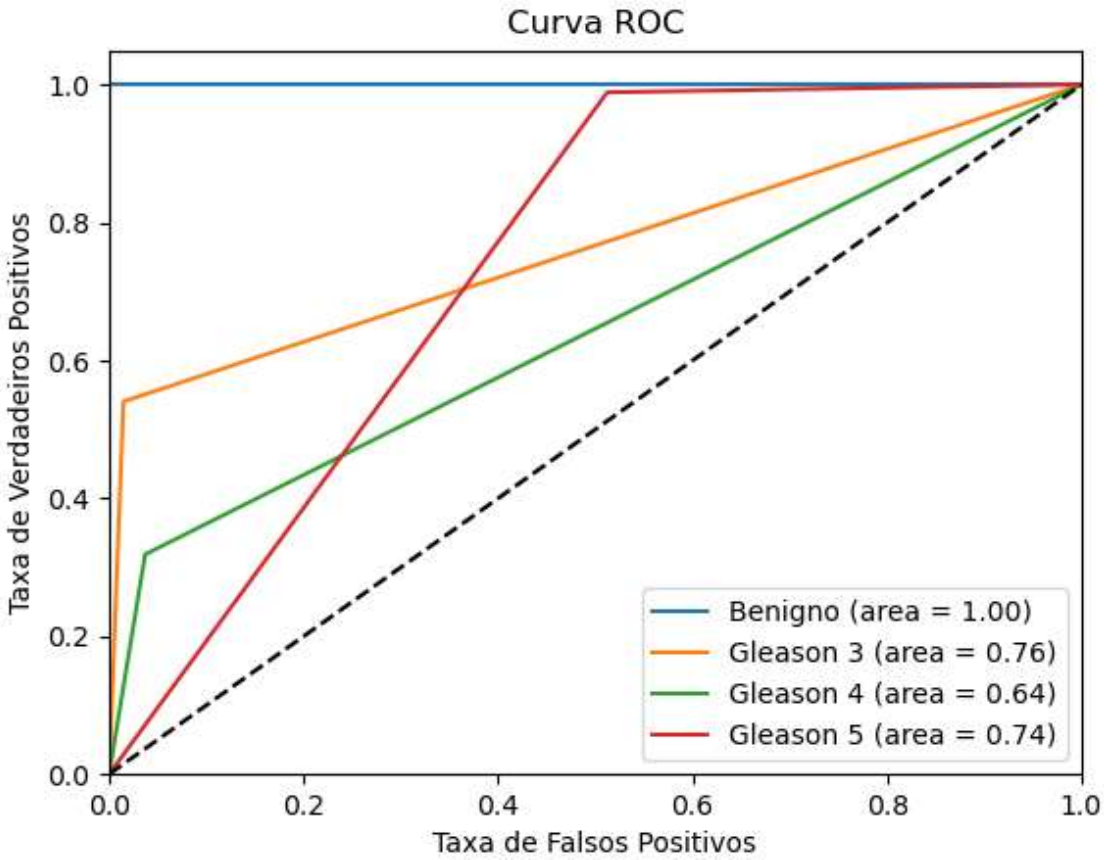

Figura 23. Análise da curva ROC quando aplicado o modelo treinado de Mask R-CNN nas lâminas de biópsia de próstata de pacientes com câncer de próstata.

A matriz de confusão revelou dados valiosos sobre o desempenho real do modelo em cada categoria. Por exemplo, na categoria Gleason 4, houve uma proporção significativa de falsos negativos (42) em comparação com os verdadeiros positivos (21). Isso pode explicar a menor ROC-AUC observada para esta categoria, indicando que o modelo, por vezes, confunde o padrão Gleason 4 com outros padrões (Figura 24).

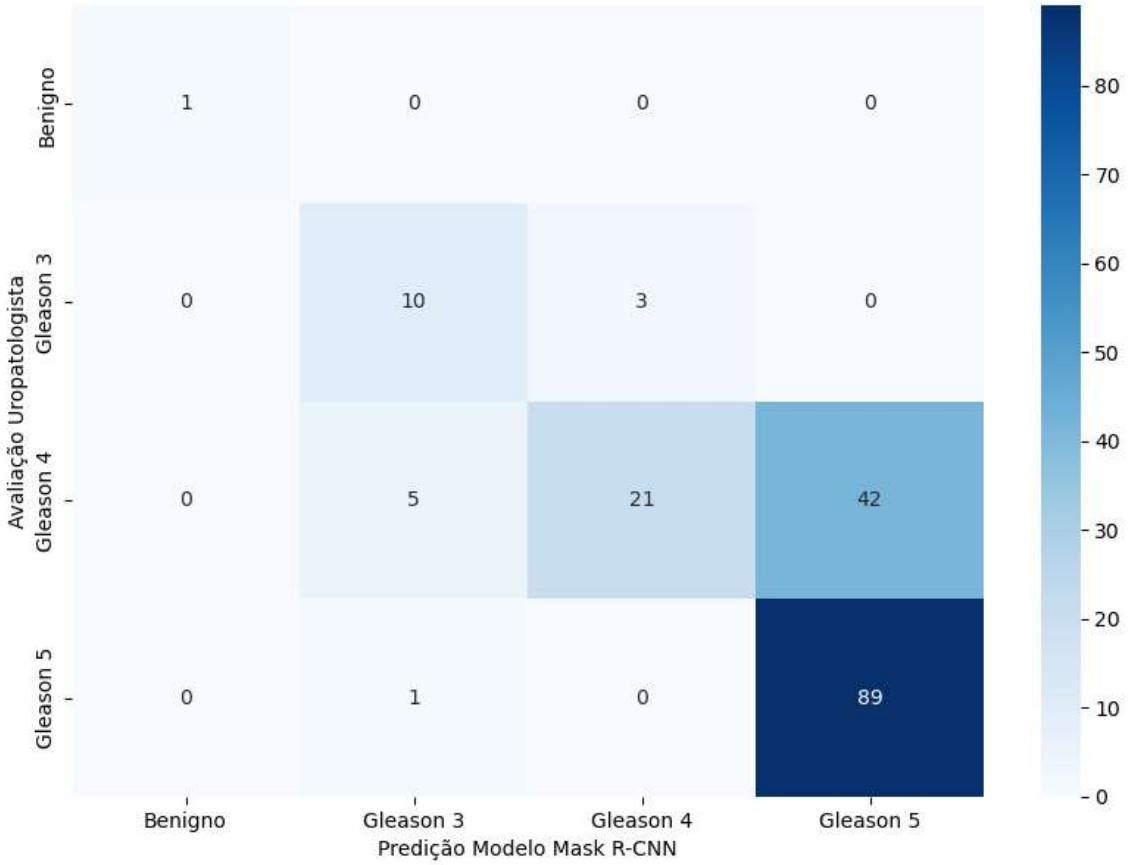

Figura 24. Matriz de confusão do modelo treinado de Mask R-CNN nas lâminas de biópsia de próstata de pacientes com câncer de próstata.

A acurácia geral do modelo foi de 0,7035, demonstrando que o modelo acertou cerca de 70,3% de suas predições em relação ao total de anotações da uropatologista.

Examinando a precisão, sensibilidade e especificidade para cada categoria:

- Gleason 3: Precisão: 0,62, Sensibilidade: 0,77, Especificidade: 0,96
- Gleason 4: Precisão: 0,88, Sensibilidade: 0,31, Especificidade: 0,97
- Gleason 5: Precisão: 0,68, Sensibilidade: 0,99, Especificidade: 0,49

4.3.3 Métricas de avaliação do modelo Mask R-CNN para cada paciente de biópsia de próstata com câncer confirmado pelo patologista

Após avaliar cada lâmina individualmente, realizamos a análise do caso completo de cada paciente utilizando não somente o padrão de Gleason visto em cada lâmina, mas sim, o escore final de Gleason para determinado paciente dado tanto pelo modelo de Mask R-CNN como pelo expert em uropatologia (Figura 25).

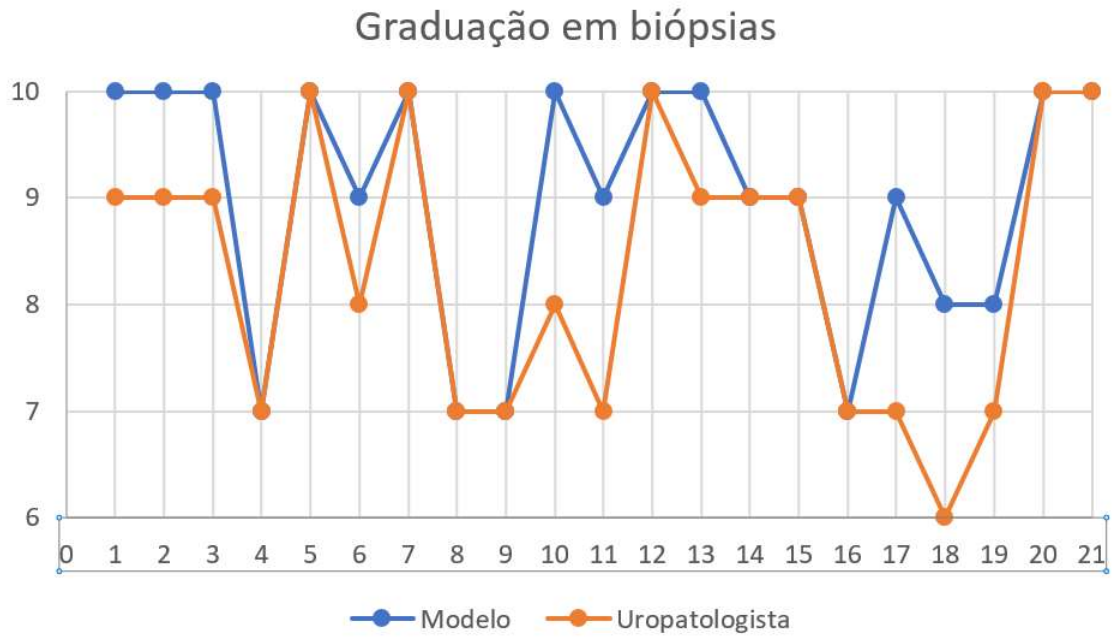

Figura 25. Escore de Gleason dos pacientes da biópsia de próstata. Em azul, predição do modelo de IA. Em laranja, análise do expert em uropatologia.

Para avaliar a concordância entre o escore de Gleason do modelo e do uropatologista, utilizamos o Coeficiente Kappa Quadrático. O valor obtido foi 0,69. Esse valor sugere uma concordância substancial entre o modelo Mask R-CNN e a avaliação do especialista em uropatologia na determinação dos escores de Gleason. Essa métrica, ao refletir tal concordância, indica que o modelo demonstrou eficácia razoável na categorização dos escores de Gleason, alinhando-se quase em paridade com a capacidade de diagnóstico de um

especialista em uropatologia. Entretanto, é primordial entender que a concordância substancial não é sinônimo de concordância perfeita.

Em relação à matriz de confusão (Figura 26), observamos uma concordância perfeita para os casos classificados como Gleason 10, indicado pelo valor 5 na matriz. No entanto, categorias como Gleason 7 e Gleason 9 apresentaram divergências, evidenciadas pelos valores fora da diagonal principal.

A acurácia obtida foi de 0,52, indicando que, em mais da metade dos casos, o modelo fez previsões alinhadas à avaliação do uropatologista. No entanto, no contexto de diagnóstico de biópsias de próstata, uma acurácia de 52% pode ser vista com ressalvas, pois os riscos associados a falsos positivos ou falsos negativos são significativos.

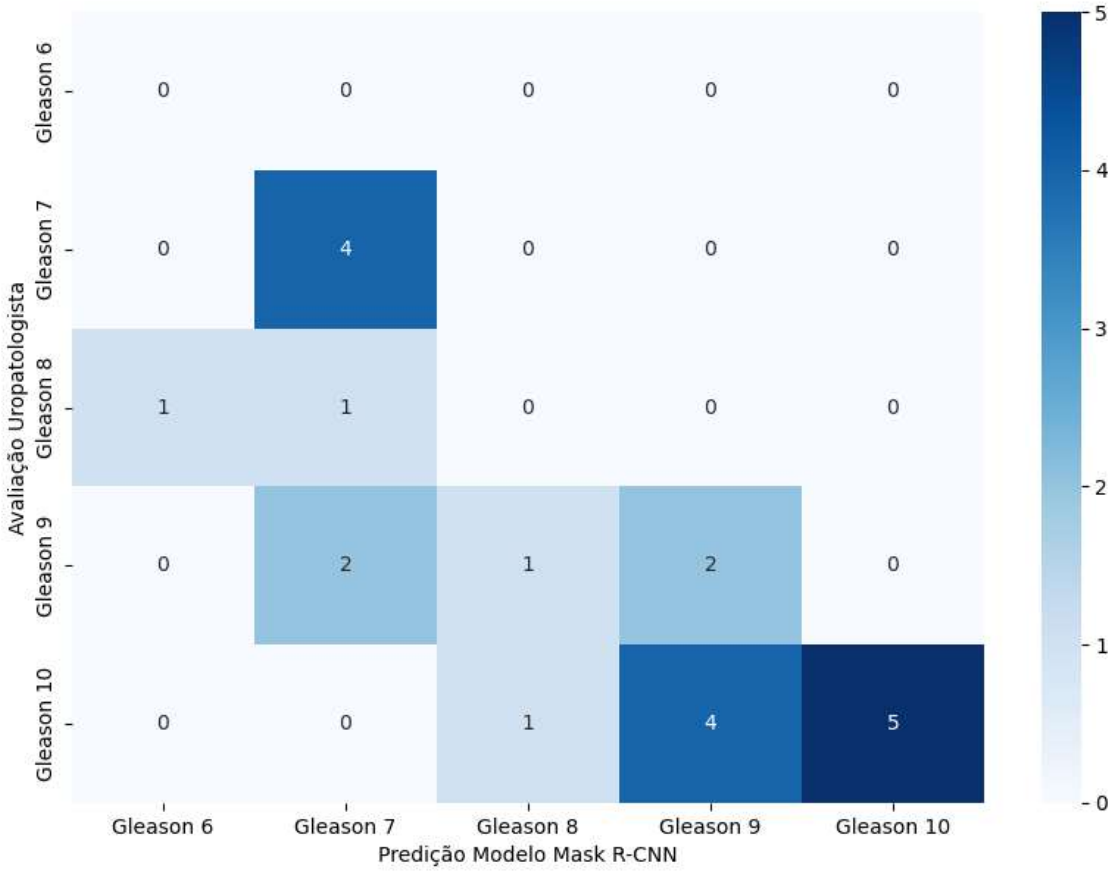

Figura 26. Matriz de confusão dos pacientes submetidos à biópsia de próstata com câncer confirmado.

4.3.4 Métricas de avaliação do modelo nas lâminas benignas de biópsia de próstata

Na análise das lâminas benignas, a precisão do modelo Mask R-CNN em distinguir tecidos benignos dos malignos, bem como dos diferentes padrões de Gleason, foi cuidadosamente avaliada.

O Coeficiente de Dice atingiu o valor de 0,9583 para as lâminas benignas. Essa métrica, notavelmente elevada em comparação com os valores observados para lâminas com tecido maligno, denota uma excelente concordância entre as segmentações feitas pelo modelo e as anotações da uropatologista em tecidos benignos.

Na matriz de confusão das lâminas benignas (Figura 27), pode-se observar que, das 120 amostras, 115 foram corretamente classificadas pelo modelo como benignas, enquanto 5 foram erroneamente classificadas como possuindo padrão de Gleason 5. Esta informação corrobora que, embora o modelo tenha demonstrado uma habilidade excepcional em identificar corretamente os tecidos benignos, existe uma pequena porcentagem de lâminas benignas que foram mal interpretadas.

4.3.5 Métricas de avaliação do modelo Mask R-CNN em todas as lâminas de biópsia de próstata

Por fim, analisamos todas as 292 lâminas selecionadas de biópsia de próstata. Ao consolidar os dados de todas as lâminas de biópsia de próstata, tanto benignas quanto malignas, buscou-se uma visão holística do desempenho do modelo Mask R-CNN.

O coeficiente de Dice global foi de 0,7962, indicando uma boa concordância entre as anotações manuais da uropatologista e as segmentações feitas pelo modelo.

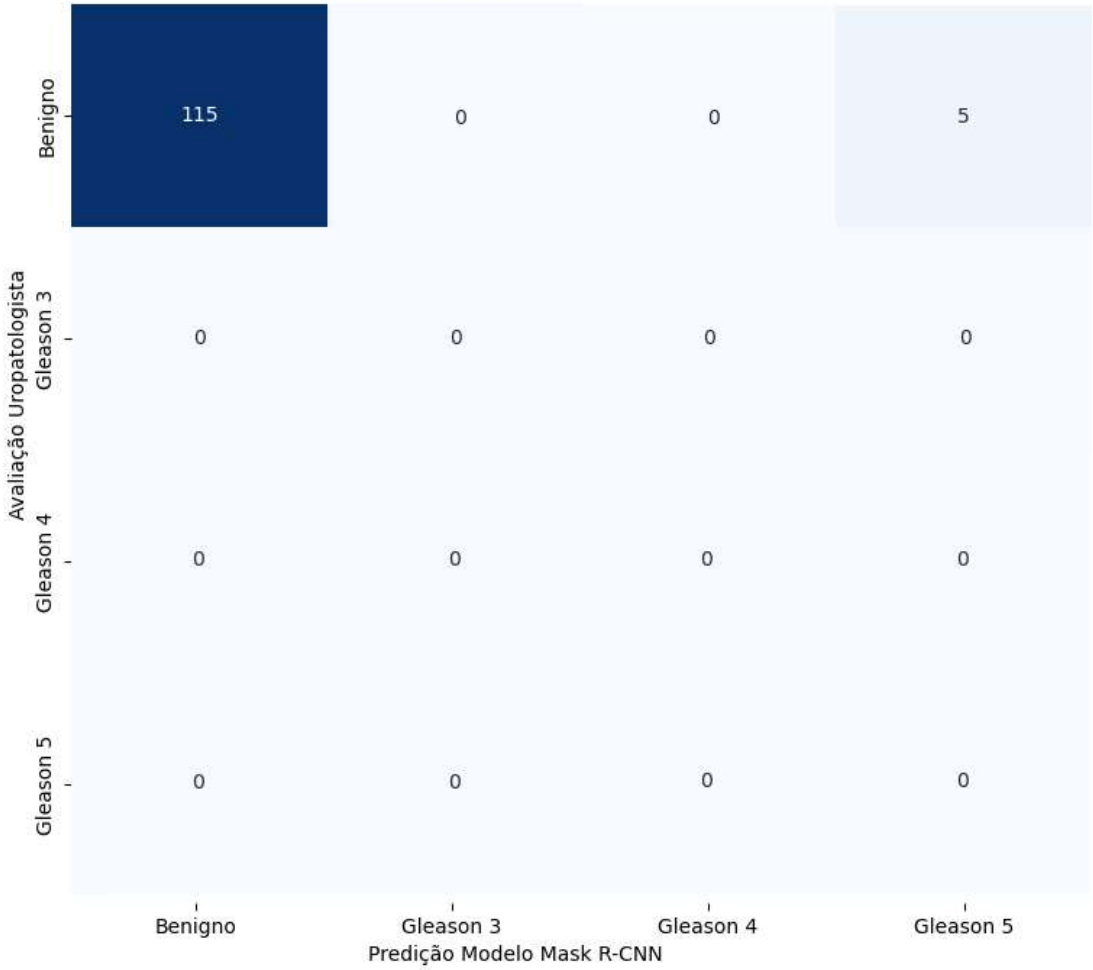

Figura 27. Matriz de confusão dos pacientes submetidos à biópsia de próstata com somente tecido benigno presente.

Na matriz de confusão envolvendo todas as biópsias (Figura 28), a categoria de tecidos benignos teve uma excelente performance com 116 classificações corretas e apenas 5 confundidos com Gleason 5. Para o padrão de Gleason 3, 10 lâminas foram corretamente classificadas, mas 3 foram confundidas com Gleason 4. No padrão de Gleason 4, 21 foram corretamente identificadas, no entanto, 42 foram classificadas equivocadamente como Gleason 5 e 5 como Gleason 3. Finalmente, para o padrão de Gleason 5, 89 lâminas foram corretamente classificadas, enquanto apenas 1 foi confundida com Gleason 3.

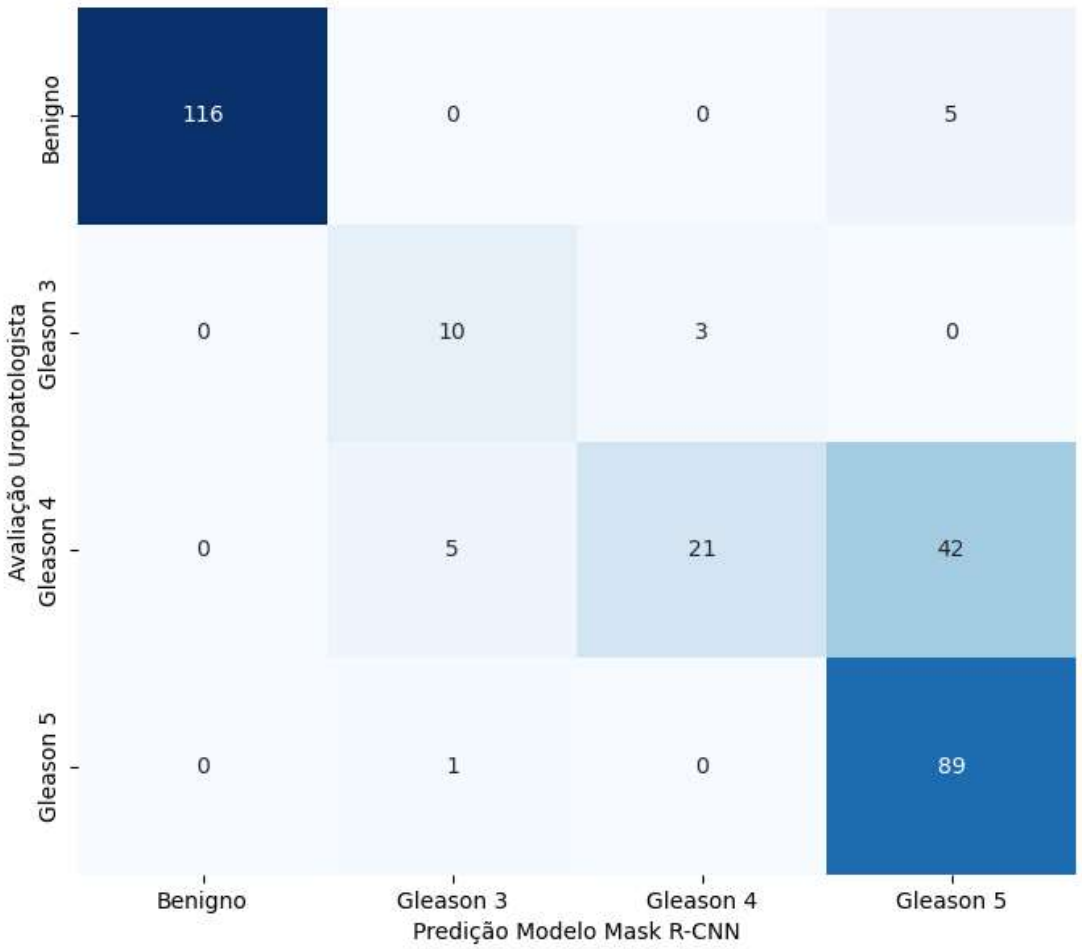

Figura 28. Matriz de confusão envolvendo todas as lâminas de biópsia de próstata avaliadas pelo modelo Mask R-CNN

Os valores de ROC-AUC (Figura 29) para cada categoria foram:

- Benigno: 0,9793
- Gleason 3: 0,7663
- Gleason 4: 0,6519
- Gleason 5: 0,8781

Esses valores confirmam a boa capacidade do modelo em diferenciar tecidos benignos dos malignos e, em menor grau, entre os diferentes padrões de Gleason.

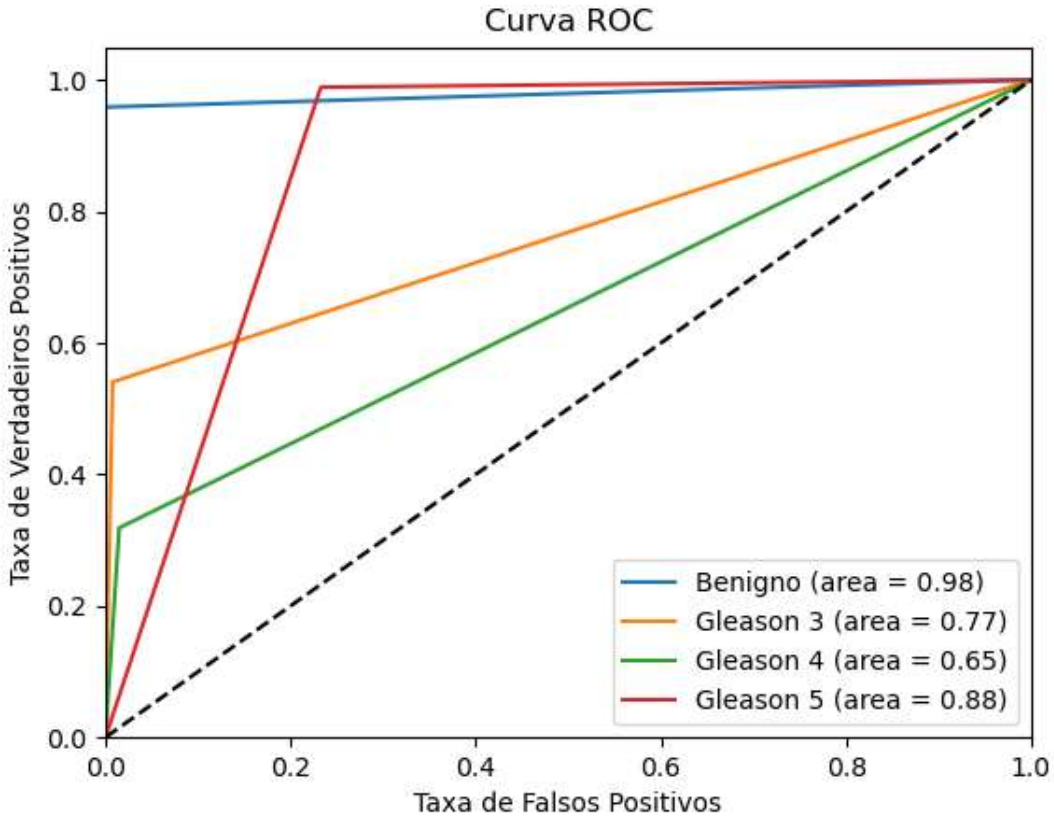

Figura 29. Análise da curva ROC quando aplicado o modelo treinado de Mask R-CNN em todas as lâminas de biópsia de próstata

A acurácia total foi de 0,8082, indicando que o modelo fez predições corretas em aproximadamente 80,8% das vezes, quando consideradas todas as lâminas.

Outras métricas por categoria:

- Benigno: Precisão: 1,00, Sensibilidade: 0,96, Especificidade: 1,00
- Gleason 3: Precisão: 0,62, Sensibilidade: 0,77, Especificidade: 0,98
- Gleason 4: Precisão: 0,88, Sensibilidade: 0,31, Especificidade: 0,99
- Gleason 5: Precisão: 0,65, Sensibilidade: 0,99, Especificidade: 0,77

5. DISCUSSÃO

A prática padrão no diagnóstico do CaP envolve a análise de lâminas histopatológicas, seja de biópsias de próstata ou de prostatectomias radicais, que tradicionalmente é executada manualmente por patologistas por meio de microscópios ópticos. Entretanto, com os avanços tecnológicos recentes, temos observado uma transição revolucionária na forma como as análises são conduzidas, particularmente com a integração de técnicas de IA no campo da patologia.

Nos últimos anos, a emergência das tecnologias relacionadas ao processamento de imagens impulsionou o desenvolvimento de técnicas de IA dedicadas a melhorar a precisão e eficiência dos diagnósticos patológicos, visando dar suporte ao trabalho dos patologistas, facilitar e complementar o trabalho desses profissionais (105). Em nossa pesquisa, o foco principal foi o emprego da IA para diferenciar imagens histopatológicas de tecido prostático entre tecido benigno e tecido contendo CaP. Para alcançar este objetivo, utilizamos dois modelos distintos de aprendizado profundo: a VGGNet e a Mask R-CNN.

Os resultados obtidos em nosso estudo demonstram que a Mask R-CNN superou notavelmente a VGGNet em termos de acurácia. Esta superioridade pode ser, em grande medida, atribuída à segmentação de instâncias realizada pelo modelo Mask R-CNN, que permite uma análise pormenorizada das características celulares e arquitetônicas nas imagens. Esse nível de detalhe é essencial, especialmente quando os patologistas buscam identificar nuances sutis que diferenciam tecidos benignos de malignos (106).

Diversos estudos têm sinalizado que a implementação da IA no diagnóstico do CaP oferece uma precisão diagnóstica significativamente elevada, e nosso estudo reforça essa tendência (107). No entanto, durante a análise dos resultados, percebemos que a acurácia do algoritmo de aprendizado profundo variou conforme o padrão de Gleason presente no tecido prostático analisado. Identificamos um desafio particular para os algoritmos ao enfrentar padrões de Gleason mais elevados, indicando que refinamentos no modelo e a inclusão de mais características discriminatórias são áreas potenciais para pesquisa futura (108).

Quando comparamos a performance dos modelos de aprendizado profundo à avaliação conduzida por um uropatologista expert, descobrimos que a IA demonstrou uma forte correlação com as avaliações humanas. Entretanto, ainda identificamos margem para melhorias (109). A combinação da IA com a avaliação humana pode, de fato, melhorar a acurácia diagnóstica e proporcionar uma maior confiabilidade no diagnóstico do câncer de próstata (67). Além disso, a IA pode auxiliar na padronização e redução do tempo necessário para a análise das lâminas histopatológicas, contribuindo para um processo diagnóstico mais eficiente e confiável (110).

5.1 Análise do desempenho do modelo de classificação categórica (VGGNet)

No decorrer do treinamento com 2.020 imagens, o modelo alcançou uma acurácia impressionante de 97,57%. Essa eficiência elevada em conjuntos de treinamento é uma característica conhecida da VGGNet, dado o seu desenho profundo e a capacidade de extração de características (96). Contudo, ao avaliar o conjunto de teste, a acurácia caiu para 45%, sugerindo um caso clássico de sobreajuste (*overfitting*) (54).

A matriz de confusão também revelou dificuldades no modelo em distinguir corretamente entre classes. Em particular, o modelo teve dificuldades em classificar o padrão de Gleason 5, apresentando precisão e sensibilidade menores. A literatura já destacou anteriormente que o padrão de Gleason 5 pode ser mais desafiador de ser identificado corretamente, devido à sua heterogeneidade (111).

O desempenho reduzido no conjunto de teste pode ter também outras causas. A representação desequilibrada das classes no conjunto de treinamento, embora aqui tenham sido cuidadosamente balanceadas, pode não refletir a distribuição real em uma configuração clínica.

Outra possível razão para a baixa acurácia é a complexidade das imagens histopatológicas, que apresentam variações significativas em termos de morfologia, cor e textura (112). Essa complexidade pode dificultar a identificação

das características relevantes para a classificação do tecido prostático, o que pode resultar em uma acurácia reduzida no conjunto de teste (113).

Ao analisar as métricas de precisão, sensibilidade e especificidade entre as diferentes classes, notamos que o modelo VGGNet enfrentou dificuldades na classificação de alguns padrões de Gleason. A sensibilidade e a especificidade na classificação do tecido benigno foram razoavelmente satisfatórias, sugerindo que o modelo foi capaz de identificar corretamente casos benignos e excluir os casos malignos. No entanto, a precisão na classificação dos diferentes padrões de Gleason foi inferior à desejada, indicando que o modelo teve dificuldade em diferenciar esses padrões com precisão.

Essa dificuldade pode ser atribuída a diversos fatores. A VGGNet, apesar de sua profundidade e capacidade de modelagem, pode não ser o modelo ideal para este tipo específico de dados (27). Variações da VGGNet ou outras arquiteturas podem eventualmente oferecer melhor desempenho (114). Além disso, a quantidade de dados disponíveis para o treinamento do modelo pode ter sido insuficiente para capturar adequadamente as variações nos padrões de Gleason, o que pode ter resultado em uma acurácia reduzida na classificação desses padrões (115).

Comparando com a literatura atual, o desempenho do modelo VGGNet neste estudo está aquém do relatado em pesquisas semelhantes (76).

Nir et al. (109) avaliou 231 pacientes que foram submetidos a prostatectomia radical entre 1997 e 2011. As lâminas de prostatectomia radical foram digitalizadas e anotadas por 6 patologistas para 4 classes (benigno e padrões de Gleason 3, 4 e 5). Essas imagens foram divididas em múltiplos pequenos patches como realizamos em nosso trabalho. O modelo treinado era capaz de classificar cada imagem em uma das classes. A acurácia obtida nesse trabalho na classificação dos patches analisados entre câncer de próstata e patches benignos foi de 97,8%.

Nagpal et al. (116) usou uma estratégia semelhante, mas não obteve resultados tão expressivamente positivos: utilizou lâminas de prostatectomia radical digitalizadas e divididas em pequenos patches para avaliar a capacidade do sistema de aprendizagem profunda de detectar e graduar o câncer de próstata, porém, com um número muito robusto de imagens. Foram utilizados

112 milhões de imagens anotadas por patologistas de 1226 lâminas de prostatectomia radical para treinamento e 331 lâminas independentes para validação. O modelo de aprendizagem profunda atingiu uma acurácia diagnóstica de 0,70 (p = 0,002).

Nosso modelo de classificação categórica não obteve resultados tão animadores quanto o desses 2 trabalhos. Um dos fatores provavelmente foi o número reduzido de imagens que utilizamos no treinamento. Os trabalhos citados anteriormente utilizaram um número de imagens mais de 10 mil vezes maior do que o utilizado no nosso trabalho. Apesar de Nir (109) ter obtido resultados surpreendentes com acurácia maior que 97%, Nagpal (116) obteve 70% mesmo usando dezenas de milhões de imagens.

Com base na análise do desempenho do modelo VGGNet em nosso estudo, algumas melhorias podem ser propostas. Primeiramente, o uso de outras arquiteturas de rede neural, como a ResNet ou a Inception, pode ser considerado, uma vez que essas arquiteturas demonstraram resultados promissores em tarefas semelhantes (117). Além disso, o emprego de técnicas avançadas de pré-processamento de dados e aumento de dados (*data augmentation*) pode ajudar a melhorar a capacidade do modelo de capturar as variações nos padrões de Gleason e, consequentemente, aumentar a acurácia na classificação desses padrões (118). Por fim, a utilização de abordagens de aprendizado de transferência (*transfer learning*) pode ser uma estratégia eficiente para melhorar o desempenho do modelo, permitindo que ele se beneficie do conhecimento adquirido em outras tarefas relacionadas (119).

5.2 Análise do desempenho do modelo de segmentação de instâncias (Mask R-CNN)

O modelo Mask R-CNN teve um desempenho notável no treinamento e validação, com uma acurácia de 95,1% e 93,2%, respectivamente. Além disso, as métricas de perda apresentadas indicam uma aprendizagem eficaz do modelo em diferentes aspectos da segmentação, desde a classificação de objetos até a determinação exata de suas bordas.

No conjunto teste, o modelo Mask R-CNN alcançou uma acurácia global de aproximadamente 89% no conjunto de teste, um resultado expressivo dada a limitação no número de imagens usadas, em comparação com padrões usuais de aprendizado profundo.

A avaliação usando o coeficiente de Dice, que mede a concordância entre duas amostras, revelou um valor de 0,89, corroborando a eficácia do modelo. Além disso, a área sob a curva para as diferentes categorias reflete um alto grau de especificidade e sensibilidade nas predições do modelo. Particularmente notável é o desempenho na identificação de padrões de Gleason 4, com uma AUC de 0,9595, demonstrando capacidade do modelo em diferenciar casos mais sutis e clinicamente relevantes.

O treinamento das imagens para criação desse tipo de modelo exige anotações acuradas e precisas em todas as imagens utilizadas, milhares de anotações manuais, que é um processo que consome muito tempo, estressante, e exige as habilidades de um profissional experiente, pois anotações de baixa qualidade e pouco precisas tornam o modelo pouco confiável. Uma base de dados com grande número de casos e de alta qualidade com boas anotações tem potencial ilimitado com as técnicas de aprendizado profunda.

A alta acurácia alcançada pela Mask R-CNN no conjunto de teste pode ser atribuída a vários fatores. Primeiramente, a arquitetura da Mask R-CNN é baseada na Faster R-CNN (94), que foi projetada para ser eficiente em tarefas de detecção e segmentação de objetos. A Mask R-CNN estende essa arquitetura adicionando uma cabeça de segmentação, permitindo a realização da segmentação de instâncias de maneira eficaz (90). Essa abordagem baseada em segmentação pode ser mais adequada para análise histopatológica, pois fornece informações espaciais mais detalhadas sobre a localização e a morfologia das células tumorais (27).

Além disso, o pré-processamento de dados e o aumento de dados (*data augmentation*) foram cruciais para melhorar o desempenho do modelo. O uso de técnicas de aumento de dados, como rotação, inversão e zoom, aumentou a diversidade de exemplos de treinamento e ajudou a aprendizagem do modelo (101). O balanceamento de classes também foi fundamental para evitar o viés

em direção às classes mais prevalentes, garantindo uma aprendizagem adequada das características associadas a cada classe (120).

A comparação das métricas de precisão, sensibilidade e especificidade entre as diferentes classes revelou que a Mask R-CNN apresentou um bom desempenho na maioria das classes de Gleason. No entanto, o modelo enfrentou dificuldades na classificação do câncer de próstata Gleason 5, que é caracterizado por um padrão de crescimento infiltrativo e indiferenciado das células tumorais (20). Essa dificuldade pode ser explicada pela maior variabilidade morfológica e pela ausência de glândulas bem formadas neste padrão, tornando a distinção entre tecido benigno e maligno mais desafiadora (67).

Outro fator que pode ter contribuído para a dificuldade na classificação do Gleason 5 é a presença de artefatos e ruído nas imagens histopatológicas, como descoloração e dobramento do tecido, que podem afetar a qualidade das características extraídas pelo modelo (121). Além disso, a presença de infiltrado inflamatório nas imagens pode aumentar a complexidade da tarefa e confundir o modelo durante a classificação (115). A maioria dos estudos na área ainda enfrenta dificuldades na classificação de padrões de Gleason mais agressivos e complexos, como o Gleason 5 (122).

Nossos resultados são consistentes com a literatura atual e estudos semelhantes que investigaram o uso de modelos de segmentação de instâncias baseados em aprendizado profundo para análise histopatológica (110).

A Mask R-CNN tem sido amplamente utilizada em estudos que envolvem a análise de imagens histopatológicas, apresentando resultados promissores na detecção e segmentação de células tumorais. Por exemplo, Ilse et al. (123) aplicaram a Mask R-CNN para segmentação de células em imagens histopatológicas de câncer de mama, obtendo resultados superiores em comparação com métodos tradicionais de segmentação. Usando a mesma técnica de Mask-RCNN, Couteaux et al. (124) obteve uma acurácia de 90,6% na detecção automática de lacerações de menisco no joelho, demonstrando que a efetividade dessa técnica e a aplicabilidade em qualquer área da medicina.

5.3 Comparação entre os modelos VGGNet e Mask R-CNN

Os resultados obtidos nos experimentos realizados nesta tese revelaram diferenças significativas no desempenho dos modelos VGGNet e Mask R-CNN na classificação e segmentação de imagens histopatológicas de tecido prostático.

Nossos resultados apresentados anteriormente mostram que o modelo VGGNet obteve uma acurácia de 45% no conjunto de teste, enquanto o modelo Mask R-CNN alcançou uma acurácia geral de 0,89. Estes resultados demonstram que Mask R-CNN superou VGGNet em termos de desempenho geral, como evidenciado por suas métricas superiores tanto no conjunto de treinamento quanto no de testes.

O Mask R-CNN mostrou uma capacidade superior de distinguir entre as diferentes categorias, particularmente nas classes malignas, quando comparado ao VGGNet. Embora o VGGNet tenha se saído bem no treinamento e na validação, o seu desempenho no conjunto de teste foi abaixo do esperado. Essa discrepância pode indicar um possível sobreajuste durante o treinamento. Por outro lado, Mask R-CNN mostrou uma performance consistente entre treinamento, validação e teste.

Enquanto VGGNet se concentra na classificação categórica de imagens, Mask R-CNN fornece uma segmentação de instâncias mais detalhada. Isso faz com que o Mask R-CNN seja especialmente valioso para análises morfológicas detalhadas em patologia, permitindo que os patologistas identifiquem e analisem características específicas das lesões (90, 96).

Uma das limitações da VGGNet no contexto do diagnóstico de CaP é a sua dependência de classificações categóricas, o que pode limitar a capacidade do modelo de capturar a complexidade e a diversidade dos padrões de Gleason e outros aspectos morfológicos importantes para o diagnóstico preciso do CaP. Além disso, a VGGNet pode ser sensível à variação na coloração das lâminas histológicas, o que pode afetar negativamente o desempenho do modelo (121). Em contraste, a Mask R-CNN é capaz de segmentar e identificar características individuais dentro das imagens, permitindo uma análise mais detalhada dos

tecidos prostáticos e, assim, alcançando uma maior acurácia na identificação do CaP e na atribuição dos padrões de Gleason (115).

No entanto, ao refinar nossa análise, percebemos que a Mask R-CNN também não é isenta de desafios. Por exemplo, ela demonstrou dificuldades na classificação do padrão Gleason 5. Esta constatação sugere que, apesar de sua capacidade de segmentação de instâncias ser um grande trunfo, ainda há espaço para otimização quando confrontada com padrões histológicos mais intricados ou atípicos.

5.4 Análise do desempenho do modelo nas lâminas de biópsia de próstata

O modelo Mask R-CNN obteve desempenho notável no conjunto de treinamento, validação e teste nas lâminas de prostatectomia radical. Porém, quando o modelo treinado previamente foi aplicado nas lâminas de biópsia de próstata, o coeficiente de Dice foi de 0,683 e a acurácia geral de 70,3%. Estes valores indicam uma boa, mas não excelente, concordância com as anotações da uropatologista. Os valores ROC-AUC para os padrões de Gleason foram variáveis, oscilando entre 0,6408 para Gleason 4 e 0,7629 para Gleason 3.

Na análise das lâminas benignas, o modelo demonstrou um alto coeficiente de Dice de 0,9583. Este resultado indica uma concordância quase perfeita na segmentação de tecidos benignos. A habilidade de distinguir precisamente entre tecidos benignos e malignos é crucial para minimizar os diagnósticos falsos positivos.

Modelos de aprendizado profundo, como o Mask R-CNN, têm sido aplicados no diagnóstico do CaP com resultados promissores. No entanto, o desempenho varia consideravelmente dependendo da base de dados, pré-processamento e ajustes do modelo.

Por exemplo, em um estudo realizado por Campanella et al., um modelo de aprendizado profundo atingiu acurácia de 94% na classificação de imagens histológicas da próstata (110). O estudo avaliou a classificação de diferentes tipos de cânceres em imagens histológicas. Resultados para o câncer de

próstata mostraram uma área sob a curva superior a 0,98. O método proposto poderia permitir aos patologistas excluírem 65 - 75% das lâminas mantendo 100% de sensibilidade. Embora nosso modelo tenha apresentado um desempenho notável, a AUC obtida por Campanella et al. demonstra uma eficiência ainda maior na classificação.

Outro estudo notável foi realizado por Marginean (125) que desenvolveu um algoritmo de aprendizado profundo com alta precisão na detecção de áreas cancerígenas (sensibilidade: 100%, especificidade: 68%). O coeficiente de correlação intraclasse para os padrões de Gleason 3 e 4 foram 0,96 e 0,94, respectivamente. Nossa pesquisa demonstrou sensibilidade reduzida (31%) para o padrão Gleason 4, indicando que o modelo de Marginean et al. pode ter uma eficácia superior em algumas métricas de desempenho.

A utilização de técnicas de aprendizado profundo no diagnóstico do CaP está se mostrando uma abordagem promissora. Nosso modelo demonstra potencial, no entanto, os resultados da literatura atual sugerem que ainda há espaço para otimizações. Fazer o treinamento do modelo usando imagens específica de biópsia de próstata seria uma delas. A busca por métodos mais precisos e consistentes para auxiliar os patologistas no diagnóstico e na gradação do câncer de próstata é de extrema importância para o futuro da patologia digital.

5.5 A influência da experiência dos patologistas e o potencial dos modelos de aprendizado profundo

A acurácia na determinação do escore de Gleason é altamente dependente da experiência e especialização do patologista (20). Estudos anteriores mostraram que a concordância entre o grau de Gleason entre patologistas gerais e uropatologistas pode ser relativamente baixa. Por exemplo, Nakai et al. (126) demonstraram que a concordância entre o grau de Gleason atribuído por patologistas gerais e uropatologistas foi de apenas 47,5%. Essa diferença na concordância destaca a complexidade inerente à classificação do

CaP e sugere que a experiência do patologista pode desempenhar um papel crucial na precisão do diagnóstico.

Além disso, o número de patologistas em muitas regiões do mundo é insuficiente para atender às necessidades crescentes, e a complexidade das subespecializações está aumentando, especialmente em países em desenvolvimento como o Brasil (127). Nesse contexto, os modelos de aprendizado profundo têm o potencial de auxiliar patologistas gerais com experiência limitada em áreas específicas, como uropatologia, a obter resultados mais precisos e consistentes.

Nagpal et al. (107) investigaram a aplicação de modelos de aprendizado profundo no diagnóstico do CaP e compararam seu desempenho com o de patologistas gerais e uropatologistas experientes. Os autores observaram que, para tecidos contendo CaP, a taxa de concordância entre o modelo de aprendizado profundo e os uropatologistas na graduação do tumor foi de 71,7% (IC 95%: 67,9%-75,3%), significativamente maior do que a taxa de concordância entre patologistas gerais e uropatologistas (58,0%; IC 95%, 54,5%-61,4%) ($p < 0,001$). No entanto, a concordância entre os especialistas e o modelo na distinção entre tecidos benignos e malignos foi semelhante à concordância entre especialistas e patologistas gerais (94,3% [IC 95%: 92,4%-95,9%] vs. 94,7% [IC 95%, 92,8%-96,3%], respectivamente, $p = 0,58$).

Esses resultados indicam que os modelos de aprendizado profundo podem ter um desempenho equivalente ao de patologistas gerais na diferenciação entre tecidos malignos e benignos e podem até superar o desempenho dos patologistas gerais na graduação do escore de Gleason. Portanto, a implementação de modelos de aprendizado profundo no diagnóstico do CaP pode oferecer uma ferramenta valiosa para melhorar a precisão e a consistência dos resultados, especialmente em cenários onde o acesso a uropatologistas experientes é limitado.

A experiência do patologista é um fator crítico no diagnóstico do CaP (27). A aplicação de modelos de aprendizado profundo no diagnóstico do CaP tem o potencial de ajudar a superar as limitações relacionadas à experiência e especialização dos patologistas, fornecendo uma ferramenta adicional para melhorar a precisão e consistência dos resultados (110). À medida que os

modelos de aprendizado profundo continuam a evoluir e a melhorar, é provável que sua implementação na prática clínica se torne cada vez mais comum, auxiliando patologistas gerais e especialistas a tomarem decisões diagnósticas mais precisas e informadas.

Além disso, a integração dos modelos de aprendizado profundo em sistemas de apoio à decisão clínica pode facilitar a colaboração entre patologistas e aumentar a eficiência do processo de diagnóstico (67). Por exemplo, os modelos podem ser usados para identificar áreas de interesse nas imagens histopatológicas e sugerir graduações preliminares de Gleason, que podem ser posteriormente confirmadas ou ajustadas pelos patologistas com base em sua experiência e conhecimento.

Ainda em relação ao potencial dos modelos de aprendizagem profundo, um estudo conduzido por Eloy et al. (128) avaliou uma ferramenta de IA que já está sendo utilizada no mundo real. O estudo avaliou a ferramenta chamada Paige Prostate, projetada especificamente para auxiliar patologistas na detecção, graduação e quantificação do câncer de próstata. Ao comparar o desempenho diagnóstico de quatro patologistas sem e com a assistência do Paige Prostate, foi observado que, embora a precisão diagnóstica se mantivesse similar em ambas as fases (95,00% na fase 1 e 93,81% na fase 2), houve benefícios claros ao integrar a ferramenta de IA. Os patologistas relataram proliferação atípica de pequenos ácinos (ASAP) cerca de 30% menos vezes na fase com assistência da IA. Notavelmente, a ferramenta resultou em cerca de 20% menos pedidos de estudos de imuno-histoquímica e 40% menos solicitações de segunda opinião. O tempo médio necessário para a leitura e emissão de relatórios por lâmina foi aproximadamente 20% menor quando assistido pelo Paige Prostate.

Importante notar que o estudo do Paige Prostate também revelou que a concordância média total com o desempenho do software foi de aproximadamente 70% dos casos, sendo significativamente mais alto em casos negativos (cerca de 90%) do que em casos de câncer (cerca de 30%). As principais discordâncias diagnósticas ocorreram ao distinguir casos negativos com ASAP de pequenos focos de adenocarcinoma acinar bem diferenciado (menos de 1,5 mm).

5.6 Limitações do estudo e direções futuras

Embora os resultados deste estudo apresentem informações relevantes sobre a aplicação de inteligência artificial e aprendizado profundo no diagnóstico do CaP, é importante reconhecer as limitações e os desafios enfrentados durante o desenvolvimento e a avaliação dos modelos propostos.

Uma das principais limitações deste estudo foi o tamanho do conjunto de dados utilizado. O conjunto de dados é um fator crítico na construção de modelos de aprendizado profundo, uma vez que a qualidade e a quantidade de dados disponíveis podem afetar significativamente o desempenho e a generalização dos modelos treinados (58). Neste estudo, a quantidade limitada de imagens histopatológicas disponíveis pode ter influenciado a capacidade dos modelos de aprender padrões robustos e generalizáveis, o que resultou em um desempenho inferior quando comparado a estudos semelhantes na literatura (129). Para abordar essa limitação, recomenda-se a obtenção de um conjunto de dados maior e mais diversificado em estudos futuros, o que poderia melhorar a capacidade dos modelos de aprender padrões complexos e aumentar a acurácia no diagnóstico do CaP.

Outra limitação deste estudo está relacionada às características das imagens utilizadas. Imagens histopatológicas podem variar em termos de qualidade, cor, contraste e resolução, o que pode tornar o processo de aprendizado mais desafiador para os modelos de aprendizado profundo (115). Além disso, a presença de artefatos e ruídos nas imagens pode dificultar a detecção e classificação das áreas contendo CaP (121). Neste contexto, a utilização de técnicas de pré-processamento de imagem e de aumento de dados pode ser benéfica para melhorar a qualidade das imagens e reduzir os efeitos de variações e artefatos nas imagens (101). Estudos futuros podem explorar diferentes abordagens de pré-processamento de imagem e aumento de dados para melhorar o desempenho dos modelos de aprendizado profundo no diagnóstico do CaP.

Em relação às investigações futuras, é importante considerar a exploração de novas arquiteturas de redes neurais convolucionais e técnicas de

aprendizado profundo que possam proporcionar melhorias no desempenho e na eficiência dos modelos. Por exemplo, a implementação de redes neurais com atenção seletiva (130) ou a utilização de aprendizado de transferência (131) podem ser abordagens promissoras para aprimorar a capacidade dos modelos de aprender padrões discriminativos e aumentar a acurácia no diagnóstico do CaP. Além disso, a integração de informações clínicas e moleculares com os dados de imagem histopatológica pode oferecer uma abordagem mais abrangente e precisa para o diagnóstico e a estratificação de risco do CaP (27).

Outro aspecto a ser investigado em estudos futuros é a análise da correlação entre as predições dos modelos de aprendizado profundo e os resultados clínicos dos pacientes, como sobrevida e resposta ao tratamento. Essa análise pode fornecer informações valiosas sobre o potencial dos modelos de aprendizado profundo para auxiliar na tomada de decisões clínicas e na identificação de pacientes que possam se beneficiar de terapias específicas. Além disso, o estudo da capacidade dos modelos em prever a agressividade do tumor e o potencial de progressão da doença pode contribuir para a personalização do manejo clínico do CaP (67).

Outra direção futura importante é a exploração da aplicação de modelos de aprendizado profundo em outras modalidades de imagem, como imagens de ressonância magnética multiparamétrica e imagens de ultrassonografia. A combinação de diferentes modalidades de imagem pode fornecer informações complementares e aumentar a acurácia no diagnóstico do CaP (132). Além disso, a integração de modelos de aprendizado profundo em sistemas de apoio à decisão clínica pode ajudar a melhorar a eficiência e a qualidade dos diagnósticos realizados por uropatologistas (133).

Por fim, este estudo apresentou uma análise detalhada do desempenho de dois modelos de aprendizado profundo, VGGNet e Mask R-CNN, no diagnóstico do CaP usando imagens histopatológicas. Embora tenham sido observadas limitações e desafios, os resultados obtidos fornecem informações importantes sobre o potencial das técnicas de inteligência artificial e aprendizado profundo na melhoria da acurácia diagnóstica do CaP. Estudos futuros podem se basear nessas descobertas para desenvolver e validar modelos mais eficientes e robustos, explorando novas arquiteturas de rede, técnicas de

aprendizado e abordagens combinadas. A investigação contínua nesta área é crucial para o avanço do conhecimento científico e para a promoção da inovação na detecção e no tratamento do câncer de próstata.

6. CONCLUSÕES

Foram construídos dois sistemas de aprendizado profundo para o diagnóstico do CaP e graduação de Gleason nas lâminas de prostatectomia radical.

O modelo Mask R-CNN foi superior ao VGGNet no diagnóstico do CaP e graduação de Gleason.

A acurácia do modelo Mask R-CNN para o diagnóstico do CaP foi de 89% e a precisão para a determinação dos padrões 3, 4 e 5 de Gleason foi 93%, 96% e 77%, respectivamente.

A maior dificuldade do sistema foi no reconhecimento do padrão 5 de Gleason.

Na era da patologia especializada, o uso da IA é promissor e pode auxiliar o patologista em sua rotina, permitindo uma maior assertividade no diagnóstico e graduação de Gleason, principalmente entre não especialistas.

7. ANEXOS

ANEXO I. PARECER CONSUBSTANCIADO DO CEP

USP - FACULDADE DE MEDICINA DA UNIVERSIDADE DE SÃO PAULO - FMUSP

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: AVALIAÇÃO DA INTELIGÊNCIA ARTIFICIAL NA ACURÁCIA DA GRADUAÇÃO HISTOLÓGICA DO CÂNCER DE PRÓSTATA (GLEASON E ISUP) E COMPARAÇÃO ENTRE A BIÓPSIA E O PRODUTO DE PROSTATECTOMIA RADICAL

Pesquisador: Katia Ramos Moreira Leite

Área Temática:

Versão: 2

CAAE: 16159019.1.0000.0065

Instituição Proponente: Faculdade de Medicina da Universidade de São Paulo

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 3.689.656

Apresentação do Projeto:

A redação do projeto agora ficou b em mais clara. Ele apresenta uma seção “Introdução” muito boa, embora um pouco extensa, sobre o câncer de próstata. Adicionalmente, a seção “Materiais e Métodos” ficou melhor com o acréscimo da descrição do método de Inteligência Artificisl. A
Bibliografia está bem apresentada e faz jus ao trabalho.

Objetivo da Pesquisa:

O objetivo da pesquisa é avaliar a exatidão do método de Aprendizagem Profunda (Deep Learning) na gradação do câncer de próstata e aplicar esse aprendizado para comparar os achados da biópsia préoperatória e do espécime da prostatectomia radical.

Avaliação dos Riscos e Benefícios:

O projeto usará imagens virtuais de blocos de parafina anonimizados e, portanto, não há risco. O benefício óbvio é a possibilidade de melhorar os diagnósticos pré-operatórios através do uso do Aprendizado de Máquina.

Endereço: DOUTOR ARNALDO 251 21º andar sala 36
Bairro: PACAEMBU **CEP:** 01.246-903
UF: SP **Município:** SAO PAULO
Telefone: (11)3893-4401 **E-mail:** cep.fm@usp.br

USP - FACULDADE DE MEDICINA DA UNIVERSIDADE DE SÃO PAULO - FMUSP

Comentários e Considerações sobre a Pesquisa:

A pesquisa é importante, dados os possíveis benefícios no diagnóstico do câncer de próstata. Constitui um bom trabalho de doutorado.

Considerações sobre os Termos de apresentação obrigatória:

Foram apresentados os documentos necessários, especialmente a carta de anuência do laboratório fornecedor das imagens. Foi solicitada dispensa do TCLE em função do uso apenas de imagens virtuais anonimizadas. Parece-me que o pedido deve ser aceito já que o número de casos de câncer de próstata é grande e não há risco de identificação do paciente por nenhum meio.

Conclusões ou Pendências e Lista de Inadequações:

O projeto agora está melhor e pode ser aprovado.

Considerações Finais a critério do CEP:

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_1370865.pdf	06/09/2019 17:27:23		Aceito
Projeto Detalhado / Brochura Investigador	Projeto_Petronio_Melo.docx	06/09/2019 17:26:31	Katia Ramos Moreira Leite	Aceito
Cronograma	Cronograma.docx	06/09/2019 17:25:44	Katia Ramos Moreira Leite	Aceito
Outros	etica_FMUSP.pdf	17/06/2019 10:33:51	Katia Ramos Moreira Leite	Aceito
Folha de Rosto	folha_rosto.pdf	17/06/2019 10:33:08	Katia Ramos Moreira Leite	Aceito
Outros	Carta_Petronio_GENOA.pdf	03/06/2019 09:51:56	Katia Ramos Moreira Leite	Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	dispensa_petronio.pdf	03/06/2019 09:50:18	Katia Ramos Moreira Leite	Aceito

Situação do Parecer:

Endereço: DOUTOR ARNALDO 251 21º andar sala 36
Bairro: PACAEMBU **CEP:** 01.246-903
UF: SP **Município:** SAO PAULO
Telefone: (11)3893-4401 **E-mail:** cep.fm@usp.br

USP - FACULDADE DE MEDICINA DA UNIVERSIDADE DE SÃO PAULO - FMUSP

Aprovado

Necessita Apreciação da CONEP:
Não

SAO PAULO, 07 de Novembro de 2019

Assinado por:
Maria Aparecida Azevedo Koike Folgueira
(Coordenador(a))

Endereço: DOUTOR ARNALDO 251 21º andar sala 36
Bairro: PACAEMBU **CEP:** 01.246-903
UF: SP **Município:** SAO PAULO
Telefone: (11)3893-4401 **E-mail:** cep.fm@usp.br

ANEXO II – PUBLICAÇÃO DE ARTIGO NA REVISTA CLINICS

CLINICS

ORIGINAL ARTICLE

Detecting and grading prostate cancer in radical prostatectomy specimens through deep learning techniques

Petronio Augusto de Souza Melo,* **Carmen Liane Neubarth Estivallet**, **Miguel Srougi**, **William Carlos Nahas**, **Katia Ramos Moreira Leite**

Laboratorio de Pesquisa Medica – LIM55, Divisao de Urologia, Faculdade de Medicina FMUSP, Universidade de Sao Paulo, Sao Paulo, SP, BR.

Melo PAS, Estivallet CLN, Srougi M, Nahas WC, Leite KRM. Detecting and grading prostate cancer in radical prostatectomy specimens through deep learning techniques. Clinics (Sao Paulo). 2021;76:e3198

*Corresponding author. E-mail: petronio_augusto@hotmail.com

OBJECTIVES: This study aims to evaluate the ability of deep learning algorithms to detect and grade prostate cancer (PCa) in radical prostatectomy specimens.

METHODS: We selected 12 whole-slide images of radical prostatectomy specimens. These images were divided into patches, and then, analyzed and annotated. The annotated areas were categorized as follows: stroma, normal glands, and Gleason patterns 3, 4, and 5. Two analyses were performed: i) a categorical image classification method that labels each image as benign or as Gleason 3, Gleason 4, or Gleason 5, and ii) a scanning method in which distinct areas representative of benign and different Gleason patterns are delineated and labeled separately by a pathologist. The Inception v3 Convolutional Neural Network architecture was used in categorical model training, and a Mask Region-based Convolutional Neural Network was used to train the scanning method. After training, we selected three new whole-slide images that were not used during the training to evaluate the model as our test dataset. The analysis results of the images using deep learning algorithms were compared with those obtained by the pathologists.

RESULTS: In the categorical classification method, the trained model obtained a validation accuracy of 94.1% during training; however, the concordance with our expert uropathologists in the test dataset was only 44%. With the image-scanning method, our model demonstrated a validation accuracy of 91.2%. When the test images were used, the concordance between the deep learning method and uropathologists was 89%.

CONCLUSION: Deep learning algorithms have a high potential for use in the diagnosis and grading of PCa. Scanning methods are likely to be superior to simple classification methods.

KEYWORDS: Prostate Cancer; Deep Learning; Radical Prostatectomy; Prostate Pathology; Artificial Intelligence.

INTRODUCTION

The high prevalence and complex management of prostate cancer (PCa) have imposed significant amounts of investment in healthcare systems (1,2). The wide spectrum of aggressiveness of PCa, ranging from an indolent disease that can be managed with surveillance to an aggressive disease with a poor prognosis, necessitates accurate diagnosis and classification. Tumor grading using the Gleason/ISUP score is the main prognostic factor, and together with staging, indicates the choice of treatment and probable outcome (3,4). Histological analysis and Gleason/ISUP grading are currently conducted subjectively by pathologists, and although there are many initiatives aiming to train as many specialists as possible, the number of pathologists is insufficient for dealing with the increasing number and complexity of the actual requirements (5–7).

Among expert uropathologists, the disagreement in determination based on the Gleason score reaches up to 12%; however, this number increases to 50% when considering generalist pathologists (8,9).

In the last few years, the field of knowledge on artificial intelligence has rapidly increased. Machine learning has become prevalent, and is present in many high-tech products, including web search results, speech recognition in smartphones, and video recommendations, among other tasks.

In 2006, Hinton et al. (10) described how to train a machine that is capable of recognizing handwritten digits with high precision (>98%), an approach they called "deep learning."

Deep learning is a branch of artificial intelligence that processes data and creates patterns for use in decision-making (11). In recent years, researchers have tried to solve the problem of PCa diagnosis and grading using deep

No potential conflict of interest was reported.

Received for publication on May 29, 2021. **Accepted for publication on** September 21, 2021

DOI: 10.6061/clinics/2021/e3198

learning techniques to overcome the current limitations of human-made diagnoses (12–14).

In this study, we used prostatectomy specimens evaluated by experienced uropathologists to train a deep learning algorithm in the detection and grading of PCa.

MATERIALS AND METHODS

Study population, slide digitization, and annotation

We randomly selected 12 whole-slide images of hematoxylin-and-eosin-stained formalin-fixed paraffin-embedded prostatectomy specimens from our database slides. Each analyzed slide belonged to a different patient. These slides were digitized at a magnification of 20× using a Pannoramic Flash II 250 scanner (3DHISTECH Ltd., Budapest, Hungary). The whole slides were segmented into 1,525 image patches with a pixel resolution of 2,000 × 2,000 using Python 3 (https://www.python.org). These image patches were then analyzed and annotated by two experienced uropathologists (K.R.M.L. and C.L.N.E.). The annotations were initially conducted separately by the pathologists, and all images were shown. When a disagreement occurred in any image, the pathologists discussed the particular image and reached a consensus. Two different analyses and annotations were employed.

- Categorical image classification method: We labeled each image according to the presence or absence of malignancy. Within the cancer images, each image was labeled according to the most prevalent Gleason pattern present on the slide. As the output, each predicted image was classified into one of four patterns: benign, Gleason 3, Gleason 4, or Gleason 5.
- Image scanning method: Using this method, we delineated and annotated specific areas in each image, rather than simply classifying the entire image with a single label. To accomplish this task, we used the co-annotator tool (https://github.com/jsbroks/coco-annotator/). Each annotation belonged to one of five categories: stroma, normal glands, or Gleason pattern 3, 4, or 5.

Development of deep learning algorithm

In the categorical classification method, we used the Inception v3 Convolutional Neural Network Architecture (https://github.com/machine-learning/Inception-v3-tensorflow) and TensorFlow library (https://www.tensorflow.org) to train the model. The image patches were divided into training and validation datasets. The training dataset is an actual dataset used to train the model. The model observes and learns from these data. Meanwhile, the validation dataset is the sample of data used for frequent evaluations of the model, the hyperparameters of which are turned. The model sees the validation dataset, but never learns from it. Because robust datasets are required for adequate network training, we applied data augmentation on all image patches of our training data: horizontal and vertical flipping, rotation, and zooming.

With the scanning method, the image patches were divided into training and validation data. The model was trained using the Mask Region-based Convolutional Neural Network (Mask R-CNN) (https://github.com/matterport/Mask_RCNN), where the model learns from the delineated areas annotated by the pathologists and generates its own bounding boxes and segmentation masks for each instance of an object in the image.

Model evaluation

After the model training, we selected three new whole-slide images that were not used in the training, to evaluate the generalization capability of the model. We prepared these images in the same way as with the training images, *i.e.*, using image patches with a pixel resolution of 2000 × 2000. From them, we randomly chose 100 different image patches for each classification method. All images were evaluated using deep learning algorithms after being read by the uropathologists, and the concordance between the two results was analyzed.

Ethics

The study was approved by the Institutional Review Board and Ethics Committee, and informed consent was considered unnecessary.

RESULTS

Using the categorical classification method, 740 images in the benign group and 785 images in the cancer group (251 for Gleason 3, 254 for Gleason 4, and 280 for Gleason 5) were categorized. The images were randomly separated into training (1,220 images) and validation (305 images) data.

With the scanning method, from the 1,525 images, the pathologists made 1,982 annotations, which were divided into 559 normal glands, 535 stroma, 273 Gleason 3, 281 Gleason 4, and 334 Gleason 5 annotations. Likewise, the images were randomly divided into two groups, *i.e.*, training (1,220) and validation (305) images.

Table 1 summarizes how the images and annotations were distributed for both classification methods.

Using the categorical classification method, the trained model obtained a 94.1% validation accuracy for determining malignant tissue and its Gleason pattern. Subsequently, we evaluated the model using 100 test images that were not used during the training process. However, the concordance with our expert uropathologist analysis was only 44%. When we separately analyzed the correct prediction between groups, we found that, when the true label was benign, the model precision was 48%, whereas, for Gleason 3, 4, and 5, it was 60%, 34.6%, and 33.3%, respectively (Table 2).

With the image scanning method, our model demonstrated a validation accuracy of 91.2%. When the test images

Table 1 - Characteristics of annotated slides.

	n (%)
Whole prostatectomy slides	12 (100)
Categorical classification method	
Total no. of slide patches generated	1,525 (100)
Only benign tissue	740 (48.5)
Gleason 3 pattern predominant	251 (16.4)
Gleason 4 pattern predominant	254 (16.7)
Gleason 5 pattern predominant	280 (18.4)
Scanning method	
Total no. of annotations generated	1,982 (100)
Stroma	535 (27.0)
Normal glands	559 (28.2)
Gleason 3 pattern	273 (13.8)
Gleason 4 pattern	281 (14.2)
Gleason 5 pattern	334 (16.8)

Table 2 - Categorical classification method—true label (pathologist label) *versus* predicted label (deep learning label) in test dataset images.

			Predicted label			
		Benign	Gleason 3	Gleason 4	Gleason 5	Total
True label	Benign	12 (48%*)	2	4	7	25
	Gleason 3	4	15 (60%*)	4	2	25
	Gleason 4	4	7	9 (34.6%*)	6	26
	Gleason 5	1	6	9	8 (33.3%*)	24

*Correct concordance between pathologist analysis and trained model prediction.

were used, the concordance between the deep learning method and uropathologists was surprisingly high; the approach correctly detected benign and cancerous tissues, including their patterns, in 89% of the never-before-seen images (Figure 1). When the annotations were evaluated individually, 117 areas were detected by the model in 100 of the images, among which 106 areas were detected correctly (90.5%) (Table 3). The correct annotation rate was as follows: 31 predictions for benign tissue (96.7% correct), 27 predictions for Gleason 3 (92.5% correct), 29 annotations for Gleason 4 (96.5% correct), and 30 predictions for Gleason 5 (76.6% correct).

DISCUSSION

A slide analysis of a biopsy or radical prostatectomy specimen is traditionally conducted manually by pathologists, using optical microscopes. In recent years, owing to rapidly evolving visual system technologies, artificial intelligence techniques have been developed to support the work of pathologists (15).

In comparison to the results recorded by experienced uropathologists, using the proposed deep learning scanning method, we demonstrated an accuracy of 89% in real-world images in the PCa diagnosis and determination of the Gleason/ISUP grading. However, our categorical method had a low global accuracy of 44% in the never-before-seen images. These findings suggest that delimitating the areas of interest in each image patch is an extremely time-consuming and stressful activity, but can generate superior results. Using Mask-RCNN, Couteaux (16) obtained a 90.6% accuracy in automatically detecting meniscal tears in the knee, demonstrating the effectiveness of this technique and its applicability in any field of medicine.

In addition, Nagpal et al. (17) used an extremely robust database, comprising 112 million pathologist-annotated image patches from 1,226 whole-slide images, and achieved a mean accuracy of 70% compared to 61% among the 29 general pathologists. Interestingly, they reported that the tumor grading evaluations by uropathologists were significantly more accurate than those of the general pathologists (71.7% *versus* 58.0%, $p<0.001$), suggesting that the deep learning model may have a higher proficiency for tumor grading than general pathologists (18).

Litjens et al. (12) introduced deep learning as a tool for improving the objectivity and efficiency of a histopathological evaluation. They studied the deep learning performance in the PCa identification during a biopsy, and their algorithm was able to detect all slides containing PCa, whereas 30–40% of the slides containing normal tissue needed human intervention to be excluded. Using specimens from radical prostatectomies segmented in a tissue microarray, Arvaniti

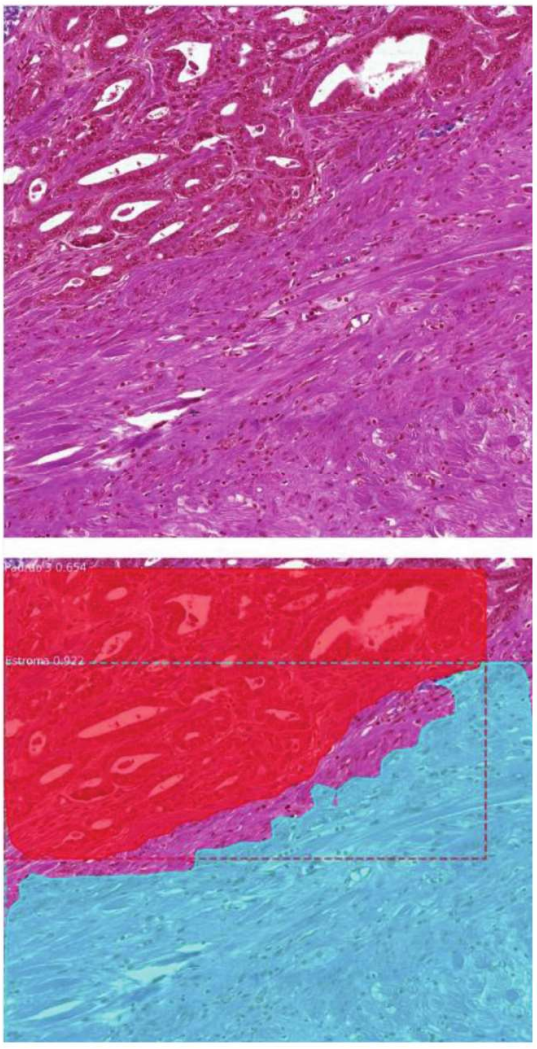

Figure 1 - Scanning method example—The upper image shows an image patch extracted from a radical prostatectomy specimen slide. The lower image demonstrates the scanning model prediction. The method automatically detected a Gleason 3 pattern area in the upper part of the image and stroma tissue in the lower part of the patch).

Table 3 - Scanning classification method—true area label (pathologist analysis) *versus* deep learning predicted area label in test dataset images.

			Predicted area label			
		Benign	Gleason 3	Gleason 4	Gleason 5	Total
True area label	Benign	30 (96.7%*)	1	0	6	37
	Gleason 3	1	25 (92.5%*)	1	0	27
	Gleason 4	0	1	28 (96.5%*)	1	30
	Gleason 5	0	0	0	23 (76.6%*)	23

*Correct concordance between pathologist analysis and trained model prediction.

et al. (13) reached an inter-annotator agreement between the model and two pathologists at 0.75 and 0.71, respectively, comparable with the inter-pathologist agreement (kappa= 0.71).

The accuracy in the determination of the Gleason/ISUP score depends directly on the experience of the pathologist. However, the number of pathologists in most parts of the world is insufficient for supporting the complexities of sub-specialization, which is more serious in lower-income countries such as Brazil.

Increasing the number of images is essential for improving the accuracy of our model. In addition, by evaluating the image sets, we noted that some morphologies are matter of confusion, such as the seminal vesicle epithelium and inflammatory infiltrate, which may be difficult for the algorithm to solve. We observed that, in addition to increasing the number of images, if we include different aspects of Gleason pattern 5, (*e.g.*, inflammation, atrophy, and post-atrophic hyperplasia), we believe our algorithm will be able to learn and distinguish the different morphological aspects that may be a matter of confusion.

The involvement of multiple uropathologists may also improve the quality of the image sets by selecting those achieving a consensus.

With our numbers, we want to reinforce the satisfactory results of deep learning algorithms in the diagnosis and grading of PCa, as well as their utility as a tool used in daily routines to improve quality and speed of pathologists, thereby benefiting the welfare of the society.

This is a new type of knowledge, and many variables should be assessed before excellence can be achieved. For example, what is the best machine learning method available? How many images are necessary to achieve a good agreement? Who should train the machine? Will the results be based exclusively on machine observations or will pathologists have to sign off on the final outcome? Such questions need to be addressed in future large-scale studies, which should be conducted globally.

CONCLUSIONS

Our data have shown that a deep learning algorithm has high potential for the detection and grading of PCa. Scanning methods are likely to be superior to simple classification methods when a limited dataset is available. Future applications of deep learning methods will be unlimited, and should therefore be studied extensively during the next few years.

AUTHOR CONTRIBUTIONS

Melo PAS and Leite KRM conceptualized the study. Melo PAS, Estivallet CLN and Leite KRM collected and analyzed the data. Melo PAS, Srougi M, Nahas WC and Leite KRM conducted the formal analysis. Melo PAS, Estivallet CLN and Leite KRM developed the methodology. Melo PAS, Estivallet CLN and Leite KRM were in charge of project administration. Melo PAS and Estivallet CLN handled the software. Melo PAS, Srougi M, Nahas WC and Leite KRM supervised the study. Melo PAS, Estivallet CLN and Leite KRM validated the data. Melo PAS, Srougi M, Nahas WC and Leite KRM visualized the study. Melo PAS, Srougi M, Nahas WC and Leite KRM wrote the manuscript.

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018; 68(6):394-424. https://doi.org/10.3322/caac.21492
2. Roehrborn CG, Black LK. The economic burden of prostate cancer. BJU Int. 2011;108(6):806-13. https://doi.org/10.1111/j.1464-410X.2011.10365.x
3. Gleason DF. Classification of prostatic carcinomas. Cancer Chemother Rep. 1966;50(3):125-8.
4. Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA, et al. The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System. Am J Surg Pathol. 2016;40(2):244-52. https://doi.org/10.1097/PAS.0000000000000530
5. Adesina A, Chumba D, Nelson AM, Orem J, Roberts DJ, Wabinga H, et al. Improvement of pathology in sub-Saharan Africa. Lancet Oncol. 2013; 14(4):e152-7. https://doi.org/10.1016/S1470-2045(12)70598-3
6. Jemal A, Center MM, DeSantis C, Ward EM. Global patterns of cancer incidence and mortality rates and trends. Cancer Epidemiol Biomarkers Prev. 2010;19(8):1893-907. https://doi.org/10.1158/1055-9965.EPI-10-0437
7. Egevad L, Ahmad AS, Algaba F, Berney DM, Boccon-Gibod L, Compérat E, et al. Standardization of Gleason grading among 337 European pathologists. Histopathology. 2013;62(2):247-56. https://doi.org/10.1111/his.12008
8. Allsbrook WC Jr, Mangold KA, Johnson MH, Lane RB, Lane CG, Amin MB, et al. Interobserver reproducibility of Gleason grading of prostatic carcinoma: urologic pathologists. Hum Pathol. 2001;32(1):74-80. https://doi.org/10.1053/hupa.2001.21134
9. Egevad L, Delahunt B, Berney DM, Bostwick DG, Cheville J, Comperat E, et al. Utility of Pathology Imagebase for standardisation of prostate cancer grading. Histopathology. 2018;73(1):8-18. https://doi.org/10.1111/his.13471
10. Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural Comput. 2006;18(7):1527-54. https://doi.org/10.1162/neco.2006.18.7.1527
11. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553): 436-44. https://doi.org/10.1038/nature14539
12. Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep. 2016;6:26286. https://doi.org/10.1038/srep26286
13. Arvaniti E, Fricker KS, Moret M, Rupp N, Hermanns T, Fankhauser C, et al. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. Sci Rep. 2018;8(1):12054. https://doi.org/10.1038/s41598-018-30535-1
14. Kott O, Linsley D, Amin A, Karagounis A, Jeffers C, Golijanin D, et al. Development of a Deep Learning Algorithm for the Histopathologic Diagnosis and Gleason Grading of Prostate Cancer Biopsies: A Pilot Study. Eur Urol Focus. 2021;7(2):347-51. https://doi.org/10.1016/j.euf.2019.11.003
15. Cui M, Zhang DY. Artificial intelligence and computational pathology. Lab Invest. 2021;101(4):412-22. https://doi.org/10.1038/s41374-020-00514-0

16. Couteaux V, Si-Mohamed S, Nempont O, Lefevre T, Popoff A, Pizaine G, et al. Automatic knee meniscus tear detection and orientation classification with Mask-RCNN. Diagn Interv Imaging. 2019;100(4):235-42. https://doi.org/10.1016/j.diii.2019.03.002
17. Nagpal K, Foote D, Liu Y, Chen PC, Wulczyn E, Tan F, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. NPJ Digit Med. 2019;2:48. https://doi.org/10.1038/s41746-019-0112-2
18. Nagpal K, Foote D, Tan F, Liu Y, Chen PC, Steiner DF, et al. Development and Validation of a Deep Learning Algorithm for Gleason Grading of Prostate Cancer From Biopsy Specimens. JAMA Oncol. 2020;6(9):1372-80. https://doi.org/10.1001/jamaoncol.2020.2485

ANEXO III – TRABALHO APRESENTADO E Ppremiado NO XXXVII CONGRESSO BRASILEIRO DE UROLOGIA

XXXVII Congresso Brasileiro Urologia
Curitiba | PR
24 a 27 AGOSTO | 2019
Expo Unimed Curitiba
A UROLOGIA COM FOCO NA MEDICINA PERSONALIZADA E HUMANIZADA

Certificamos que os Drs. Petronio Augusto de Souza Melo, Carmen Liane Neubarth Estivallet e Katia Ramos Moreira Leite ficaram em **2º lugar** com o trabalho científico intitulado **"A Inteligência Artificial como Ferramenta no Diagnóstico Histopatológico do Câncer de Prostata"**.

XXXVII Congresso Brasileiro de Urologia 2019
Diretoria Biênio 2018-2019
Curitiba, 25 de agosto de 2019

Sebastião Westphal
Presidente da SBU Gestão 2018/2019

Gustavo Franco Carvalhal
Presidente da Comissão Científica

Iniciativa e Realização
SBU
SOCIEDADE BRASILEIRA DE UROLOGIA

8. REFERÊNCIAS

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209-49.

2. Hsing AW, Chokkalingam AP. Prostate cancer epidemiology. Front Biosci. 2006;11:1388-413.

3. Cuzick J, Thorat MA, Andriole G, Brawley OW, Brown PH, Culig Z, et al. Prevention and early detection of prostate cancer. Lancet Oncol. 2014;15(11):e484-92.

4. Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA). Estimativa 2020: incidência de câncer no Brasil. Rio de Janeiro: INCA; 2020.

5. Guimarães R, Muzi C, Trend S. Mortalidade por câncer de próstata no Brasil: tendência temporal por escolaridade e local de residência. Rev Panam Salud Publica; 2019;43:e62.

6. Roehrborn CG, Black LK. The economic burden of prostate cancer. BJU Int. 2011;108(6):806-13.

7. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med. 2000;343(2):78-85.

8. Narod SA, Neuhausen S, Vichodez G, Armel S, Lynch HT, Ghadirian P, et al. Rapid progression of prostate cancer in men with a BRCA2 mutation. Br J Cancer. 2008;99(2):371-4.

9. Ewing CM, Ray AM, Lange EM, Zuhlke KA, Robbins CM, Tembe WD, et al. Germline mutations in HOXB13 and prostate-cancer risk. N Engl J Med. 2012;366(2):141-9.

10. Roddam AW, Allen NE, Appleby P, Key TJ, Group EHaPCC. Endogenous sex hormones and prostate cancer: a collaborative analysis of 18 prospective studies. J Natl Cancer Inst. 2008;100(3):170-83.

11. Khera M. Male hormones and men's quality of life. Curr Opin Urol. 2016;26(2):152-7.

12. Ryan CJ, Smith MR, Fizazi K, Saad F, Mulders PF, Sternberg CN, et al. Abiraterone acetate plus prednisone versus placebo plus prednisone in chemotherapy-naive men with metastatic castration-resistant prostate cancer (COU-AA-302): final overall survival analysis of a randomised, double-blind, placebo-controlled phase 3 study. Lancet Oncol. 2015;16(2):152-60.

13. Di Sebastiano KM, Mourtzakis M. The role of dietary fat throughout the prostate cancer trajectory. Nutrients. 2014;6(12):6095-109.

14. Mottet N, van den Bergh RCN, Briers E, Van den Broeck T, Cumberbatch MG, De Santis M, et al. EAU-EANM-ESTRO-ESUR-SIOG Guidelines on Prostate Cancer-2020 Update. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent. Eur Urol. 2021;79(2):243-62.

15. Cao Y, Ma J. Body mass index, prostate cancer-specific mortality, and biochemical recurrence: a systematic review and meta-analysis. Cancer Prev Res (Phila). 2011;4(4):486-501.

16. Huncharek M, Haddock KS, Reid R, Kupelnick B. Smoking as a risk factor for prostate cancer: a meta-analysis of 24 prospective cohort studies. Am J Public Health. 2010;100(4):693-701.

17. Barry MJ. Clinical practice. Prostate-specific-antigen testing for early diagnosis of prostate cancer. N Engl J Med. 2001;344(18):1373-7.

18. Sartor O, de Bono JS. Metastatic Prostate Cancer. N Engl J Med. 2018;378(7):645-57.

19. Kasivisvanathan V, Rannikko AS, Borghi M, Panebianco V, Mynderse LA, Vaarala MH, et al. MRI-Targeted or Standard Biopsy for Prostate-Cancer Diagnosis. N Engl J Med. 2018;378(19):1767-77.

20. Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA, et al. The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System. Am J Surg Pathol. 2016;40(2):244-52.

21. Allsbrook WC, Mangold KA, Johnson MH, Lane RB, Lane CG, Amin MB, et al. Interobserver reproducibility of Gleason grading of prostatic carcinoma: urologic pathologists. Hum Pathol. 2001;32(1):74-80.

22. Gleason DF, Mellinger GT. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. J Urol. 1974;111(1):58-64.

23. Epstein JI, Allsbrook WC, Amin MB, Egevad LL, Committee IG. The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma. Am J Surg Pathol. 2005;29(9):1228-42.

24. Hernandez DJ, Nielsen ME, Han M, Trock BJ, Partin AW, Walsh PC, et al. Natural history of pathologically organ-confined (pT2), Gleason score 6 or less, prostate cancer after radical prostatectomy. Urology. 2008;72(1):172-6.

25. Pierorazio PM, Walsh PC, Partin AW, Epstein JI. Prognostic Gleason grade grouping: data based on the modified Gleason scoring system. BJU Int. 2013;111(5):753-60.

26. Truong M, Hollenberg G, Weinberg E, Messing EM, Miyamoto H, Frye TP. Impact of Gleason Subtype on Prostate Cancer Detection Using Multiparametric Magnetic Resonance Imaging: Correlation with Final Histopathology. J Urol. 2017;198(2):316-21.

27. Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep. 2016;6:26286.

28. Pelzer AE, Bektic J, Berger AP, Halpern EJ, Koppelstätter F, Klauser A, et al. Are transition zone biopsies still necessary to improve prostate cancer detection? Results from the tyrol screening project. Eur Urol. 2005;48(6):916-21; discussion 21.

29. Kweldam CF, Kümmerlin IP, Nieboer D, Verhoef EI, Steyerberg EW, van der Kwast TH, et al. Disease-specific survival of patients with invasive cribriform and intraductal prostate cancer at diagnostic biopsy. Mod Pathol. 2016;29(6):630-6.

30. Epstein JI. Prognostic significance of tumor volume in radical prostatectomy and needle biopsy specimens. J Urol. 2011;186(3):790-7.

31. Sehdev AE, Pan CC, Epstein JI. Comparative analysis of sampling methods for grossing radical prostatectomy specimens performed for

nonpalpable (stage T1c) prostatic adenocarcinoma. Hum Pathol. 2001;32(5):494-9.

32. Trock BJ, Guo CC, Gonzalgo ML, Magheli A, Loeb S, Epstein JI. Tertiary Gleason patterns and biochemical recurrence after prostatectomy: proposal for a modified Gleason scoring system. J Urol. 2009;182(4):1364-70.

33. Bill-Axelson A, Holmberg L, Garmo H, Rider JR, Taari K, Busch C, et al. Radical prostatectomy or watchful waiting in early prostate cancer. N Engl J Med. 2014;370(10):932-42.

34. Coughlin GD, Yaxley JW, Chambers SK, Occhipinti S, Samaratunga H, Zajdlewicz L, et al. Robot-assisted laparoscopic prostatectomy versus open radical retropubic prostatectomy: 24-month outcomes from a randomised controlled study. Lancet Oncol. 2018;19(8):1051-60.

35. Donovan JL, Hamdy FC, Lane JA, Mason M, Metcalfe C, Walsh E, et al. Patient-Reported Outcomes after Monitoring, Surgery, or Radiotherapy for Prostate Cancer. N Engl J Med. 2016;375(15):1425-37.

36. Sridhar AN, Cathcart PJ, Yap T, Hines J, Nathan S, Briggs TP, et al. Recovery of Baseline Erectile Function in Men Following Radical Prostatectomy for High-Risk Prostate Cancer: A Prospective Analysis Using Validated Measures. J Sex Med. 2016;13(3):435-43.

37. Hoskin PJ, Rojas AM, Bownes PJ, Lowe GJ, Ostler PJ, Bryant L. Randomised trial of external beam radiotherapy alone or combined with high-dose-rate brachytherapy boost for localised prostate cancer. Radiother Oncol. 2012;103(2):217-22.

38. Gillessen S, Attard G, Beer TM, Beltran H, Bjartell A, Bossi A, et al. Management of Patients with Advanced Prostate Cancer: Report of the Advanced Prostate Cancer Consensus Conference 2019. Eur Urol. 2020;77(4):508-47.

39. Harris WP, Mostaghel EA, Nelson PS, Montgomery B. Androgen deprivation therapy: progress in understanding mechanisms of resistance and optimizing androgen depletion. Nat Clin Pract Urol. 2009;6(2):76-85.

40. Sweeney CJ, Chen YH, Carducci M, Liu G, Jarrard DF, Eisenberger M, et al. Chemohormonal Therapy in Metastatic Hormone-Sensitive Prostate Cancer. N Engl J Med. 2015;373(8):737-46.

41. Nguyen PL, Alibhai SM, Basaria S, D'Amico AV, Kantoff PW, Keating NL, et al. Adverse effects of androgen deprivation therapy and strategies to mitigate them. Eur Urol. 2015;67(5):825-36.

42. Kirby M, Hirst C, Crawford ED. Characterising the castration-resistant prostate cancer population: a systematic review. Int J Clin Pract. 2011;65(11):1180-92.

43. Petrylak DP, Tangen CM, Hussain MH, Lara PN, Jones JA, Taplin ME, et al. Docetaxel and estramustine compared with mitoxantrone and prednisone for advanced refractory prostate cancer. N Engl J Med. 2004;351(15):1513-20.

44. Bahn DK, Silverman P, Lee F, Badalament R, Bahn ED, Rewcastle JC. Focal prostate cryoablation: initial results show cancer control and potency preservation. J Endourol. 2006;20(9):688-92.

45. Tay KJ, Polascik TJ. Focal Cryotherapy for Localized Prostate Cancer. Arch Esp Urol. 2016;69(6):317-26.

46. Gelet A, Chapelon JY, Poissonnier L, Bouvier R, Rouvière O, Curiel L, et al. Local recurrence of prostate cancer after external beam radiotherapy: early experience of salvage therapy using high-intensity focused ultrasonography. Urology. 2004;63(4):625-9.

47. Ahmed HU, Hindley RG, Dickinson L, Freeman A, Kirkham AP, Sahu M, et al. Focal therapy for localised unifocal and multifocal prostate cancer: a prospective development study. Lancet Oncol. 2012;13(6):622-32.

48. Valerio M, Ahmed HU, Emberton M, Lawrentschuk N, Lazzeri M, Montironi R, et al. The role of focal therapy in the management of localised prostate cancer: a systematic review. Eur Urol. 2014;66(4):732-51.

49. Topalian SL, Hodi FS, Brahmer JR, Gettinger SN, Smith DC, McDermott DF, et al. Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. N Engl J Med. 2012;366(26):2443-54.

50. Klotz L, Vesprini D, Sethukavalan P, Jethava V, Zhang L, Jain S, et al. Long-term follow-up of a large active surveillance cohort of patients with prostate cancer. J Clin Oncol. 2015;33(3):272-7.

51. Johansson JE, Andrén O, Andersson SO, Dickman PW, Holmberg L, Magnuson A, et al. Natural history of early, localized prostate cancer. JAMA. 2004;291(22):2713-9.

52. Russell S, Norvig P. Artificial Intelligence: A Modern Approach. 3rd ed. Upper Saddle River, NJ: Prentice Hall;2016.

53. Mitchell T. Machine Learning. 1st ed. New York, NY: McGraw-Hill;1997.

54. Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge, MA: MIT Press;2016.

55. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56.

56. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115-8.

57. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 2016;316(22):2402-10.

58. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-44.

59. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018;15(141).

60. Aliper A, Plis S, Artemov A, Ulloa A, Mamoshina P, Zhavoronkov A. Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data. Mol Pharm. 2016;13(7):2524-30.

61. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. JAMA. 2018;319(13):1317-8.

62. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med. 2019;380(14):1347-58.

63. Price WN, Gerke S, Cohen IG. Potential Liability for Physicians Using Artificial Intelligence. JAMA. 2019;322(18):1765-6.

64. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60-88.

65. Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84-90.

66. Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. Lancet Oncol. 2019;20(5):e253-e61.

67. Bulten W, Pinckaers H, van Boven H, Vink R, de Bel T, van Ginneken B, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. Lancet Oncol. 2020;21(2):233-41.

68. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med. 2018;24(10):1559-67.

69. Sirinukunwattana K, Ahmed Raza SE, Yee-Wah Tsang, Snead DR, Cree IA, Rajpoot NM. Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images. IEEE Trans Med Imaging. 2016;35(5):1196-206.

70. Wang D, Khosla A, Gargeya R, Irshad H, Beck A. Deep Learning for Identifying Metastatic Breast Cancer. arXiv:160605718 [cs]. 2016.

71. Ribeiro M, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery; 2016. p. 1135-44.

72. Montavon G, Samek W, Müller K-R. Methods for interpreting and understanding deep neural networks. Digit Signal Process. 2018;73:1-15.

73. Cireşan DC, Giusti A, Gambardella LM, Schmidhuber J. Mitosis detection in breast cancer histology images with deep neural networks. Med Image Comput Comput Assist Interv. 2013;16(Pt 2):411-8.

74. Liu S, Zheng H, Feng Y, Li W, editors. Prostate cancer diagnosis using deep learning with 3D multiparametric MRI. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series; 2017 March 01, 2017.

75. Teixeira S, Ribeiro J, Oliveira A, Santos J. Artificial intelligence in prostate cancer diagnosis: a systematic review of the past 10 years. Expert Review of Anticancer Therapy, 20(10), 835-847;. 2020.

76. Arvaniti E, Fricker KS, Moret M, Rupp N, Hermanns T, Fankhauser C, et al. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. Sci Rep. 2018;8(1):12054.

77. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368:m689.
78. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2(4):230-43.
79. Castelvecchi D. Can we open the black box of AI? Nature. 2016;538(7623):20-3.
80. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov. 2019;9(4):e1312.
81. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: Addressing ethical challenges. PLoS Med. 2018;15(11):e1002689.
82. Matheny ME, Whicher D, Thadaney Israni S. Artificial Intelligence in Health Care: A Report From the National Academy of Medicine. JAMA. 2020;323(6):509-10.
83. Balthazar P, Harri P, Prater A, Safdar NM. Protecting Your Patients' Interests in the Era of Big Data, Artificial Intelligence, and Predictive Analytics. J Am Coll Radiol. 2018;15(3 Pt B):580-6.
84. Shaw J, Rudzicz F, Jamieson T, Goldfarb A. Artificial Intelligence and the Implementation Challenge. J Med Internet Res. 2019;21(7):e13659.
85. Meskó B, Drobni Z, Bényei É, Gergely B, Győrffy Z. Digital health is a cultural transformation of traditional healthcare. Mhealth. 2017;3:38.
86. Bostwick D, Cheng L. Urologic Surgical Pathology. 4th ed. Philadelphia: Elsevier;2020.
87. Van Rossum G, Drake FL. Python 3 Reference Manual: CreateSpace; 2009.
88. Brooks J. COCO Annotator. 2019.
89. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998;86(11):2278-324.
90. He K, Gkioxari G, Dollar P, Girshick R. Mask R-CNN. IEEE Trans Pattern Anal Mach Intell. 2020;42(2):386-97.
91. Géron A. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow. 2ª ed. Sebastopol: O'Reilly Media; 2019.

92. Boureau Y, Ponce J, LeCun Y. A hierarchical model for learning invariant features. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (pp. 1455-1462). IEEE.
93. Nair V, Hinton G. Rectified Linear Units Improve Restricted Boltzmann Machines Vinod Nair2010. 807-14 p.
94. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2015;39.
95. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778. 2016.
96. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 14091556. 2014.
97. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation2015. 3431-40 p.
98. Abdulla W. Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow. Github repository: Github; 2017.
99. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow : Large-Scale Machine Learning on Heterogeneous Distributed Systems2015.
100. Perez L, Wang J. The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621. 2017.
101. Shorten C, Khoshgoftaar T. A survey on Image Data Augmentation for Deep Learning. Journal of Big Data. 2019;6.
102. Kingma D, Ba J. Adam: A Method for Stochastic Optimization. International Conference on Learning Representations. 2014.
103. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research. 2014;15:1929-58.
104. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer;2009.
105. Cui M, Zhang DY. Artificial intelligence and computational pathology. Lab Invest. 2021;101(4):412-22.

106. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. 2015.
107. Nagpal K, Foote D, Liu Y, Chen PC, Wulczyn E, Tan F, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. NPJ Digit Med. 2019;2:48.
108. Litjens G, Debats O, Barentsz J, Karssemeijer N, Huisman H. Computer-aided detection of prostate cancer in MRI. IEEE Trans Med Imaging. 2014;33(5):1083-92.
109. Nir G, Karimi D, Goldenberg SL, Fazli L, Skinnider BF, Tavassoli P, et al. Comparison of Artificial Intelligence Techniques to Evaluate Performance of a Classifier for Automatic Grading of Prostate Cancer From Digitized Histopathologic Images. JAMA Netw Open. 2019;2(3):e190442.
110. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med. 2019;25(8):1301-9.
111. Egevad L, Swanberg D, Delahunt B, Ström P, Kartasalo K, Olsson H, et al. Identification of areas of grading difficulties in prostate cancer and comparison with artificial intelligence assisted grading. Virchows Arch. 2020;477(6):777-86.
112. Gurcan MN, Boucheron LE, Can A, Madabhushi A, Rajpoot NM, Yener B. Histopathological image analysis: a review. IEEE Rev Biomed Eng. 2009;2:147-71.
113. Komura D, Ishikawa S. Machine Learning Methods for Histopathological Image Analysis. Comput Struct Biotechnol J. 2018;16:34-42.
114. Japkowicz N, Stephen S. The Class Imbalance Problem: A Systematic Study. Intell Data Anal. 2002;6:429-49.
115. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. J Pathol Inform. 2016;7:29.
116. Nagpal K, Foote D, Tan F, Liu Y, Chen PC, Steiner DF, et al. Development and Validation of a Deep Learning Algorithm for Gleason Grading of Prostate Cancer From Biopsy Specimens. JAMA Oncol. 2020.
117. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions2015. 1-9 p.

118. Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions2017. 1800-7 p.
119. Pan S, Yang Q. A Survey on Transfer Learning. Knowledge and Data Engineering, IEEE Transactions on. 2010;22:1345-59.
120. Chawla N, Bowyer K, Hall L, Kegelmeyer W. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res (JAIR). 2002;16:321-57.
121. Tellez D, Litjens G, Bándi P, Bulten W, Bokhorst JM, Ciompi F, et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Med Image Anal. 2019;58:101544.
122. Ström P, Kartasalo K, Olsson H, Solorzano L, Delahunt B, Berney DM, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. Lancet Oncol. 2020;21(2):222-32.
123. Ilse M, Tomczak J, Welling M. Attention-based Deep Multiple Instance Learning. 2018.
124. Couteaux V, Si-Mohamed S, Nempont O, Lefevre T, Popoff A, Pizaine G, et al. Automatic knee meniscus tear detection and orientation classification with Mask-RCNN. Diagn Interv Imaging. 2019;100(4):235-42.
125. Marginean F, Arvidsson I, Simoulis A, Christian Overgaard N, Åström K, Heyden A, et al. An Artificial Intelligence-based Support Tool for Automation and Standardisation of Gleason Grading in Prostate Biopsies. Eur Urol Focus. 2021;7(5):995-1001.
126. Nakai Y, Tanaka N, Shimada K, Konishi N, Miyake M, Anai S, et al. Review by urological pathologists improves the accuracy of Gleason grading by general pathologists. BMC Urol. 2015;15:70.
127. Kim KH, Lim SK, Shin TY, Lee JY, Chung BH, Rha KH, et al. Upgrading of Gleason score and prostate volume: a clinicopathological analysis. BJU Int. 2013;111(8):1310-6.
128. Eloy C, Marques A, Pinto J, Pinheiro J, Campelos S, Curado M, et al. Artificial intelligence-assisted cancer diagnosis improves the efficiency of pathologists in prostatic biopsies. Virchows Arch. 2023;482(3):595-604.

129. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. Nat Rev Clin Oncol. 2019;16(11):703-15.
130. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention Is All You Need. 2017.
131. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C. A Survey on Deep Transfer Learning: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4–7, 2018, Proceedings, Part III. 2018. p. 270-9.
132. Turkbey B, Rosenkrantz AB, Haider MA, Padhani AR, Villeirs G, Macura KJ, et al. Prostate Imaging Reporting and Data System Version 2.1: 2019 Update of Prostate Imaging Reporting and Data System Version 2. Eur Urol. 2019;76(3):340-51.
133. Montalto M. Artificial intelligence in medical imaging: the future is now. Radiol Bras. 2019;52(5):VII-VIII.