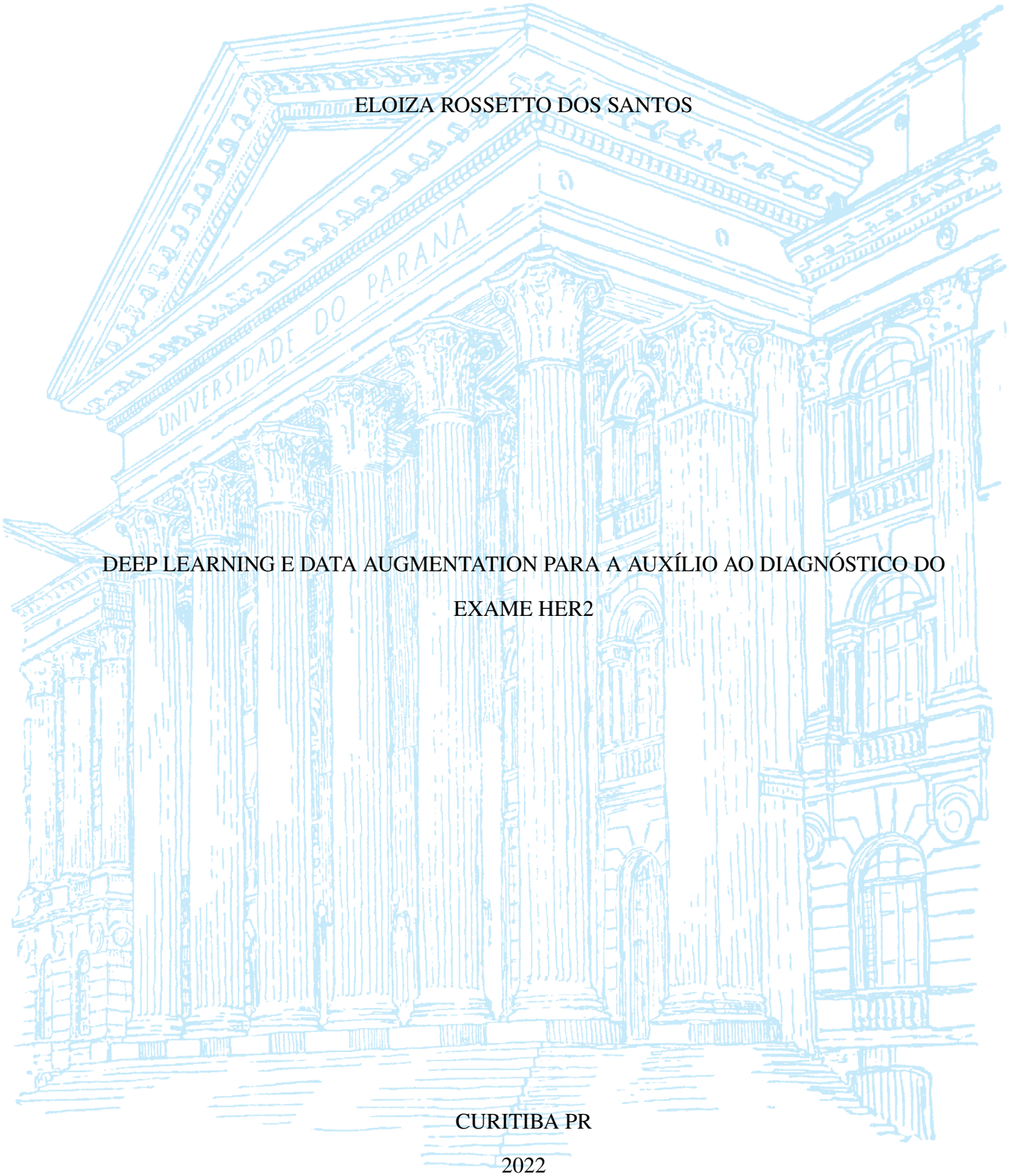
UNIVERSIDADE FEDERAL DO PARANÁ

ELOIZA ROSSETTO DOS SANTOS



DEEP LEARNING E DATA AUGMENTATION PARA A AUXÍLIO AO DIAGNÓSTICO DO EXAME HER2

CURITIBA PR

2022

ELOIZA ROSSETTO DOS SANTOS

DEEP LEARNING E DATA AUGMENTATION PARA A AUXÍLIO AO DIAGNÓSTICO DO EXAME HER2

Trabalho apresentado como requisito parcial à conclusão do Curso de Bacharelado em Informática Biomédica, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: David Menotti.

CURITIBA PR

2022

UNIVERSIDADE FEDERAL DO PARANÁ

ATA DE REUNIÃO

Universidade Federal do Paraná

Setor de Ciências Exatas

Curso de Informática Biomédica

Ata de Apresentação de Trabalho de Conclusão de Curso em Infromática - DMG - CI262

Título do Trabalho: DEEP LEARNING E DATA AUGMENTATION PARA A AUXÍLIO AO DIAGNÓSTICO DO EXAME HER2

Autor:

GRR 20173522 Nome: ELOIZA ROSSETO DOS SANTOS

Apresentação: Data: 18 / 05 / 2022 Hora: 14:00 Local: https://bbb.c3sl.ufpr.br/b/dav-mnm-kre-yvg

AVALIAÇÃO – Produto escrito	ORIENTADOR	MEMBRO 1	MEMBRO 2	MÉDIA
Conteúdo (00-40)	30	30	30	30
Referência Bibliográfica (00-10)	00	00	00	8
Formato (00-05)	00	00	00	2
AVALIAÇÃO – Produto escrito				
Domínio do Assunto (00-15)	00	00	00	11
Desenvolvimento do Assunto (00-05)	00	00	00	4
Técnica de Apresentação (00-03)	00	00	00	3
Uso do Tempo (00-02)	00	00	00	2
AVALIAÇÃO - Desenvolvimento				
Nota do Orientador (00-20)	15	**********	**********	15
NOTA FINAL	**********	**********	**********	**75**

Pesos indicados são sugestões.

Conforme decisão do colegiado do curso de Informática Biomédica, a entrega dos documentos comprobatório de trabalho de conclusão de curso deve deve respeitar os seguintes procedimentos: Orientador deve abrir um processo no Sistema Eletrônico de Informações (SEI – UFPR); Selecionar o tipo: Graduação/Ensino Técnico: Avaliação de Aluno (provas, exames e trabalhos, inclusive verificações suplementares); informar os interessados: nome do aluno e o nome do orientador; incluir esta ata escaneada ou assinada eletrônicamente no SEI e a versão final do pdf da monografia do aluno.; Tramitar o processo para UFPR/R/ET/CCIB - Coordenação do Curso de Informática Biomédica.

Orientador: DAVID MENOTTI GOMES

Membro 1: LUCAS FERRARI DE OLIVEIRA

Membro 2: RAYSON BARTOSKI LAROCA DOS SANTOS

Documento assinado eletronicamente por **DAVID MENOTTI GOMES**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 18/05/2022, às 15:08, conforme art. 1º, III, "b", da Lei 11.419/2006.

Dedico esse trabalho aos futuros informatas biomédicos que assim como eu amam misturar computação com biologia e saúde

AGRADECIMENTOS

Agradeço meus colegas de curso, cm quem convivi durante os últimos anos, pelo companheirismo e pela troca de experiências que me permitiram crescer não só como pessoa, mas também como formando. Aos colegas do laboratório VRI (Visão Robótica e Imagem) por me ajudarem a compreender mais sobre o mundo das redes neurais profundas e terem proporcionado um ambiente amistoso para meu desenvolvimento. Por último, agradeço meus familiares, em especial meu namorado por ter me dado suporte emocional e cuidado de mim durante os anos que estive na instituição.

LISTA DE FIGURAS

3.1 Exemplo de lâmina inteira disponível no *dataset* 19

3.2 Exemplos de *patches* para cada classe da base de dados utilizada. 0/1+ HER2 negativo, 2+ indeterminado, 3+ HER2 positivo e Ruído.. 20

3.3 Imagens aumentadas. A - imagem original, B - resultado após flip, C - resultado após transformação elástica, D - resultado de rotação. 21

3.4 *Pipeline* de processamento de imagens proposto. 22

LISTA DE TABELAS

2.1 Caracterização dos trabalhos e análise comparativa. *Cl.Lâm - Classificação de toda a lâmina, RN - Rede Neural, RF - Random Forest, RNR - Rede neural recorrente, MIL - Multiple Instance Learning, DL - Deep learning, TF - Transfer Learning. 17

3.1 Distribuição de classes entre pacientes da base de dados disponibilizada pela Universidade Warwick. (Qaiser et al., 2018) 19

3.2 Distribuição da base de *patches* rotuladas pelo patologista especialista. 20

3.3 Distribuição das bases original e aumentadas. 21

4.1 Total de pacientes por *fold* criada. 24

4.2 Acurácia e função *loss* da CNN por cada etapa de treinamento. O - *dataset Original*, A1 - *dataset DataAug1*, A2 - *dataset DataAug2* 25

4.3 Métricas de classificação para cada base testada obtida na etapa de teste. A letra P para a métrica precisão, R para *recall* e F1 para *F1 Score*.. 26

4.4 Acurácia média de todas as *folds* de cada classificador. 27

LISTA DE ACRÔNIMOS

HER2	Receptor de crescimento epidérmico humano tipo 2
ER	Estrogênio
PR	Progesterona
IHC	Imuno-histoquímica
HE	Hematoxilina-Eosina
DT	Decision Tree
MLP	Multi-Layer Perceptron
SVM	Support Vector Machines
RF	Random Forest
KNN	K-Nearest Neighbors
CNN	Rede Neural de Convolução
MIL	Multiple Instance Learning

SUMÁRIO

1 INTRODUÇÃO 9
1.1 MOTIVAÇÃO 10
1.2 OBJETIVOS 11
1.3 CONTRIBUIÇÕES 11
1.4 ORGANIZAÇÃO DO DOCUMENTO 11
2 TRABALHOS RELACIONADOS 12
2.1 RESUMO TRABALHOS RELACIONADOS 12
2.2 ANÁLISE COMPARATIVA 15
3 MATERIAIS E MÉTODOS 18
3.1 *DATASET* 18
3.2 PRÉ-PROCESSAMENTO 18
3.3 *DATA AUGMENTATION* 21
3.4 MÉTODO PROPOSTO 22
3.5 ARQUITETURA DOS CLASSIFICADORES 22
4 RESULTADOS E DISCUSSÃO 24
4.1 METODOLOGIA DE AVALIAÇÃO 24
4.2 RESULTADOS EM NÍVEL DE *PATCH* 25
4.3 RESULTADOS EM NÍVEL DE LÂMINA 27
4.4 COMPARAÇÃO COM A LITERATURA 28
5 CONCLUSÃO 29
REFERÊNCIAS 30

1 INTRODUÇÃO

Em 2020, cerca de 19,3 milhões de novos casos de câncer foram diagnosticados e quase 10 milhões de mortes pela doença foram relatadas (Sung et al., 2021). O câncer de mama diagnosticado em mulheres superou o câncer de pulmão como o tipo mais diagnosticado, com uma estimativa de 2,3 milhões de casos, representando 11,7% do total de diagnósticos feitos no ano seguido pelo câncer de pulmão (11,4%), colorretal (10,0%), próstata (7,3%) e estomago (5,6%). O câncer de mama é o que mais acomete mulheres em todo o mundo, em 2020 foram diagnosticados cerca de 2 milhões de novos casos representando 25,8% do total de câncer diagnosticados em mulheres de todas as idades. A doença também ocupa uma posição elevada quanto a mortalidade, sendo o quinto tipo de câncer mais mortal do mundo, responsável por 6,9% do total de mortes em 2020 em decorrência de câncer. Entre mulheres este é o tipo de câncer mais letal sendo responsável por 15,6% das mortes no mesmo período.

O cenário global apresentado também é visto no Brasil, em 2020 o câncer de mama foi o mais diagnosticado com cerca de 97 mil novos casos representando 16,4% dos 592 mil de novos casos de câncer no país (Ferlay et al., 2020). Entre as mulheres, o câncer de mama ocupa a terceira posição quanto a mortalidade, foram registrados cerca de 20 mil óbitos representando 8% do total de mortes pela doença no gênero feminino.

Neste cenário, o diagnóstico e tratamento precoce do câncer de mama tem se tornado um novo desafio para sistemas de saúde em todo o mundo. Um nódulo ou outro sintoma suspeito nas mamas deve ser investigado clinicamente para confirmar se trata de um câncer de mama (Instituto Nacional do Câncer (INCA) - Ministério da Saúde, 2021).

Além do exame clínico, para o diagnóstico também é recomendado exames de imagens como mamografia, ultrassonografia ou ressonância magnética. A confirmação diagnóstica só é feita por meio de biópsia, exame que consiste na coleta de um fragmento do nódulo por meio de punções ou uma pequena cirurgia. Com o material recolhido são feitos também exames para determinar a natureza molecular do câncer.

Para obter a caracterização molecular do câncer de mama são feitos exames imuno-histoquímicos para detectar a presença de receptores de estrogênio (RE), receptores de progesterona (RP) e receptor tipo 2 do fator de crescimento epidérmico humano (HER2) (Jafari et al., 2018). A combinação desses três fatores é utilizada para determinar o subtipo molecular do câncer de mama que pode ser útil no planejamento terapêutico e prognóstico do paciente (Cirqueira et al., 2011).

Os receptores HER2 estão envolvidos na comunicação célula-célula e célula-estroma por meio de um processo conhecido como transdução de sinal. Nele, fatores de crescimento externos ou ligantes afetam a transcrição de vários genes, ativando ou desativando outras proteínas transmembrana e intermediários de sinalização intracelular (Ross et al., 2009). As rotinas celulares de proliferação, sobrevivência, mortalidade e adesão são diretamente afetadas pela cadeia de reações inicializadas no receptor de HER2. A amplificação do gene HER2 no câncer de mama está relacionada com aumento na proliferação, mortalidade celular, agressividade tumoral, maiores chances de metástase, angiogênese (crescimento de vasos sanguíneos) acelerada e apoptose (morte celular) reduzida.

Tumores com alta expressão de HER2, chamados HER2 positivo, são mais agressivos e resistentes a tratamentos quimioterápicos (Loibl e Gianni, 2017). Estima-se que entre 15% e 20% do total de cânceres de mama tenha uma super expressão do receptor HER2. Para este subtipo molecular o tratamento é feito com quimioterapia em conjunto com o medicamento Trastuzumab.

Este tratamento combinado mostrou redução na chance de metástase e morte. Entretanto, a medicação mostrou efeitos colaterais no coração podendo ser necessário o uso de medicação adicional para reduzir possíveis danos.

O diagnóstico de HER2 pode ser feito em nível de proteína, RNA ou DNA (Moelans et al., 2011). O tipo mais comum de diagnóstico é feito em nível de proteína utilizando exames imuno-histoquímicos por serem uma opção mais barata e acessível aos laboratórios. De acordo com as recomendações do Colégio de Patologistas Americanos (CAP) e a Sociedade Americana de Oncologia Clínica (ASCO), um tumor é considerado HER2 positivo se o número de células tumorais mostrando forte super expressão de HER2 (células 3+) excede 10% do total do tecido tumoral; equívoco se o número de células tumorais com expressão moderada de super expressão de HER2 (células 2+) excede 10% do total da população celular tumoral; negativo em outros casos (Wolff et al., 2013). Pacientes diagnosticados como HER2 equívoco necessitam realizar o exame complementar de hibridização *in situ* por fluorescência (FISH) para determinar o status do receptor.

1.1 MOTIVAÇÃO

Uma discordância e variabilidade de diagnóstico entre patologistas tem sido reportada (Vogel et al., 2011; Kaufman et al., 2014; Orlando et al., 2016). Estima-se que cerca de 4% dos diagnósticos negativos e 18% dos diagnósticos positivos são equivocados (Memon et al., 2022; Perez et al., 2006). Essa discordância pode ter origem em decorrência dos preparativos da lâmina, devido a fatores, humanos por se tratar de um exame subjetivo.

Tendo em vista a importância do diagnóstico precoce do câncer de mama, se torna imperativo o desenvolvimento de ferramentas computacionais para auxílio do diagnóstico do exame HER2. Na área da computação, redes neurais e redes neurais convolucionais (CNN) têm sido aplicadas em diversos problemas, principalmente relacionados com visão e processamento de linguagem natural (Li et al., 2021). Sistemas de auxílio ao diagnóstico podem ser encontrados na área de radiologia (Yamashita et al., 2018) e patologia (Srinidhi et al., 2021).

Um dos desafios encontrados ao desenvolver CNNs para exames patológicos é a falta de *datasets* públicos e dados rotulados por especialistas. Uma das características das CNNs é a necessidade abundante de dados rotulados para que o modelo aprenda a generalizar de maneira correta. A falta de dados pode causar o efeito de *overfitting*, em que ao utilizar poucos dados a capacidade de classificação da rede fica prejudicada. Outro empecilho para o desenvolvimento e operação destes sistemas é o tamanho das imagens histopatológicas. Por serem captadas a partir de *scanners* de alta resolução, as imagens histopatológicas podem chegar a ordem de *gigabytes*. Armazenar e processar uma grande quantidade de imagens requer alto poder computacional.

Dado os problemas citados, a principal motivação deste trabalho é criar um sistema de auxilio ao diagnóstico capaz de acelerar o processo diagnóstico e tornar a prática mais precisa. Para mitigar o problema da falta de dados abordagens de *data augmentation* tem se popularizado. Data augmentation trata-se de um conjunto de técnicas capazes de aumentar o tamanho e a qualidade de *datasets* de treinamento (Shorten e Khoshgoftaar, 2019). Quanto as imagens de grande resolução, são necessárias etapas de pré-processamento em que as imagens são repartidas em imagens menores. Essas imagens menores acabam sendo convertidas para padrões de escala de cores populares e ficam sujeitas a perda de informação e resolução.

1.2 OBJETIVOS

Considerando a forma manual como o exame de HER2 é conduzido atualmente e a importância do diagnóstico precoce do câncer de mama, o presente trabalho tem como objetivo a criação de um sistema de auxílio ao diagnóstico do exame HER2. Para tal técnicas de redes neurais profundas, redes neurais e *data augmentation* foram combinadas.

Os objetivos específicos são:

- Otimizar as redes neurais e redes neurais profundas implementadas;
- Implementar e testar a eficiência de técnicas de *data augmentation*;

1.3 CONTRIBUIÇÕES

Ao cumprir os objetivos citados anteriormente, o presente trabalho terá como contribuição um novo sistema de auxílio ao diagnóstico ao exame HER2.

1.4 ORGANIZAÇÃO DO DOCUMENTO

O trabalho foi organizado em cinco capítulos: o Capítulo 2 discute trabalhos relacionados disponíveis na literatura e faz uma análise comparativa com o trabalho atual; o Capítulo 3 apresenta os materiais e métodos discutindo o *dataset* utilizado, etapa de pré-processamento e método proposto; o Capítulo 4 apresenta os resultados obtidos; o Capítulo 5 conclui o trabalho e discute trabalho futuros.

2 TRABALHOS RELACIONADOS

Esta seção tem como objetivo apresentar de maneira resumida trabalhos disponíveis na literatura com a mesma premissa do trabalho atual de auxílio ao diagnóstico do exame HER2 a partir de sistemas computacionais. Os trabalhos foram organizados de maneira cronológica e estão resumidos na Seção 2.1. Na Seção 2.2 é feita uma análise comparativa entre os trabalhos apresentados e a presente proposta. Para essa análise, foram levados em conta fatores como técnica computacional, tipo de lâminas, base de dados e se é feito uma pontuação de HER-2 para toda a lâmina.

2.1 RESUMO TRABALHOS RELACIONADOS

Vandenberghe et al. (2017) propõem uma abordagem para automaticamente determinar a expressão de HER2 em células tumorais. Nessa abordagem, primeiro é feita a deconvolução das imagens nos canais Hematoxilina e HER2. Após a deconvolução, o algoritmo de *watershed* é usado para detectar células. As células detectadas são então classificadas por uma rede neural convolucional (CNN) e por algoritmos clássicos de *machine learning*, no caso, *Random Forest* (RF) e *Support Vector Machine* (SVM). As células podem pertencer as classes: estromais, imunes, tumorais com forte super-expressão HER2, tumorais com super-expressão moderada de HER2, tumorais com super-expressão fraca de HER2, tumorais que não demonstram super-expressão de HER2 e artefatos (tecidos ou células que foram erroneamente identificadas com células). Ao final do experimento, o autor compara o desempenho dos algoritmos clássicos de *machine learning* com o algoritmo de rede neural convolucional. O desempenho obtido pela CNN foi melhor do que o atingido por métodos clássicos, atingindo uma acurácia geral de 0,78 em comparação a acurácia de 0,68 e 0,70 obtidos pelo RF e SVM respectivamente.

No trabalho de Cordeiro et al. (2018), cada imagem é repartida em segmentos chamados de *patches*. Cada *patch* passa por um processo de extração de características baseadas em cor. Esses descritores são classificados pelos algoritmos de aprendizado de máquina: SVM, *Decision Tree*, *Multilayer Perceptron* (MLP) e *K-Nearest Neighbors* (KNN). As predições feitas para cada *patch* são usadas para criar um histograma com a quantidade de cada classe. O histograma gerado é classificado novamente pelo mesmo pelos algoritmos de aprendizado de máquina e ficam responsáveis por determinar o rótulo para a lâmina toda. O método proposto mostrou 90% de concordância com o diagnóstico feito por patologistas treinados.

Singh e Mukundan (2018) propõem uma arquitetura de rede neural com um conjunto reduzido de características. Especificamente, foram usadas quatro tipos diferentes de características baseadas em variações de intensidade de cor, textura, variações morfológicas e estatísticas recolhidas de histogramas. O modelo final possui três camadas escondidas de mesmo tamanho. Com essa configuração os autores alcançaram acurácia de 91,10%, precisão média de 90,75% e *recall* de 91,50%.

Saha e Chakraborty (2018) propõem uma arquitetura de rede neural especificamente desenhada para o auxilio ao diagnostico de exames HER2. Nesta rede é utilizada camadas de convolução, de-convolução e memoria trapezoidal de longo prazo (TLSMTM). Para tal é feito a extração de *patches* seguido pela classificação na rede proposta em conjunto com a sumarização dos resultados obtidos para determinar a classe da lâmina de exame. Ao utilizar esta abordagem, os autores obtiveram 98,33% de acurácia, 96,64% de precisão, 96,79% de *recall* e 96,71% *F-score*.

Mukundan (2019) explora diferentes características apropriadas para a classificação de lâminas inteiras de HER2. Foram testadas características de intensidade, de textura baseadas no descritor *Uniform Local Binary Patterns* (ULBPs), baseadas na morfologia e por último características de natureza estatística de distribuição de intensidade geral. Além destas características, foram utilizados métodos para reduzir a redundância de informação E maximizar a separabilidade intra-classes. As características extraídas foram usadas com os classificadores: regressão logística e SVM. O classificador de maior acurácia foi o regressor alcançando acurácia, precisão, *recall* e F1-Score médio de 93%.

Khameneh et al. (2019) fazem a classificação de toda a lâmina histopatológica a partir de três etapas: identificação de regiões epiteliais e estromais, segmentação das regiões epiteliais com o uso de CNN e finalmente mescla de regiões classificadas e avaliação da pontuação da lâmina inteira. Ao final, se obteve 94% de acurácia durante a etapa de segmentação da imagem e 87% de acurácia para todo o processo de classificação.

Qaiser e Rajpoot (2019) trazem uma nova abordagem para a classificação de exames HER2. Neste trabalho, os autores consideram a tarefa como sendo sequencial e assim sugerem o uso de um modelo de *deep reinforcement learning*. Em sua abordagem, os autores conseguem filtrar regiões de interesse sem ter de analisar toda a lâmina histológica, mimetizando o comportamento de patologistas. Ao final dos experimentos, o modelo proposto obteve 79% de acurácia.

Anand et al. (2020) propõem uma classificação em cascata que utiliza três redes neurais convolucionais (CNNs) para obter a classe de uma lâmina. Primeiro, são extraídos *patches* rejeitando regiões do fundo da lâmina. Os núcleos celulares então são detectados utilizando uma CNN e passados para outro classificador de tumor ou não-tumor. Este classificador determina se os núcleos encontrados são manchas tumorais. Se sim, estas manchas são testadas para HER2+ ou HER2-. Finalmente é feita a aplicação de um limiar para determinar o status HER2 de um paciente. O método proposto alcançou um valor de 0,82 de AUC (área sob a curva) em um intervalo de confiança de 0,65 a 0,98.

La Barbera et al. (2020) utilizam um classificador em cascata que une técnicas como *multiple instance learning* (MIL) e redes profundas para a classificação de lâminas do exame HER2 preparadas com Hematoxilina-Eosina. Cada imagem é decomposta em segmentos, os quais são usados em duas abordagens: classificar todos os segmentos para HER2 ou filtrar apenas aqueles que contêm tecido tumoral usando uma CNN. Cada segmento é então classificado e recebe uma probabilidade de ser HER2+. Os resultados são agregados de três formas pela função de média, considerando todas as probabilidades, apenas aquelas maiores que 0,5 e se 35% dos segmentos possuem probabilidade igual ou maior a 0,5. Para todos os casos exceto o último o limiar para casos positivos de HER2 adotado foi de 0,5 e 0,6 respectivamente. Com as classes obtidas pelos três cálculos da média é escolhida a classe com maioria dos votos ou utilizando aprendizado tabular. Na abordagem tabular, segundo os autores, um classificador aprende os melhores valores de peso para cada um dos índices calculados. A abordagem de melhor resultado foi a que utilizou todos os segmentos junto ao voto majoritário totalizando 68% de acurácia, 57% de precisão, 88% de *recall* e 69% *F1-Score*.

Gamble et al. (2021) desenvolveram três sistemas independentes baseados em *deep learning* para determinar o status de ER/PR/HER2 para imagens contendo regiões focais de tecidos (*patches*) e lâminas inteiras. As imagens utilizadas foram coloridas com hematoxilina e eosina (H&E). Os modelos treinados foram avaliados usando lâminas anotadas por patologistas. Para a classificação usando *patches* o AUC alcançado foi de 0,939 em um intervalo de confiança de 95% para o exame ER, de 0,938 para PR e 0,813 para o exame HER2. Para a classificação de lâminas inteiras os resultados de AUC foram 0,86, 0,75 e 0,60 para ER/PR/HER2.

Tewary e Mukhopadhyay (2021) aplicaram a técnica de *transfer-learning* para classificar imagens de HER2 a partir de arquiteturas de redes convolucionais já existentes. Os autores testaram as arquiteturas *VGG16, VGG19, ResNet50, MobileNetV2,* e *NASNetMobile*. Para determinar a classe da lâmina inteira foi utilizada a moda dos resultados obtidos por cada classificador. Ao final do experimento, a melhor arquitetura observada foi a *VGG19* com acurácia de 93% para classificação de *patches* seguido por 98% de acurácia ao classificar toda a lâmina utilizado a moda das predições feitas.

No trabalho de (Ma et al., 2021) são utilizadas técnicas de *data augmentation* para gerar imagens artificiais do exame HER2 e assim obter redes neurais sem *overfitting*. Os autores geraram três bases artificiais baseadas no conjunto de dados utilizando as técnicas clássicas (operações de *flip*), *Deep Convolutional Generative Adversarial Network* (DCGAN), *Variational Auto-encoder* (VAE). Tendo as diferentes bases em mão, uma rede neural profunda foi treinada e testada com cada um das bases criadas. A técnica clássica obteve 94,29% e 88,85% de acurácia no treinamento e teste respectivamente. Os modelos treinados com as bases DCGAN e VAE alcançaram 94,87% e 93,3%, 94,43% e 92,45% de acurácia no treinamento e teste.

Wang et al. (2021) propõem um método de aprendizado fracamente supervisionado que combina as técnicas de *deep learning* e *transfer learning*. Os autores uma arquitetura baseada no modelo VGG16 e o treinaram com os *datasets*: CIFAIR-10 de objetos do dia-a-dia, IDC de imagens patológicas de câncer de mama e HAM1000 de imagens de lesões dermatoscópico. Os modelos foram avaliados com um *dataset* privado de imagens de IHC de HER2. A melhor arquitetura atingiu 97% de acurácia, precisão e *recall* e foi a pré-treinada com o *dataset* IDC. Ao utilizar somente as imagens de HER2, o desempenho foi de 88% de acurácia, 82% de precisão e 88% de *recall*.

Yang et al. (2022) propõem um *framework* baseado em *deep learning* multimodal para determinar a probabilidade de recorrência e metástase de tumores HER2 positivos. Nesse *framework* imagens inteiras de lâminas tingidas com Hematoxilina e Eosina são processadas por uma CNN e então as características extraídas são somadas as características clinicas do paciente como estágio tumoral, idade e resultados de outros exames. Após a avaliação da lâmina através de *patches* uma probabilidade é atrelada a cada um e então agregadas pela média. Se o total obtido por um paciente for igual ou superior a 0,5, o paciente é considerado com chance de recorrência. Ao final dos experimentos se obteve um valor AUC de 0,72 e com isso, os pesquisadores conseguiram concluir que o uso de imagens em conjunto com informações clínicas se mostrou promissor para determinar a chance de metástase de pacientes HER2 positivo.

Tewary e Mukhopadhyay (2022) propõem duas arquiteturas de redes neurais profundas para a classificação de lâminas HER2 em três possíveis classes: negativo, equívoco e positivo. A primeira rede proposta é um *framework* de *transfer learnig* feito com a rede *Xception* e a segunda, chamada de *AutoIHCNet*, foi desenvolvida pelos autores. O processo proposto é composto por uma etapa de pré-processamento no qual regiões de interesse são extraídas a partir das lâminas inteiras. Para determinar a classe da região de interesse é usada a função de moda para agregar as classificações feitas. Ao final dos experimentos, a rede *AutoIHCNet* desenvolvida pelos autores teve melhor desempenho se comparada com a rede *Xception*, alcançando 96% de acurácia ao classificar *patches* e 98% de acurácia ao classificar imagens. Segundo os autores, a rede *AutoIHCNet* teve melhor acurácia e menor tempo de processamento por imagem ao ser comparada com a rede *Xception*. Dessa forma concluiu-se que a nova rede poderia ser utilizada em aplicações em tempo real.

Wang et al. (2022) propõem uma arquitetura de rede neural residual para a classificação de lâminas imuno-histoquímicas de HER2. Os autores, ao utilizarem uma rede neural residual,

buscaram obter um sistema acurado mesmo com muitas camadas. O trabalho obteve 93% de acurácia, 91% de precisão, 92% de *recall* e 91% de *F1-Score*.

2.2 ANÁLISE COMPARATIVA

Os trabalhos resumidos na Seção 2.1 foram comparados entre si e a proposta atual. Foram levados em conta os seguintes fatores: técnica computacional, tipo de lâmina, base de dados, e se é feita a pontuação de HER2 para toda a lâmina. A Tabela 2.2 apresenta os resultados da análise comparativa descrita. O restante da Seção tem como objetivo aprofundar essa análise avaliando cada fator.

Ao compararmos os trabalhos encontrados, estes se dividem entre o uso de lâminas IHC ou H&E para determinar a classe de exames de HER2. Há pouca variabilidade de bases de dados usadas, sendo as bases propostas em Qaiser et al. (2018) e em Codella et al. (2019) as principais utilizadas em praticamente todos os trabalhos dessa área. Dessa forma, torna-se possível a comparação direta dos resultados entre os métodos propostos que compartilhem da mesma base de dados e o mesmo protocolo de avaliação (teste). Além deste ponto em comum, é possível também observar que todos os trabalhos utilizam técnicas de aprendizado de máquina. O uso de múltiplos classificadores é corriqueiro nos trabalhos analisados, o que pode indicar a presença de um *pipeline* de classificação ou a comparação entre os resultados obtidos por cada classificador. Os classificadores mais usados para a tarefa de pontuar exames HER2 foram CNNs e SVM.

Ao avaliarmos os trabalhos que utilizaram lâminas H&E para a classificação, aquele que obteve o melhor resultado foi o de (Anand et al., 2020). Nele, uma arquitetura de CNN em formato U é desenvolvida. Essa abordagem em U, segundo os autores, permite com que a rede extraia mais informações dos dados sem que haja a perda de conhecimento durante as iterações do algoritmo de aprendizado. Ao utilizarem essa abordagem, o resultado obtido foi de 98% de acurácia. Em contrapartida o trabalho de menor performance foi o de La Barbera et al. (2020), no qual é feita uma classificação em cascata, utilizando uma CNN para determinar quais *patches* possuem tecido cancerígeno. Os *patches* selecionados são classificados por outro modelo de CNN e então as predições são avaliadas utilizando funções estatísticas ou aprendizado tabular. Ao desenvolverem essa abordagem, os autores obtiveram apenas 68% de acurácia.

Para os trabalhos que utilizaram lâminas IHC, os que obtiveram o melhor resultado foram Saha e Chakraborty (2018), Tewary e Mukhopadhyay (2022) ambos com 98% de acurácia. No trabalho proposto por Saha e Chakraborty (2018) é desenvolvida uma CNN misturada com elementos de uma TLSTM. Já no trabalho de Tewary e Mukhopadhyay (2022), uma arquitetura de CNN é proposta a fim de obter o mesmo poder de classificação utilizando poucas camadas. Enquanto isso, o trabalho de Qaiser e Rajpoot (2019) obteve o pior resultado ao classificar lâminas IHC, em que apenas 79% de acurácia foi obtido. Os autores propõem um método para reduzir a área necessária para a avaliação da lâmina, assim mimetizando o comportamento de patologistas com técnicas de *deep reinforcement learning* e CNN para determinar regiões com maior valor de diagnóstico.

Ao avaliarmos os trabalhos com propostas similares aos objetivos propostos na Seção 1.2, podemos compreender de maneira aprofundada por qual caminho o trabalho pode seguir e os resultados esperados. Em nossa abordagem, propomos um *pipeline* de pré-processamento simplificado e uma classificação em duas etapas: uma em nível de *patches* e outra em nível de paciente. Na etapa de pré-processamento, seguimos os mesmos passos originalmente propostos por Cordeiro et al. (2018). Nele, as lâminas são repartidas em *patches* de mesmo tamanho e então apenas *patches* contendo tecido são levados para a etapa de classificação. Na primeira etapa de

classificação, utilizamos uma CNN para determinar a classe atrelada a cada *patch*. Em sequência um histograma com as classes identificadas é criado para cada paciente. Para determinarmos a classe de toda a lâmina utilizamos a informação obtida pelos histogramas como entrada para os algoritmos: KNN, SVM, MLP, *Random Forest*.

A fim de reduzir as chances de *overfitting* e *underfitting* aumentamos o tamanho da base de *patches* original em até 15 vezes de forma artificial usando técnicas de *data augmentation*. Ao final do processo, obtivemos uma base equilibrada em número de amostras por classe assim reduzindo as chances de *overfitting* e *underfiting*.

Trabalho	Téc. Computacionais	T.Lâmina	Base de dados	*Lâm.
Vandenberghe et al. (2017)	CNN, SVM, RF *watershed*	IHC	Qaiser et al. (2018)	Sim
Cordeiro et al. (2018)	SVM, RF, MLP, KNN	IHC	Qaiser et al. (2018)	Sim
Singh e Mukundan (2018)	RN	IHC	Qaiser et al. (2018)	Sim
Saha e Chakraborty (2018)	CNN, LSTM, TLSMTM	IHC	Qaiser et al. (2018)	Sim
Mukundan (2019)	SVM, ULBP e Regressão Logística	IHC	Qaiser et al. (2018)	Sim
Khameneh et al. (2019)	CNN, SVM, LBP e U-net	IHC	Qaiser et al. (2018)	Sim
Qaiser e Rajpoot (2019)	*deep reinforcement learning*, CNN,	IHC	Qaiser et al. (2018)	Sim
Anand et al. (2020)	CNN, U-net	IHC, H&E	Qaiser et al. (2018) e Clark et al. (2013)	Sim
La Barbera et al. (2020)	CNN, MIL, aprendizado tabular	H&E	Conde-Sousa et al. (2021)	Sim
Gamble et al. (2021)	CNN, *clusteri*ng	H&E	Clark et al. (2013) e dados privados	Sim
Tewary e Mukhopadhyay (2021)	CNN, TF	IHC	Qaiser et al. (2018)	Sim
Ma et al. (2021)	*data augmentation*, DCGAN, VAE, CNN	IHC	dados privados	Sim
Wang et al. (2021)	CNN, TF, aprendizado fracamente supervisionado	IHC	Janowczyk e Madabhushi (2016), Cruz-Roa et al. (2014), Tschandl et al. (2018), Codella et al. (2019)	Sim
Yang et al. (2022)	CNN, DL multimodal	H&E	Clark et al. (2013) e dados privados	Sim
Tewary e Mukhopadhyay (2022)	CNN, *decision fusion*	IHC	Qaiser et al. (2018)	Sim
Wang et al. (2022)	RNR	IHC	Marinelli et al. (2007)	Sim
Trabalho atual	CNN, *data augmentation*, RF, SVM, MLP, KNN	IHC	Qaiser et al. (2018)	Sim

Tabela 2.1: Caracterização dos trabalhos e análise comparativa. *Cl.Lâm - Classificação de toda a lâmina, RN - Rede Neural, RF - Random Forest, RNR - Rede neural recorrente, MIL - Multiple Instance Learning, DL - Deep learning, TF - Transfer Learning

3 MATERIAIS E MÉTODOS

Neste capítulo, serão abordados os materiais e métodos utilizados para alcançar o objetivo do trabalho de classificar lâminas do exame HER2 usando lâminas inteiras e redes neurais profundas.

3.1 *DATASET*

Em 2016, a Universidade de Warwick na Inglaterra criou uma competição para classificar lâminas inteiras do exame HER2 (Qaiser et al., 2018). Para isso a universidade criou e disponibilizou um *dataset* com os exames de HER2 em IHC e H&E de 86 pacientes diferentes, totalizando 176 imagens de lâminas inteiras. As lâminas de IHC são geralmente utilizadas para o diagnóstico de HER2 e as lâminas de H&E são utilizadas para auxiliar o patologista a determinar a região tumoral.

Os rótulos da base foram retirados de relatórios clínicos de caso que foram avaliados por patologistas no Hospital da Universidade de Nottingham (NHS Trust). Cada caso clínico foi avaliado por pelo menos dois patologistas especialistas. Foram definidas quatro classes baseadas no protocolo sugerido para o diagnóstico de HER2. Dentre elas, as classes 0 e 1 como HER2 negativo, a classe 2 como HER2 indeterminada e a classe 3 como HER2 positivo. A Figura 3.2 apresenta exemplos de imagens de *patches* para cada classe. Observe que não tivemos acesso a base que originou estes *patches*.

As lâminas foram digitalizadas com um *scanner* Hamamatsu NanoZoomer C9600, permitindo que as imagens sejam visualizadas de 94 a 940 pontos de magnificação, podendo ser comparada com as condições de diagnóstico de um microscópio padrão. As imagens são armazenadas a partir de uma estrutura piramidal em que a maior resolução possui 940 de magnificação ou seja o objeto real pode ser aumentado em até 940 vezes através da imagem. A Figura 3.1 é um exemplo de lâmina inteira disponível no *dataset* em que uma região foi ampliada. Devido ao formato de armazenamento e aquisição as imagens podem chegar a *gigabytes* e por isso o seu processamento e armazenamento fica limitado ao *hardware* disponível.

O trabalho atual utilizou somente as lâminas de IHC para a classificação de HER2. Do total das 86 lâminas disponibilizadas 51 delas foram colocadas a disposição para os competidores utilizarem para o treinamento e 34 foram mantidas em sigilo pela Universidade para a validação dos modelos enviados para a competição. Para o trabalho atual foi utilizado o conjunto de lâminas de treinamento da competição que foi dividido em treino e teste. Além disso uma das lâminas foi descartada por possuir pouca informação tecidual necessária para a classificação. A distribuição das lâminas da base entre as classes pode ser vista na Tabela 3.1.

3.2 PRÉ-PROCESSAMENTO

Devido a ordem de armazenamento das imagens ser em *gigabytes*, estratégias de pré-processamento para reduzir e tornar as lâminas processáveis foram utilizadas. As imagens foram repartidas em regiões não sobrepostas de resolução 250×250. Estas sub-imagens são chamadas de *patches* e ao final dessa etapa de processamento muitos *patches* não continham informação relevante para o diagnóstico por conterem pouca amostra tecidual ou somente o fundo da lâmina. Essas imagens contribuem para que o conjunto de dados seja mais ruidoso e de menor qualidade e além disso adicionam tempo de processamento necessário para determinar o

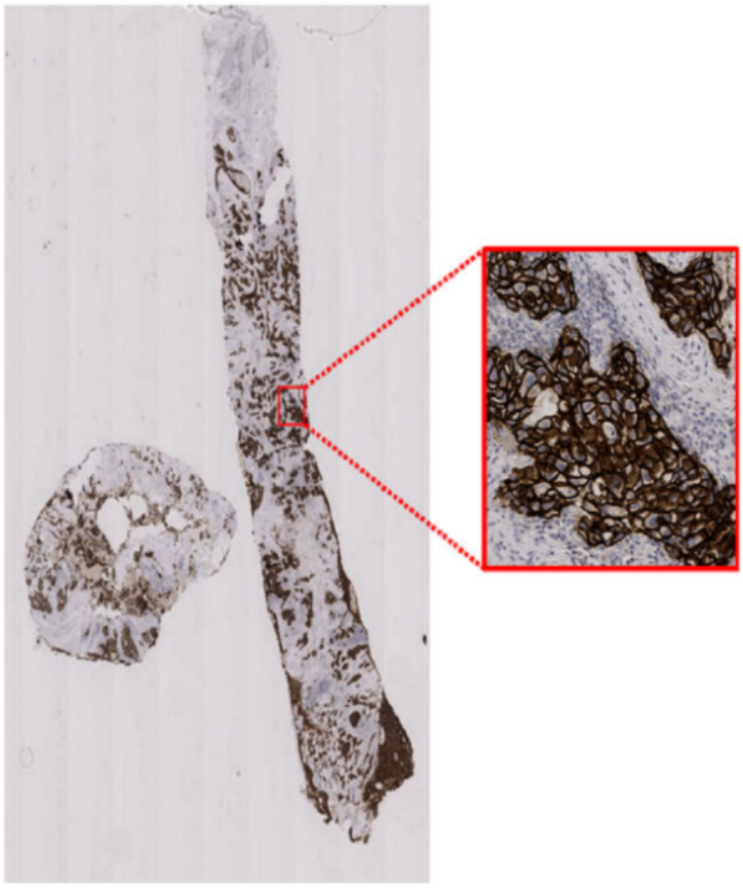


Figura 3.1: Exemplo de lâmina inteira disponível no *dataset*

Tabela 3.1: Distribuição de classes entre pacientes da base de dados disponibilizada pela Universidade Warwick. (Qaiser et al., 2018)

Classe	**Treino**	**Teste**
0	9	4
1	15	7
2	14	13
3	13	10
Total	51	34

rótulo de uma lâmina. Dessa forma os *patches* que foram considerados sem informação relevante foram removidos. Para determinar quais possuíam região histológica suficiente foi utilizado um algoritmo de binarização baseado em *threshold*. Todas as etapas de pré-processamento citadas até então seguem os mesmos passos feitos no trabalho de Cordeiro et al. (2018)

Após todo o processo de pré-processamento descrito se obteve ao todo 230.809 *patches*. A criação desses *patches* ocasionou outro problema: a falta de rótulos para as novas imagens. Na base de dados original, tem-se o rótulo de todos os pacientes, mas ao repartir uma lâmina, talvez um *patch* não tenha o mesmo rótulo atribuído a lâmina inteira. A solução encontrada foi a de rotular, com a ajuda de um especialista, os *patches*. Entretanto a quantidade de *patches* gerados também foi um empecilho pois rotulá-los em tempo hábil se tornou uma tarefa praticamente impossível humanamente. Foram então selecionados os *patches* de maior representatividade de cada classe, ou seja, aqueles que podem ser utilizados como exemplo para a classe que pertencem.

Além das quatro classes originárias do exame HER2, foi incluída uma classe de ruído composta por imagens do fundo da lâmina ou por segmentos de tecido sem a coloração ideal. A Figura 3.2 mostra exemplos de*patches* para cada classe. Após todo o processo descrito, foi obtida uma base de dados com 1867 *patches* divididos entre cinco classes. A distribuição entre classes da base pode ser vista na Tabela 3.2.

Ao analisarmos a Tabela 3.2 há poucas imagens para a criação de um modelo de rede neural profunda. Tais redes neurais por possuírem muitos parâmetros treináveis tornam-se imprescindíveis para a boa performance do modelo em bases ricas em dados. A falta de dados pode ocasionar problemas de *overfitting* durante o treinamento, efeito que reduz a capacidade de generalização de redes neurais.

Outra característica inerente da base final obtida é a falta de balanceamento entre as classes. A classe 0 é a de menor representatividade com 277 amostras representando apenas 14% do total de imagens enquanto a classe 3 possui 420 amostras representando 22% da base, ou seja, praticamente o dobro de amostras que as da classe 0. Bases desbalanceadas também implicam em problemas de performance em que novamente a capacidade de generalização do modelo é reduzida.

Tabela 3.2: Distribuição da base de *patches* rotuladas pelo patologista especialista.

Classe	**Patches**	**Porcentagem**
0	277	14,83%
1	387	20,72%
2	419	22,44%
3	420	22,50%
Ruído	364	19,50%
Total	1867	100%

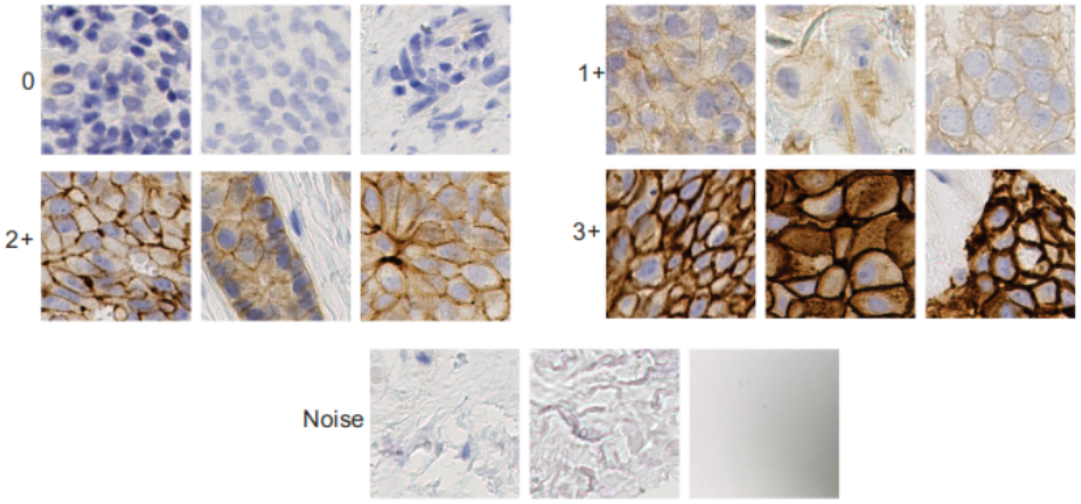


Figura 3.2: Exemplos de *patches* para cada classe da base de dados utilizada. 0/1+ HER2 negativo, 2+ indeterminado, 3+ HER2 positivo e Ruído.

3.3 *DATA AUGMENTATION*

Tendo em vista a distribuição de amostras por classes apresentado na Tabela 3.2 e os pontos levantados, sobre a base de dados final, nesta seção apresentamos o uso de estratégias para aumentar o conjunto de dados visando a conclusão do objetivo do trabalho atual. Para mitigar o problema da falta de dados, abordagens de *data augmentation* tem se popularizado na área. *Data augmentation* trata-se de um conjunto de técnicas capazes de aumentar o tamanho e a qualidade de *datasets* de treinamento (Shorten e Khoshgoftaar, 2019).

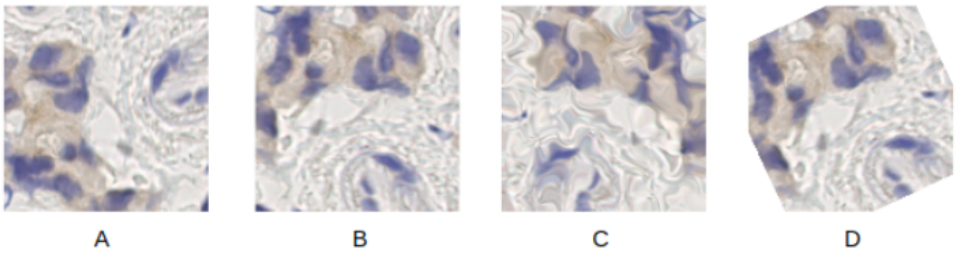


Figura 3.3: Imagens aumentadas. A - imagem original, B - resultado após flip, C - resultado após transformação elástica, D - resultado de rotação.

Desta forma técnicas de operações de *flip* horizontal (esquerda para a direita), *flip* vertical (topo para baixo), rotações nos ângulos de: 90, 180, 270, 45, 99, 178, 127 graus, transposta da imagem e transformação elástica foram usadas para gerar novas amostras a partir da base de dados original. A Figura 3.3 mostra estas operações aplicadas a uma imagem de exemplo.

Essas operações foram aplicadas em todos os *patches* da base original a fim de obter uma base balanceada com mais amostras por classe. Dessa forma, foram confeccionadas duas bases de dados "aumentadas", chamadas de *DataAug1* e *DataAug2*. A primeira aumentou a base de dados original em até sete vezes e a segunda em até 15 vezes. Para se obter uma base com uma distribuição homogênea de amostras entre as classes, algumas classes foram aumentadas mais vezes. A Tabela 3.3 apresenta a distribuição das bases de dados aumentadas em comparação com a base original.

Ao analisarmos a Tabela 3.3 as bases aumentadas possuem um total de amostras balanceado entre as classes. A classe de maior aumento dentre todas foi a classe 0 em que foi multiplicada em até 15 vezes em *DataAug2* para permitir balanceamento entre as classes. Enquanto as classes 2 e 3 de maior representatividade foram aumentadas menos vezes. As novas bases confeccionadas foram utilizadas para o treinamento do classificador com o intuito de reduzir chances de *overfitting* e se obter melhor desempenho do modelo.

Tabela 3.3: Distribuição das bases original e aumentadas.

Classe	**Original**	***DataAug1***	***DataAug2***
0	277	2493 (x9)	4155 (x15)
1	387	2709 (x7)	4257 (x11)
2	419	2514 (x6)	4190 (x10)
3	420	2520 (x6)	4200 (x10)
Ruído	364	2548 (x7)	4368 (x12)
Total	1867	12584	21170

3.4 MÉTODO PROPOSTO

O presente trabalho se propõe a auxiliar o diagnóstico do exame HER2 a partir de um sistema computacional baseado em visão computacional. Para tal foi utilizada a abordagem proposta por Cordeiro et al. (2018) em que há a classificação a nível de *patch* e a nível de paciente. Para a classificação a nível de *patch* propomos o uso de uma rede neural convolucional (CNN). Após a classificação individual de todos os *patches* de um paciente, estes são agregados por meio de um histograma. O histograma contém o total de *patches* classificados para cada classe. Para a classificação a nível de paciente esse histograma gerado é usado como entrada para classificadores clássicos de *machine learning*, sendo eles: *Suport Vector Machines* (SVM), *Random Forest* (RF), *K-Nearest Neighbors* (KNN) e *Multi-Layer Perceptron* (MLP). Estes classificadores são então responsáveis por determinar a classe que o paciente pertence. É importante ressaltar que na classificação a nível de paciente as classes possíveis são: HER2+, HER2- e indeterminado. Neste caso, o pipeline de processamento de imagens adotado pelo presente trabalho pode ser resumido na Figura 3.4.

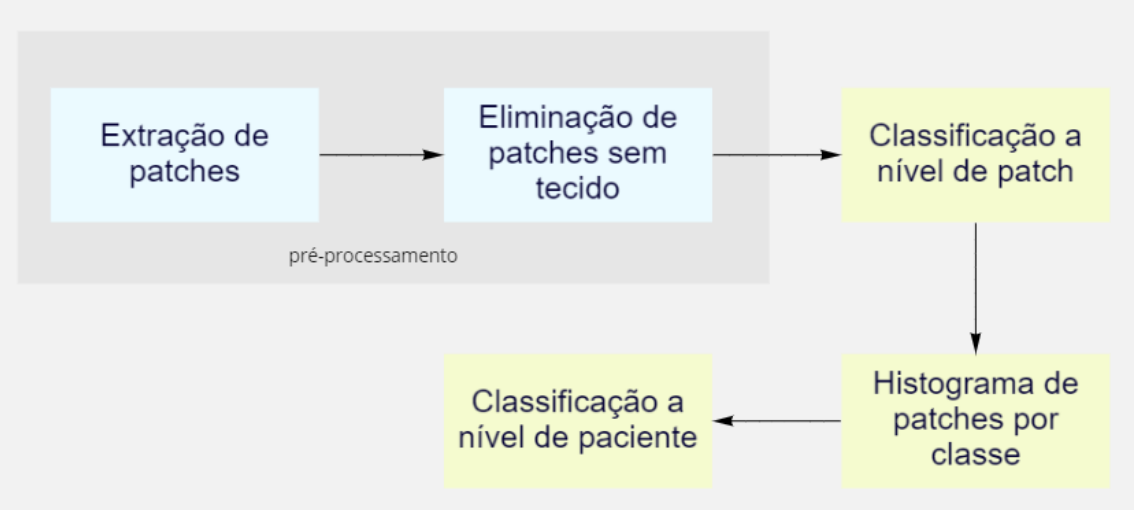


Figura 3.4: *Pipeline* de processamento de imagens proposto.

A principal diferença da proposta do trabalho atual em relação à proposta do trabalho de Cordeiro et al. (2018) é a classificação em nível de *patches*. Em (Cordeiro et al., 2018) são utilizados descritores de cor e textura para criar vetores de características que são classificados por algoritmos clássicos de *machine learning*. Enquanto o trabalho presente se propõe a utilizar uma rede neural profunda para a extração de características e classificação de *patches*. Além disso há o udo de técnicas de *data augmentation* para melhorar a qualidade e o total de amostras da base de dados usada. Para a classificação a nível de paciente se manteve os mesmos algoritmos de *machine learning* utilizados por Cordeiro et al. (2018).

3.5 ARQUITETURA DOS CLASSIFICADORES

Para definir uma arquitetura de CNN, foram feitas avaliações exaustivas a fim de obter os melhores hiperparâmetros para a tarefa de classificação das imagens. A arquitetura de CNN de maior eficiência encontrada possui sete camadas de convolução, sendo as três primeiras com 32 filtros e *kernel* de tamanho 3×3, seguido por uma camada com 64 filtros e *kernel* de mesmo tamanho, terminando com duas camadas de convolução de 64 filtros com *kernel* de tamanho 7×7 e uma última camada de convolução com 64 filtros e kernel de tamanho 3×3. Todas as

camadas de convolução utilizaram como função de ativação a função *relu*. Após as camadas de convolução, o modelo possui uma camada com 125 neurônios e outra com 85 com a função de ativação *relu* e uma última camada com cinco neurônios (total de classes) com ativação sigmoide. Durante o treinamento foi utilizado uma taxa de aprendizado de 10^{-3} e um total de 20 épocas. Também foi utilizado o *callback* de *early stop*, responsável por monitorar a *loss* da etapa de validação, com paciência de três épocas e um *baseline* de 0,005.

Os parâmetros dos classificadores SVM, RF, KNN e MLP foram sempre mantidos como os *default* da biblioteca *scikit-learning*. Para comparar o desempenho entre os classificadores, foram utilizadas métricas de avaliação como acurácia, *recall*, precisão e F1-Score

4 RESULTADOS E DISCUSSÃO

O presente capítulo tem como finalidade apresentar os resultados obtidos na classificação em nível de *patch* e de lâmina e comparar com os resultados disponíveis na literatura. A Seção 4.1 apresenta a política de testes usadas para a avaliação dos modelos, a Seção 4.2 apresenta e discute os resultados obtidos em nível de *patch* e a Seção 4.3 os resultados em nível de lâmina, a Seção 4.4 compara os resultados obtidos com os demais trabalhos disponíveis na literatura.

4.1 METODOLOGIA DE AVALIAÇÃO

Para testar o método proposto de classificação em nível de *patch* e classificação em nível de paciente, foi criada uma metodologia de avaliação. As bases artificiais *DataAug1* e *DataAug2* de *patches* foram utilizadas separadamente para o treinamento de modelos de CNN. Um terceiro modelo treinado com a base sem *data augmentation* foi criado para servir como controle. Em todos estes casos, 20% dos dados foram separados para o conjunto de validação que visa interromper o processo de aprendizado visando evitar a especialização do modelo mantendo a generalização. Todos os três modelos foram avaliados no mesmo conjunto de teste.

Para garantir que os valores de performance obtidos não dependessem da seleção de sub-conjuntos de treinamento e teste, o esquema de avaliação descrito acima, foi feito utilizando um esquema de validação cruzada chamado *K-fold* estratificado. Neste esquema, uma base de dados é dividida em K porções ou *folds* de forma que cada uma delas tenha a mesma distribuição estatística entre as classes vista na base original. Para o presente trabalho foi utilizado k como 5 e a divisão foi feita a partir dos rótulos a nível de paciente. Dessa forma, todos os *patches* de um paciente só podem ser encontrados na etapa de treinamento ou de teste em uma mesma *fold*. Por exemplo, os pacientes 1, 2 e 3 foram utilizados no treinamento e os pacientes 4, 5 e 6 foram utilizados no teste para a *fold* 1 enquanto na *fold* 2 os pacientes 4, 5 e 6 foram utilizados no treinamento e os pacientes 1, 2 e 3 no teste.

A Tabela 4.1 mostra o total de pacientes pertencentes a cada *fold* utilizada para os experimentos. O mesmo conjunto de *folds* utilizada para treinamento e teste a nível de *patch* foram repetidos para treinar e testar os classificadores a nível de paciente. O total de pacientes utilizados no treinamento e no teste variam, pois a divisão feita na base procura manter cada *fold* balanceado. Por este motivo, há discrepância no total de pacientes usado entre cada *fold*.

Tabela 4.1: Total de pacientes por *fold* criada.

Fold	Pacientes Treino	Pacientes Teste
0	40	12
1	41	11
2	41	11
3	42	10
4	44	8

4.2 RESULTADOS EM NÍVEL DE *PATCH*

A arquitetura de rede neural convolucional mostrada na Seção 3.5 foi treinada e avaliada seguindo a metodologia descrita na Seção 4.1. Ao final dos testes realizados se obteve os resultados apresentados nas Tabelas 4.2 e 4.3. A Tabela 4.2 contém os valores de *loss*, acurácia média e desvio padrão nas etapas de treino, validação e teste. A Tabela 4.3 contém as métricas de precisão, *recall* e *F1-Score* feitas no teste.

Tabela 4.2: Acurácia e função *loss* da CNN por cada etapa de treinamento. O - *dataset Original*, A1 - *dataset DataAug1*, A2 - *dataset DataAug2*

	Treino			Validação			Teste		
	O	*A1*	*A2*	*O*	*A1*	*A2*	*O*	*A1*	*A2*
Acurácia	0,26	0,97	0,95	0,12	0,88	0,85	0,27	0,87	0,83
Dp Acur.	0,09	0,02	0,08	0,16	0,05	0,11	0,04	0,03	0,16
Loss	0,19	0,01	0,01	0,19	0,03	0,03	0,19	0,04	0,05
Dp Loss.	0,02	0,005	0,01	0,02	0,01	0,03	0,004	0,01	0,04

Ao analisar a Tabela 4.2, é possível ver que o modelo treinado apenas com o conjunto de dados *Original* teve uma pontuação menor ao comparar-lo com os demais modelos. Ao utilizar somente a base *Original* a acurácia média foi de 0,26 (treino), 0,12 (validação) e 0,27 (teste). Houve pouca diferença entre o visto no treino e teste indicando um padrão de certa forma consistente. Entretanto ao observar o desvio padrão que foi de 0,09 (treino), 0,16 (validação) e 0,04 (teste) há uma flutuação considerável no treinamento e validação do modelo. Outra métrica importante de se destacar são os valores médios obtidos de *loss* que foram de 0,19 para todas as etapas avaliadas. Esse valor é considerado alto já que expressa o erro do modelo. Dessa forma a flutuação na acurácia e valores altos de *loss* deixam evidente os prejuízos causados pela falta de dados.

Os modelos treinados com as bases aumentadas tiveram uma pontuação maior que a obtida pelo modelo treinado apenas com a base de dados *Original*. Em *DataAug1* a acurácia foi de 0,97 (treino), 0,88 (validação) e 0,87 (teste) e a *loss* foi de 0,01 (treino), 0,03 (validação) e 0,04 (teste). Há uma queda de 0,10 de acurácia entre as etapas de treino e teste que foi observada na etapa de validação. Essa queda pode ser um indicador de *overfitting* do modelo causada tanto por excesso de épocas de treinamento ou pelas transformações feitas para a confecção das novas imagens. Quanto a *loss* há um aumento de 0,03 entre treino e teste. Os valores obtidos se mostram estáveis por possuírem um desvio padrão baixo tanto para a acurácia quanto para a *loss*.

Em *DataAug2* o comportamento de perda de acurácia no treino e teste se repete como visto em *DataAug1*. A base teve acurácia média de 0,95 (treino), 0,85 (validação) e 0,83 (teste). Para *DataAug2* a diferença de acurácia entre treino e teste foi de 0,12 se mostrando maior do que a vista em *DataAug1*. Quanto ao desvio padrão da acurácia, este foi o maior do que o visto em *DataAug1* sendo igual a 0,08 (treino), 0,11 (validação) e 0,16 (teste) para a base *DataAug2*. A *loss* de *DataAug2* no treino e validação foi igual ao visto em *DataAug1*, porém no teste essa foi de 0,04 apresentando um aumento de 0,01. De maneira geral ambas as bases aumentadas tiveram pontuação próxima mas a base *DataAug1* obteve valores menores de desvio padrão entre as métricas avaliadas e uma *loss* média menor.

Devido as bases aumentadas terem pontuação próxima entre si foi realizado o teste estatístico t-pareado para assim determinar se há significância estatística entre os resultados

obtidos. Para o teste, foi considerado um intervalo de confiança de 95% e valor de significância de 5%. O teste t-parado entre a base *Original* e a base *DataAug1* teve valor p menor que 0,0001, já para o teste entre a base *Original* e *DataAug2* o valor p foi de 0,0052 e por último o teste entre as bases aumentadas teve valor p de 0,6414.

Os testes mostram que há significância estatística ao compararmos as bases aumentadas com a base *Original*, mas entre as bases aumentadas não há significância estatística nos resultados obtidos. Dessa forma, podemos afirmar com 95% de confiança que as diferenças vistas entre as pontuações das bases aumentadas foi devido ao acaso e que as diferenças entre as bases aumentadas e a base *Original* não foram ao acaso.

Tabela 4.3: Métricas de classificação para cada base testada obtida na etapa de teste. A letra P para a métrica precisão, R para *recall* e F1 para *F1 Score*.

Dataset/ Classe	*Original*			*DataAug1*			*DataAug2*		
	P	R	F1	P	R	F1	P	R	F1
0	0,47	0,24	0,31	0,95	0,96	0,95	0,90	0,91	0,90
1	0,60	0,04	0,07	0,79	0,67	0,72	0,84	0,62	0,71
2	0,26	0,81	0,40	0,73	0,87	0,79	0,75	0,91	0,82
3	0	0	0	0,99	0,93	0,95	0,98	0,95	0,96
4	0,23	0,23	0,23	0,97	0,97	0,97	0,91	0,94	0,92

Para a avaliação dos modelos também foram calculadas as métricas de precisão, *recall* e *F1-Score* referentes a etapa de teste, cujos valores estão disponíveis na Tabela 4.3. Ao analisarmos os resultados da base *Original*, salta aos olhos o valor 0 para todas as métricas na classe 3. Esse resultado levou à acurácia baixa vista anteriormente e demonstra a falta de generalização do modelo para essa classe específica. As demais classes apresentam valores variados de precisão e *recall*, mantendo maiores valores de precisão para as classes 0 e 1 e valores maiores de *recall* para a classe 2. Há uma falta de equilíbrio entre as duas métricas que é evidenciada pelos valores de *F1-Score*. Para a classe 4 todas as métricas foram iguais.

Para a base *DataAug1*, os valores calculados foram maiores do que os vistos na base *Original*. A proporção entre precisão e *recall* é mais balanceada do que a vista na base *Original* mostrando um *F1-Score* acima de 0,90 para as classes 0, 3 e 4. Para as classes 1 e 2 há uma queda na métrica causada pela diminuição na precisão. A classe de maior pontuação foi a classe 4 obtendo 0,97 em todas as métricas e a classe de menor pontuação foi a classe 1 que obteve 0,79, 0,67 e 0,72 de precisão, *recall* e *F1-Score*, respectivamente.

As métricas calculadas para a base *DataAug2* foram próximas aos valores vistos para a base *DataAug1*. Para *DataAug2* as classes 0, 3, 4 tiveram valores de F1-Score igual ou acima a 0,90 assim como visto em *DataAug1*. Os valores de F1-Score diminuem quando se avalia as classes 1 e 2 sendo a métrica igual a 0,71 para a classe 1 e 0,82 para a classe 2. Para a classe 1 a precisão foi maior do que o *recall* e para a classe 2 o acontecido é reverso tendo o *recall* maior que a precisão. Dessa forma podemos observar o comportamento similar entre as duas bases aumentadas.

O teste estatístico t-pareado foi repetido novamente para as métricas apresentadas na Tabela 4.3. Como no teste anterior, foi utilizado intervalo de confiança de 95% e valor *alpha* de 5%. Os testes realizados indicaram significância estatística ao compararmos as métricas das bases aumentadas com a base *Original*. Ao compararmos as métricas entre as bases aumentadas não há significância estatística. Dessa forma podemos afirmar com 95% de confiança que os

resultados mostrados pelas bases aumentadas ao compararmos com a base *Original* não foram ao acaso, mas os resultados diferentes entre as bases aumentadas foram causados pelo acaso.

4.3 RESULTADOS EM NÍVEL DE LÂMINA

Os *patches* classificados na etapa anterior por um modelo de CNN seguem pelo *pipeline* proposto para a classificação em nível de lâmina. Para classificar as lâminas, foi utilizada a mesma abordagem proposta por Cordeiro et al. (2018) em que é feito o cálculo de um histograma para cada lâmina baseado na proporção de *patches* pertencentes a cada classe. Os histogramas são utilizados para o treinamento e teste de algoritmos de *machine learning* que são responsáveis por determinar o rótulo final da lâmina inteira de HER2.

Para a etapa de classificação em nível de lâmina foram utilizados os classificadores: SVM, *Random Forest*, KNN e MLP. A política de testes adotada foi a mesma da etapa anterior, com o uso do algoritmo *k-fold* estratificado com k igual a 5. A Tabela 4.4 apresenta a acurácia obtida com os classificadores para cada base de treinamento.

Tabela 4.4: Acurácia média de todas as *folds* de cada classificador.

Classificador	**Original**		**DataAug1**		**DataAug2**	
	Média	DP	Média	DP	Média	DP
SVM	0,30	0,18	0,76	0,07	0,72	0,08
MLP	0,28	0,16	0,63	0,17	0,74	0,10
Random Forest	0,32	0,21	0,76	0,08	0,75	0,11
KNN	0,27	0,18	0,70	0,17	0,65	0,14

Ao examinar os valores apresentados na Tabela 4.4 para a base *Original*, os classificadores tiveram desempenho muito próximos entre si ao analisarmos a acurácia média. Todos os classificadores tiveram acurácia média em torno de 0,30, com destaque para o *Random Forest* que alcançou 0,32 de acurácia média. Entretanto esse classificador também foi aquele com maior desvio padrão dentre os demais sendo igual a 0,21.

Analisando os valores referentes a base *DataAug1*, os classificadores tiveram acurácia média em torno de 0,70. Os classificadores SVM e *Random Forest* tiveram performance igual a 0,76 mas valores de desvio padrão diferentes. Para o SVM o desvio padrão foi de 0,07 e para o *Random Forest* foi de 0,08. Os classificadores MLP e KNN tiveram o mesmo valor de desvio padrão de 0,17, sendo o maior entre os classificadores.

Para a base *DataAug2* o desempenho foi próximo ao visto na base *DataAug1*. Os classificadores tiveram acurácia média próximas sendo de 0,72 (SVM), 0,74 (MLP), 0,75 (*Random Forest*) e 0,65 (KNN). O desvio padrão foi próximo também exceto para o KNN que teve o maior desvio padrão do grupo. Os classificadores de destaque foram o *Random Forest* e a MLP por terem os maiores valores de acurácia e menores valores de desvio padrão.

Para avaliar estatisticamente os resultados obtidos, o teste estatístico t-pareado foi novamente aplicado. O teste levou em consideração um intervalo de confiança de 95% e valor *alpha* de 5%. Após aplicar o teste nos pares podemos concluir que apenas o classificador SVM ao utilizar a base de dados *DataAug2* obteve resultados estatisticamente significativos ao ser comparado com os valores obtidos pela base Original. Os demais classificadores não mostraram significância estatística nos resultados de seus testes. Dessa forma, os valores obtidos, menos pelo SVM, estão atrelados ao acaso.

4.4 COMPARAÇÃO COM A LITERATURA

O trabalho atual, como visto na Seção 4.4, se baseia no conjunto de passos primeiramente proposto por (Cordeiro et al., 2018). Em ambos os trabalhos é utilizado o mesmo conjunto de dados, a mesma etapa de pré-processamento (na qual as lâminas do exame HER2 são repartidas em *patches*) e a mesma forma de classificação da lâmina inteira a partir de histogramas de frequência. O que difere o trabalho atual do feito por Cordeiro et al. (2018) é a etapa de extração de características e o uso da técnica de *data augmentation*. No atual trabalho, a etapa de extração de características é feita por uma CNN e em (Cordeiro et al., 2018) por descritores variados. Tendo em vista a similaridade entre os trabalhos se mostra necessária a comparação entre os resultados obtidos pelos *pipelines* de cada um deles.

Para a classificação de lâminas, (Cordeiro et al., 2018) teve o melhor resultado obtido ao utilizar as características extraídas pelo modelo Resnet50 em conjunto com uma MLP, e ao utilizar para a classificação destas características o classificador *Decision Tree*. Na fase de teste da etapa de extração de características, o trabalho de (Cordeiro et al., 2018). alcançou uma acurácia de 89% (desvio padrão não informado em (Cordeiro et al., 2018)). Na mesma fase do trabalho atual, foi alcançado 87% de acurácia ao utilizar uma CNN, com desvio padrão de 0,03. Na etapa de classificação de lâminas, o pipeline de (Cordeiro et al., 2018) alcançou 90,20% de acurácia na etapa de teste (desvio padrão não informado por (Cordeiro et al., 2018)). Já no trabalho atual foi alcançado 76% de acurácia (desvio padrão de 0,07) ao utilizar o classificador SVM, também na etapa de teste.

Os resultados obtidos por ambos os *pipelines* foram próximos na etapa de extração de características mas tiveram uma grande divergência nos resultados da etapa seguinte. Esta disparidade pode se dar pelas características obtidas anteriormente ou até mesmo pela ordem e dados utilizados em cada etapa de desenvolvimento do modelo. Também é importante ressaltar que em (Cordeiro et al., 2018), as métricas que avaliam o erro obtido são reportadas somente na segunda etapa, enquanto que na primeira etapa é somente reportada a acurácia do modelo. Isso dificulta a análise e comparação do *pipeline* do trabalho referenciado com outros trabalhos relacionados.

5 CONCLUSÃO

Neste trabalho de conclusão de curso, foi apresentado um sistema de auxilio ao diagnóstico de lâminas inteiras do exame HER2. Como visto, o diagnóstico atualmente é feito manualmente por especialistas da área que podem cometer erros devido a fadiga e outras causas. O auxílio ao diagnóstico pode contribuir para a redução desses erros e também possibilitar o processamento mais rápido das amostras, dado que o HER2 é um exame corriqueiro durante o diagnóstico do câncer de mama.

Para a implementação do sistema, foi utilizado como base o *pipeline* de classificação de lâminas inteiras proposto por Cordeiro et al. (2018). Nele as lâminas passam por uma etapa de pré-processamento antes da classificação em que são divididas em *patches*. A classificação ocorre em duas etapas: em nível de *patch* e em nível de lâmina. Na classificação em nível de *patch* cada *patch* passa por um processo de extração de características que visa utilizar descritores de cor e intensidade para representar-los. Em seguida é realizada a classificação por algoritmos de *machine learning* e um histograma de frequência de cada classe é calculado. Com este histograma é feita a classificação em nível de lâmina utilizando também algoritmos de *machine learning*.A extração de características feita com descritores de cor e intensidade foi substituído por um modelo de CNN, além de utilizar técnicas de *data augmentation* para enriquecer o conjunto de dados utilizado.

Ao utilizar técnicas de *data augmentation* foi possível criar um conjunto de dados mais rico e de maior volume assim permitindo o uso de redes neurais profundas. Ao treinarmos um modelo somente a base de dados Original os resultados obtidos não foram tão satisfatórios quanto os vistos nos modelos que foram treinados com as bases de dados aumentadas. Os resultados alcançados pelas bases aumentadas mostram como o uso de técnicas de *data augmentation* são capazes de melhorar o desempenho do modelo. Em problemas em que se há falta de dados essas técnicas tornam-se essenciais para evitar o *underfitting*.

Dessa forma levando-se em consideração o que foi discutido ao longo do texto, as técnicas utilizadas neste trabalho se mostraram promissoras para a criação do sistema de auxílio ao diagnóstico do exame HER2. Entretanto, ainda é necessário um estudo mais abrangente sobre aspectos como o tamanho da base de dados, outras técnicas de *data augmentation* propícias para imagens biomédicas, entre outros.

Como trabalhos futuros, pode se aprofundar em técnicas avançadas de *data augmentation* (redes neurais adversárias); transferência de conhecimento entre redes neurais para o estudo da extração de características; aplicação de outros paradigmas de *machine learning*, como o *multiple instance learning* (MIL).

REFERÊNCIAS

Anand, D., Kurian, N. C., Dhage, S., Kumar, N., Rane, S., Gann, P. H. e Sethi, A. (2020). Deep learning to estimate human epidermal growth factor receptor 2 status from hematoxylin and eosin-stained breast tissue images. *Journal of Pathology Informatics*, 11:19.

Cirqueira, M. B., Moreira, M. A. R., Soares, L. R. e Freitas-Júnior, R. (2011). Subtipos moleculares do câncer de mama. *Femina*.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L. e Prior, F. (2013). The cancer imaging archive (tcia): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057.

Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., Kittler, H. e Halpern, A. (2019). Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic).

Conde-Sousa, E., Vale, J., Feng, M., Xu, K., Wang, Y., Mea, V. D., Barbera, D. L., Montahaei, E., Baghshah, M. S., Turzynski, A., Gildenblat, J., Klaiman, E., Hong, Y., Aresta, G., Araújo, T., Aguiar, P., Elo, C. e Polónia, A. (2021). Herohe challenge: assessing her2 status in breast cancer without immunohistochemistry or in situ hybridization.

Cordeiro, C. Q., Ioshii, S. O., Alves, J. H. e Oliveira, L. F. (2018). An automatic patch-based approach for her-2 scoring in immunohistochemical breast cancer images using color features.

Cruz-Roa, A., Basavanhally, A., González, F., Gilmore, H., Feldman, M., Ganesan, S., Shih, N., Tomaszewski, J. e Madabhushi, A. (2014). Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. Em Gurcan, M. N. e Madabhushi, A., editores, *Medical Imaging 2014: Digital Pathology*, volume 9041, páginas 1–15. International Society for Optics and Photonics, SPIE.

Ferlay, M, E., F, L., M, C., L, M., M, P., A, Z., I, S. e F, B. (2020). Global cancer observatory: Cancer today.

Gamble, P., Jaroensri, R., Wang, H., Tan, F., Moran, M., Brown, T., Flament-Auvigne, I., Rakha, E. A., Toss, M., Dabbs, D. J., Regitnig, P., Olson, N., Wren, J. H., Robinson, C., Corrado, G. S., Peng, L. H., Liu, Y., Mermel, C. H., Steiner, D. F. e Chen, P.-H. C. (2021). Determining breast cancer biomarker status and associated morphological features using deep learning. *Commun Med*, 1(14).

Instituto Nacional do Câncer (INCA) - Ministério da Saúde (2021). Tipos de câncer - câncer de mama. `https://www.inca.gov.br/tipos-de-cancer/cancer-de-mama`.

Jafari, S. H., Saadatpour, Z., Salmaninejad, A., Momeni, F., Mokhtari, M., Nahand, J. S., Rahmati, M., Mirzaei, H. e Kianmehr, M. (2018). Breast cancer diagnosis: Imaging techniques and biochemical markers. *Journal of Cellular Physiology*, 233(7):5200–5213.

Janowczyk, A. e Madabhushi, A. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7:29–29. 27563488[pmid].

Kaufman, P. A., Bloom, K. J., Burris, H., Gralow, J. R., Mayer, M., Pegram, M., Rugo, H. S., Swain, S. M., Yardley, D. A., Chau, M., Lalla, D., Yoo, B., Brammer, M. G. e Vogel, C. L. (2014). Assessing the discordance rate between local and central her2 testing in women with locally determined her2-negative breast cancer. *Cancer*, 120(17):2657–2664.

Khameneh, F. D., Razavi, S. e Kamasak, M. (2019). Automated segmentation of cell membranes to evaluate her2 status in whole slide images using a modified deep learning network. *Computers in Biology and Medicine*, 110:164–174.

La Barbera, D., Polónia, A., Roitero, K., Conde-Sousa, E. e Della Mea, V. (2020). Detection of her2 from haematoxylin-eosin slides through a cascade of deep learning classifiers via multi-instance learning. *Journal of Imaging*, 6(9).

Li, Z., Liu, F., Yang, W., Peng, S. e Zhou, J. (2021). A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, páginas 1–21.

Loibl, S. e Gianni, L. (2017). Her2-positive breast cancer. *The Lancet*, 389(10087):2415–2429.

Ma, Y., Yu, F., Feng, P., Zheng, Y. e Wang, Z. (2021). Generated model based data augmentations for classification of her2 immunohistochemical pathological images in breast cancer. Em *2021 6th International Conference on Mathematics and Artificial Intelligence*, ICMAI 2021, página 45–50, New York, NY, USA. Association for Computing Machinery.

Marinelli, R. J., Montgomery, K., Liu, C. L., Shah, N. H., Prapong, W., Nitzberg, M., Zachariah, Z. K., Sherlock, G. J., Natkunam, Y., West, R. B., van de Rijn, M., Brown, P. O. e Ball, C. A. (2007). The stanford tissue microarray database. *Nucleic Acids Research*, 36:D871–D877.

Memon, R., Prieto Granada, C. N., Harada, S., Winokur, T., Reddy, V., Kahn, A. G., Siegal, G. P. e Wei, S. (2022). Discordance between immunohistochemistry and in situ hybridization to detect her2 overexpression/gene amplification in breast cancer in the modern age: A single institution experience and pooled literature review study. *Clinical Breast Cancer*, 22(1):e123–e133.

Moelans, C., de Weger, R., Van der Wall, E. e van Diest, P. (2011). Current technologies for her2 testing in breast cancer. *Critical Reviews in Oncology/Hematology*, 80(3):380–392.

Mukundan, R. (2019). Analysis of image feature characteristics for automated scoring of her2 in histology slides. *Journal of Imaging*, 5(3).

Orlando, L., Viale, G., Bria, E., Lutrino, E. S., Sperduti, I., Carbognin, L., Schiavone, P., Quaranta, A., Fedele, P., Caliolo, C., Calvani, N., Criscuolo, M. e Cinieri, S. (2016). Discordance in pathology report after central pathology review: Implications for breast cancer adjuvant treatment. *The Breast*, 30:151–155.

Perez, E. A., Suman, V. J., Davidson, N. E., Martino, S., Kaufman, P. A., Lingle, W. L., Flynn, P. J., Ingle, J. N., Visscher, D. e Jenkins, R. B. (2006). Her2 testing by local, central, and reference laboratories in specimens from the north central cancer treatment group n9831 intergroup adjuvant trial. *Journal of Clinical Oncology*, 24(19):3032–3038. PMID: 16809727.

Qaiser, T., Mukherjee, A., Reddy PB, C., Munugoti, S. D., Tallam, V., Pitkäaho, T., Lehtimäki, T., Naughton, T., Berseth, M., Pedraza, A., Mukundan, R., Smith, M., Bhalerao, A., Rodner, E., Simon, M., Denzler, J., Huang, C.-H., Bueno, G., Snead, D., Ellis, I. O., Ilyas, M. e Rajpoot, N.

(2018). Her2 challenge contest: a detailed assessment of automated her2 scoring algorithms in whole slide images of breast cancer tissues. *Histopathology*, 72(2):227–238.

Qaiser, T. e Rajpoot, N. M. (2019). Learning where to see: A novel attention model for automated immunohistochemical scoring. *IEEE Transactions on Medical Imaging*, 38(11):2620–2631.

Ross, J. S., Slodkowska, E. A., Symmans, W. F., Pusztai, L., Ravdin, P. M. e Hortobagyi, G. N. (2009). The her-2 receptor and breast cancer: Ten years of targeted anti–her-2 therapy and personalized medicine. *The Oncologist*, 14(4):320–368.

Saha, M. e Chakraborty, C. (2018). Her2net: A deep framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation. *IEEE Transactions on Image Processing*, 27(5):2189–2200.

Shorten, C. e Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60.

Singh, P. e Mukundan, R. (2018). A robust her2 neural network classification algorithm using biomarker-specific feature descriptors. Em *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, páginas 1–5.

Srinidhi, C. L., Ciga, O. e Martel, A. L. (2021). Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67:101813.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A. e Bray, F. (2021). Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249.

Tewary, S. e Mukhopadhyay, S. (2021). Her2 molecular marker scoring using transfer learning and decision level fusion. *Journal of Digital Imaging*, 34:667–677.

Tewary, S. e Mukhopadhyay, S. (2022). Autoihcnet: Cnn architecture and decision fusion for automated her2 scoring. *Applied Soft Computing*, 119:108572.

Tschandl, P., Rosendahl, C. e Kittler, H. (2018). The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):180161.

Vandenberghe, M. E., Scott, M. L. J., Scorer, P. W., Söderberg, M., Balcerzak, D. e Barker, C. (2017). Relevance of deep learning to facilitate the diagnosis of her2 status in breast cancer. *Scientific Reports*, 7(1):45938.

Vogel, C., Bloom, K., Burris, H., Gralow, J., Mayer, M., Pegram, M., Rugo, H., Swain, S., Yardley, D., Chau, M., Lalla, D., Brammer, M. e Kaufman, P. (2011). P1-07-02: Discordance between central and local laboratory her2 testing from a large her2-negative population in virgo, a metastatic breast cancer registry. *Cancer Research*, 71(24_Supplement):P1–07–02–P1–07–02.

Wang, X., Shao, C., Liu, W., Liang, H. e Li, N. (2022). Her2-resnet: A her2 classification method based on deep residual network. *Technology and Health Care*, 30:215–224. S1.

Wang, Z., Ma, Y., Zheng, Y., Feng, P. e Yu, F. (2021). *Weakly-Supervised Learning Using Pretraining for Classification in HER2 Immunohistochemistry Image of Breast Cancer*, página 66–71. Association for Computing Machinery, New York, NY, USA.

Wolff, A. C., Hammond, M. E. H., Hicks, D. G., Dowsett, M., McShane, L. M., Allison, K. H., Allred, D. C., Bartlett, J. M., Bilous, M., Fitzgibbons, P., Hanna, W., Jenkins, R. B., Mangu, P. B., Paik, S., Perez, E. A., Press, M. F., Spears, P. A., Vance, G. H., Viale, G. e Hayes, D. F. (2013). Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American society of clinical oncology/college of american pathologists clinical practice guideline update. *Archives of Pathology & Laboratory Medicine*, 138(2):241–256.

Yamashita, R., Nishio, M., Do, R. K. G. e Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 9(4):611–629.

Yang, J., Ju, J., Guo, L., Ji, B., Shi, S., Yang, Z., Gao, S., Yuan, X., Tian, G., Liang, Y. e Yuan, P. (2022). Prediction of her2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning. *Computational and Structural Biotechnology Journal*, 20:333–342.