UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de São José do Rio Preto

Igor Kauê Gouveia Sugawara

IA Explicável como Reforço de Treinamento em CNN: Uma Investigação no Contexto de Imagens H&E

São José do Rio Preto
2024

Igor Kauê Gouveia Sugawara

IA Explicável como Reforço de Treinamento em CNN: Uma Investigação no Contexto de Imagens H&E

Trabalho de Conclusão de Curso (TCC) apresentado como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação, junto ao Curso de Bacharelado em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de São José do Rio Preto.

Orientador:

Prof. Dr. Leandro Alves Neves

São José do Rio Preto, Novembro de 2024

Igor Kauê Gouveia Sugawara

IA Explicável como Reforço de Treinamento em CNN: Uma Investigação no Contexto de Imagens H&E

Trabalho de Conclusão de Curso (TCC) apresentado como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação, junto ao Curso de Bacharelado em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de São José do Rio Preto.

Comissão Examinadora:

Prof. Dr. Leandro Alves Neves
UNESP - Câmpus de São José do Rio Preto
Orientador

Profa. Dra. Rogéria Cristiane Gratão de Souza
UNESP - Câmpus de São José do Rio Preto

Prof. Dr. Rodrigo Capobianco Guido
UNESP - Câmpus de São José do Rio Preto

São José do Rio Preto, Novembro de 2024

S947i

Sugawara, Igor

IA Explicável como Reforço de Treinamento em CNN: Uma Investigação no Contexto de Imagens H&E / Igor Sugawara. -- São José do Rio Preto, 2024

48 p. : il., tabs., fotos

Trabalho de conclusão de curso (Bacharelado - Ciência da Computação) - Universidade Estadual Paulista (UNESP), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto

Orientador: Leandro Alves Neves

1. Redes Neurais Explicáveis. 2. Inteligência Artificial Explicável. 3. Reforço de Treinamento. 4. Imagens H&E. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

Agradecimentos

Agradeço a Deus, pela minha vida e por me permitir ultrapassar todos os obstáculos encontrados ao longo da minha graduação.

Agradeço imensamente à minha família e amigos por todo apoio que tive durante a minha vida, em especial a minha mãe Izabel, ao meu pai Mauro, e a minha irmã Samantha, pois graças a este suporte consegui realizar este trabalho.

Agradeço à Unesp IBILCE por fornecer um ambiente propício ao aprendizado e pela infraestrutura que tornou possível a realização deste trabalho.

E agradeço aos meus professores que passaram seus conhecimentos nesse meu período de graduação, em especial ao meu orientador, Prof. Dr. Leandro Alves Neves, pela ajuda com a qual me guiou no desenvolvimento deste meu trabalho.

Resumo

As redes neurais convolucionais (CNNs) têm se mostrado extremamente eficazes no processamento de imagens nos últimos anos, obtendo excelentes resultados em tarefas de classificação e de segmentação. Este trabalho foi desenvolvido visando explorar a aplicação de CNN integradas ao XAI (Inteligência Artificial Explicável) no processamento de imagens histológicas, sendo este de importância significativa em diversas áreas como a patologia e no diagnóstico médico. Devido à alta complexidade e dimensionalidade das imagens, pode ocorrer uma dificuldade de interpretação e de confiabilidade dos resultados. Para mitigar essas limitações, foram integradas técnicas de interpretabilidade, como regiões explicáveis e mapas de ativação, que identificam as áreas mais relevantes das imagens para as decisões do modelo, e a abordagem de reforço de treinamento, que ajusta iterativamente os pesos da CNN. Embora a integração do XAI não tenha mostrado uma melhora significativa no desempenho quantitativo, como acurácia (98,18%) e *loss* (0,052), ela proporcionou uma maior compreensão das decisões do modelo ao identificar as regiões das imagens mais relevantes para as classificações. A abordagem de reforço de treinamento também foi explorada, buscando ajustar os pesos da CNN. Esses achados reforçam a importância de técnicas explicáveis para aumentar a confiança no uso de CNNs em áreas críticas como a medicina, onde a interpretabilidade pode ser tão valiosa quanto o desempenho numérico.

Palavras-chave: Redes neurais convolucionais, inteligência artificial explicável, reforço de treinamento, reconhecimento de padrões, mapas de ativação, imagens histológicas.

Abstract

Convolutional Neural Networks (CNNs) have proven to be highly effective in image processing tasks in recent years, achieving remarkable results in classification and segmentation. This work explores the application of CNNs integrated with Explainable Artificial Intelligence (XAI) in processing histological images, a field of significant importance in areas such as pathology and medical diagnosis. The high complexity and dimensionality of these images often pose challenges in interpretation and reliability of results. To address these limitations, interpretability techniques were integrated, including explainable regions and activation maps, which identify the most relevant areas of the images for the model's decisions. Additionally, a training reinforcement approach was applied to iteratively adjust the CNN weights. While the integration of XAI did not significantly improve quantitative performance metrics, such as accuracy (98.18%) and loss (0.052), it provided greater insight into the model's decision-making process by highlighting key regions of the images for classification. The training reinforcement strategy further aimed to refine the CNN's performance. These findings underscore the importance of explainability techniques to enhance trust in CNN applications in critical fields such as medicine, where interpretability can be as valuable as numerical performance.

***Keywords**: Convolutional neural networks, explainable artificial intelligence, training reinforcement, pattern recognition, activation mapping, histological images.*

Lista de Ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

CAD	Computer-Aided Diagnosis
CAM	Class Activation Maps
CNN	Convolutional Neural Networks
DA	Data Augmentation
GBD	Global Burden of Disease
$H\&E$	Hematoxilina-eosina
$IARC$	International Agency for Research on Cancer
$LIME$	Local Interpretable Model-Agnostic Explanations
OMS	Organização Mundial da Saúde
RL	Reinforcement Learning
XAI	Explainable Artificial Intelligence

Sumário

1
Introdução

Nos últimos anos, o avanço da tecnologia desempenha um papel crucial no campo da medicina, proporcionando novas ferramentas e diferentes abordagens para o diagnóstico e tratamento de diversas doenças. No contexto da análise de imagens histológicas, a aplicação de técnicas de inteligência artificial em redes neurais convolucionais tem se mostrado promissora (TAVANAEI, 2020) para o auxílio de profissionais da saúde em suas decisões clínicas.

O processamento de imagens histológicas envolve a análise microscópica de tecidos e de células para a identificação de características patológicas no diagnóstico de doenças, como, por exemplo, o câncer. Este processo necessita de habilidade e experiência consideráveis por parte de especialistas, além de também sofrer o risco de erros humanos e subjetividades na interpretação dos resultados dos dados.

Portanto, neste contexto, as CNNs aparecem como uma técnica de aprendizado profundo capaz de aprender automaticamente características complexas e distintivas em imagens, fornecendo assim resultados mais precisos e consistentes. Elas são projetadas para simular a organização do córtex visual humano, em que cada camada da rede extrai características específicas (KATTENBORN et al., 2021), possibilitando assim a detecção e classificação de padrões importantes nas imagens histológicas.

No entanto, buscando melhorar ainda mais a capacidade e precisão das Redes Neurais Convolucionais (CNNs), este projeto foi realizado visando adicionar um reforço dentro das CNNs convencionais. Nesse sentido, foi explorado o uso de técnicas de aprendizado de máquina, como o *transfer learning*, para avaliar se elas podem melhorar o desempenho e os resultados de uma rede. Essas técnicas têm recebido destaque em trabalhos anteriores devido aos resultados promissores que proporcionam, facilitando o desenvolvimento de arquiteturas amigáveis ao usuário,

além de melhorar o desempenho e a precisão, ao direcionar a atenção para a aplicação específica de interesse (ISLAM et al., 2021).

Há também o uso do reforço de treinamento, que consiste em incorporar sinais adicionais durante o processo de treinamento de um modelo de aprendizado de máquina, para melhorar seu desempenho e fornecer explicações compreensíveis. Essas técnicas combinam ajustes específicos no treinamento do modelo com métodos de XAI, permitindo que os usuários entendam e confiem nas decisões tomadas pelo modelo. Isso possibilita aprimorar a interpretabilidade, transparência e confiabilidade do modelo.

1.1 Motivação e Justificativas

Um dos motivos para o estudo desse campo é o avanço tecnológico e a demanda por diagnósticos médicos mais precisos. Com o desenvolvimento de técnicas de inteligência artificial, como as CNNs, e o crescente acesso de dados histológicos, cria uma oportunidade de aprimoramento da precisão e a eficiência dos diagnósticos médicos. Logo, a aplicação dessas tecnologias pode auxiliar os profissionais de saúde na identificação precoce de doenças, permitindo assim intervenções mais efetivas e o melhoramento dos resultados clínicos.

Outro ponto é a complexidade das imagens histológicas, a análise das imagens histológicas necessitam um conhecimento especializado para identificação e interpretação de características sutis e complexas presentes nos tecidos e nas células. Assim, a utilização de CNNs permite a extração automatizada das informações relevantes (TAVANAEI, 2020), permitindo que haja uma análise mais precisa e objetiva, reduzindo assim a dependência da experiência individual do patologista.

Adicionalmente, ao se modificar uma CNN usando regiões explicáveis, é possível tornar o processo de classificação mais transparente, identificando as áreas-chave da imagem que contribuem para a decisão do modelo. Isso ajuda os patologistas a entenderem o raciocínio do modelo e a obterem *insights* valiosos sobre a análise de imagem.

O projeto visou trazer o uso de regiões explicáveis durante o *transfer learning*, isso ajuda a aumentar a confiabilidade das classificações de imagens H&E. Ao identificar e enfatizar as regiões mais relevantes para a classificação, o modelo de CNN pode fornecer resultados mais precisos e confiáveis. Isso é particularmente importante em aplicações médicas, onde diagnósticos precisos são cruciais para o tratamento adequado dos pacientes.

Existe também a necessidade do tempo e de recursos para a análise dessas imagens histológicas, pois a análise manual de grandes volumes de imagens histológicas é uma tarefa demorada e custosa (XIE et al., 2020). Portanto, trazendo a automação desta tarefa por meio de CNNs acabará acelerando este processo, permitindo assim uma análise mais rápida e eficiente, além de também liberar o tempo dos profissionais de saúde para outras atividades médicas e reduzindo também os custos associados ao diagnóstico histopatológico.

Usar regiões explicáveis em imagens H&E durante o *transfer learning*, podem permitir reduzir a necessidade de anotações detalhadas e extensivas de especialistas. Em vez de rotular manualmente todas as regiões de interesse em cada imagem, a CNN pode aprender a identificar as regiões mais relevantes automaticamente. Isso pode economizar tempo e esforço na anotação de dados e facilitar a aplicação da CNN em conjuntos de dados de grande escala.

Por fim, o reforço de aprendizado envolve a incorporação de informações adicionais durante o treinamento do modelo, para melhorar sua precisão e eficiência. No caso das técnicas de XAI, o reforço de aprendizado pode ser realizado por meio do uso de regiões explicáveis nas imagens histológicas durante o processo de transfer learning. Essas regiões podem ser identificadas automaticamente pela CNN, permitindo que o modelo aprenda a enfatizar as áreas mais relevantes para a classificação.

Isso resulta em diagnósticos mais precisos e confiáveis, aumentando a confiança dos profissionais de saúde na utilização desses modelos como ferramentas de apoio à decisão. Exemplo disso é a XCNN (TAVANAEI, 2020), sendo uma rede neural

convolucional explicável modificada para poder representar os recursos visuais em uma arquitetura de rede. Trazendo assim uma melhor explicabilidade com os mapas de calor interpretáveis resultantes.

Portanto, modificar uma CNN pode fornecer benefícios importantes, como interpretabilidade aprimorada, maior confiabilidade diagnóstica, poupar tempo e esforço, e trazer diagnósticos com maior precisão.

1.2 Objetivos

Este trabalho visa explorar a aplicação das Redes Neurais Convolucionais (CNN) integradas à Inteligência Artificial Explicável (XAI) no processamento de imagens H&E, buscando contribuir no avanço da área de diagnóstico clínico, no reconhecimento de características histopatológicas, além de trazer uma melhor interpretabilidade das classificações. Os principais objetivos deste TCC são:

1. Implementar duas redes neurais pré-treinadas extraindo índices, resultados e mapas das regiões mais exploradas das imagens classificadas;
2. Utilizar técnicas de Inteligência Artificial Explicável, como o LIME e CAM, para gerar regiões explicáveis das imagens classificadas;
3. Implementar duas redes neurais convolucionais com transferência de aprendizado com o uso das regiões explicáveis;
4. Explorar técnicas de aprendizado por reforço para otimizar os modelos de CNN desenvolvidos;
5. Extrair as métricas resultantes e realizar comparações desses resultados;
6. Aplicar os modelos de CNNs resultantes em imagens histológicas com relevância científica.

1.3 Organização do trabalho

O presente trabalho está organizado em cinco capítulos, onde no capítulo 1 se encontra esta introdução. Em sequência temos no capítulo 2 a apresentação de

conceitos aplicados no desenvolvimento do trabalho, como redes neurais convolucionais (CNN), *transfer learning*, *fine-tuning*, *epochs*, técnicas para gerar *split* dos dados no processo de *fine-tuning*, definição de loss e acurácia, XAI, imagens histológicas, aprendizado por reforço e trabalhos relacionados. No capítulo 3 é apresentada a metodologia usada no projeto das redes neurais convolucional integrado ao XAI. No capítulo 4 são descritos os testes feitos e os resultados dos experimentos. No capítulo 5 é feita a conclusão obtida deste projeto.

2
Referencial Teórico e Trabalhos Relacionados

Neste capítulo são apresentados os conceitos teóricos que serviram como base para o desenvolvimento deste trabalho, além de também apresentar trabalhos relacionados presentes na literatura.

Na seção 2.1 estão detalhes sobre as Redes Neurais Convolucionais (CNN), além de suas características estruturais. Na seção 2.2 é introduzido o conceito de *transfer learning* e os benefícios de seu uso. Na seção 2.3 é introduzido o conceito de *fine-tuning* e como seu uso auxilia na execução do trabalho. Na subseção 2.4 são apresentadas técnicas para geração de *split* dos dados no processo de *fine-tuning*.

Na seção 2.5 é apresentado o conceito de *epochs*. Na subseção 2.6 são definidos os conceitos de loss e acurácia, e como estas são aplicadas no contexto de CNN. Na seção 2.7 está a descrição da fundamentação técnica e teórica do XAI, além dos parâmetros comumente utilizados. Na seção 2.8 é apresentado a descrição de imagens histológicas. Na seção 2.9 estão algumas informações envolvendo o aprendizado por reforço. Por fim, na seção 2.10 é feita uma análise com trabalhos relacionados.

2.1 Redes Neurais Convolucionais

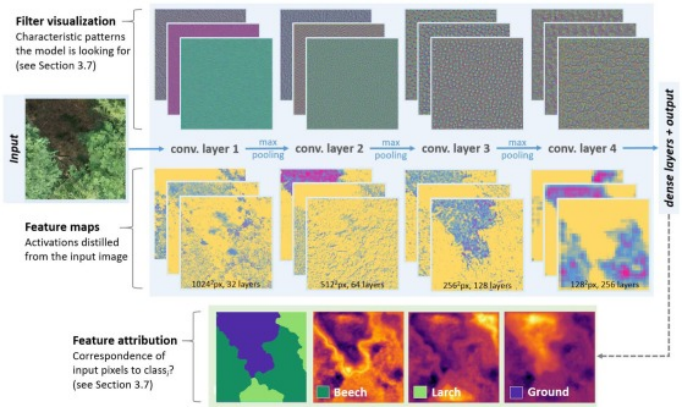
Uma Rede Neural Convolucional (CNN) é um tipo especializado de rede neural artificial projetado para o processamento de dados com uma estrutura em grade (imagens, por exemplo). As CNNs são amplamente utilizadas em tarefas de visão computacional, reconhecimento de padrões e processamento de imagens.

A estrutura das CNNs é inspirada no processo biológico, considerando abstrações de neurônios organizados em camadas para aprender representações hierarquicamente, em que os neurônios entre as camadas são interconectadas por meio de pesos e vieses (KATTENBORN et al., 2021).

A rede neural convolucional é constituída de várias camadas, sendo elas, a camada de entrada (*input layer*), as camadas internas ocultas (*hidden layers*), e a camada de saída (*output layer*) (KATTENBORN et al., 2021). A Figura 2.1 mostra o exemplo de um esquema de uma CNN, sua estrutura e operações realizadas nela.

Na camada de entrada, são inseridos os dados de uma imagem. Em seguida, as camadas internas ocultas realizam a transformação dos dados, em que se inclui pelo menos uma camada convolucional, e através destas camadas é possível fazer a transformação dos dados e explorar padrões, para assim produzir mapas de características (*feature maps*). Por fim, a camada de saída retorna a devida classificação da imagem (KATTENBORN et al., 2021).

Figura 2.1: Ilustração do esquema de uma CNN composta por quatro camadas convolucionais e operações treinadas de *pooling*.



Fonte: (KATTENBORN et al., 2021)

2.2 Transfer Learning

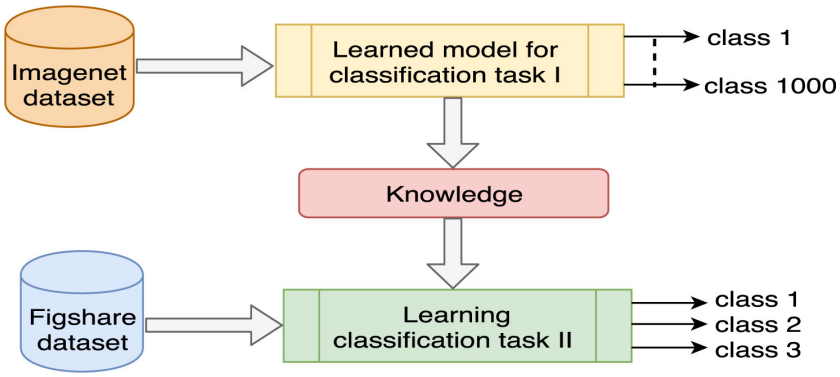
Recentemente, a classe especial de aprendizado profundo (*deep learning*) chamada de transferência de aprendizado profundo (*deep transfer learning*) domina os estudos em diversos campos, como na categorização visual, reconhecimento de objetos e em problemas de classificação de imagens (DEEPAK et al., 2019).

Através da transferência de aprendizado (*transfer learning*) é permitido o uso de um modelo de CNN pré-treinado no qual foi desenvolvido para outro trabalho

relacionado (ZHOU et al. 2019). Além disso, a transferência de aprendizado mostrou seu potencial em diagnóstico assistido por computador (CAD), como no uso de um modelo de Inception V3 pré-treinado para diferenciação de tumores renais benignos e malignos.

Outro exemplo é no uso de *transfer learning* em um modelo VGG-16 pré-treinado para o AlexNet como um classificador de câncer de mama sobre imagens histopatológicas (DENIZ et al. 2019), onde através disso é feita a extração de características classificadas via um vetor de suporte de máquina (*Support Vector Machine*). A Figura 2.2 mostra como o *transfer learning* é feito no exemplo citado.

Figura 2.2: Ilustração de um exemplo de configuração de transferência de aprendizado.



Fonte: (DEEPAK et al., 2019)

2.3 Fine-Tuning

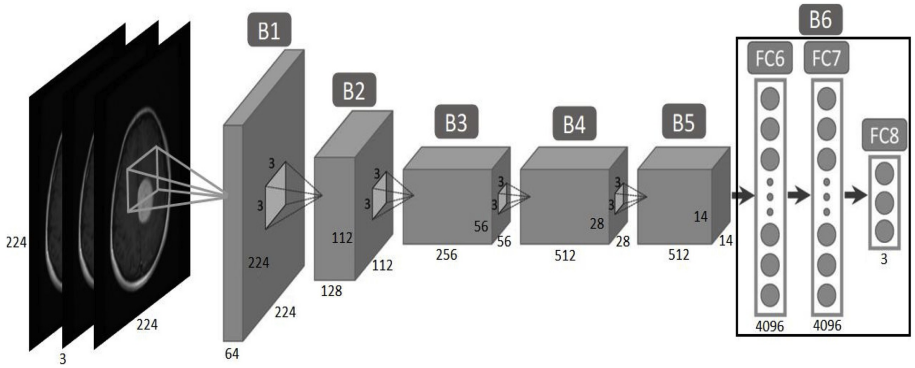
Como descrito anteriormente, o *transfer learning* permite que uma tarefa de origem afete a ideia indutiva da tarefa alvo. A maneira típica de conduzir essa transferência de aprendizado com redes neurais profundas é realizar o *fine-tuning* em um modelo pré-treinado na tarefa de origem, usando os dados da tarefa de destino. Comparando com o treinamento do zero, o *fine-tuning* em um modelo de rede neural convolucional pré-treinado em um conjunto de dados de destino pode melhorar

muito o desempenho (GUO et al. 2019).

Porém, para o treinamento e otimização de uma rede muito profunda como na arquitetura VGG19 (SWATI et al. 2019), que contém 19 camadas e 144 milhões de parâmetros treináveis, é necessário um conjunto de dados extenso, logo, há a necessidade de decidir a melhor estratégia de *fine-tuning* para o algoritmo.

Por exemplo, se aplica o *fine-tuning* em camadas, adicionando uma camada por vez, haveria 19 camadas para realizar o *fine-tuning*, no caso da arquitetura VGG19, o que demanda muito tempo e não é tão eficiente. Logo, utiliza-se a estratégia de *fine-tuning* em blocos, onde o VGG19 é dividido em 6 blocos baseados nas camadas de *pooling*, levando assim a uma estratégia mais rápida e otimizada. A Figura 2.3 exibe como essa estratégia de *fine tuning* é aplicada.

Figura 2.3: Ilustração de um exemplo de configuração de *fine tuning*.



Fonte: (SWATI et al., 2019)

2.4 Técnicas de geração de split dos dados

Com as técnicas de *deep learning*, é possível explicar problemas complexos, aprendendo a partir de representações simples (ISLAM et al., 2021). As principais características que tornaram os métodos de *deep learning* tão populares são a capacidade de aprender as representações exatas, e a propriedade de aprender os dados de maneira profunda, em que múltiplas camadas são utilizadas sequencialmente.

Os métodos de aprendizado profundo são amplamente utilizados em sistemas médicos, como no campo da análise de imagens médicas, por exemplo. Normalmente, os sistemas de *deep learning* são compostos de várias etapas, como a coleta e preparação dos dados, a extração e classificação de características e avaliação de desempenho (ISLAM et al., 2021).

Porém, considerando o passo da preparação dos dados, que converte os dados em um formato apropriado, temos o pré-processamento, que inclui diversas operações, como a remoção de ruídos, o redimensionamento, etc. Na etapa de particionamento dos dados há a divisão (*split*) dos dados em conjuntos de treinamento, validação e teste para o experimento. Assim, os dados de treinamento são usados para desenvolver o modelo particular, a sua avaliação é feita pelos dados de validação, e por fim, o desempenho do modelo desenvolvido é avaliado pelos dados de teste (ISLAM et al., 2021).

2.5 Epochs

A *epoch* refere-se a uma unidade de medida que representa a passagem completa de todos os dados de treinamento por um algoritmo de aprendizado de máquina, ou seja, o ciclo no qual o algoritmo de aprendizado processa todos os exemplos de treinamento disponíveis. Um número maior de *epochs* permite que o modelo explore mais exemplos de treinamento e refine seus parâmetros de maneira mais precisa.

Porém, um número excessivo de epochs pode ocasionar um *overfitting*, o modelo se ajusta demais aos dados de treinamento com um desempenho ruim em dados não vistos anteriormente. Logo, a determinação do número adequado de epochs é um aspecto importante para o treinamento de modelos de aprendizado de máquina.

Pelos modelos de aprendizado profundo estarem sendo treinados em grandes conjuntos de dados, os métodos existentes podem acabar não utilizando todas as informações de diferentes *epochs* de maneira eficaz. Por isso, um método de lidar com este problema, é aproveitar as informações de cada *epoch* de treinamento para

suprimir os mapas de predição das epochs subsequentes.

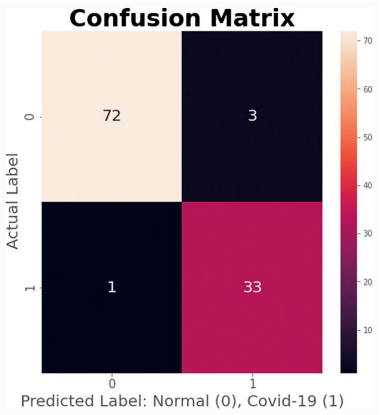
Um exemplo desta técnica é o uso da arquitetura FANet (TOMAR et al., 2022), que unifica a máscara da *epoch* anterior com o mapa de características da *epoch* de treinamento atual. Assim, a máscara da *epoch* anterior é usada para fornecer atenção aos mapas de recursos aprendidos em diferentes camadas convolucionais.

2.6 Loss e acurácia

A acurácia e a perda (*loss*) são métricas usadas para a avaliação do desempenho de modelos de aprendizado de máquina, em especial aos problemas de classificação. A acurácia indica a proporção de previsões corretas feitas pelo modelo em relação ao total de previsões, enquanto o loss indica o quão distante as previsões do modelo estão dos valores verdadeiros.

Por isso, uma forma de melhorar a precisão dos casos relatados e previsão com precisão a doenças de radiografias de tórax, foi feito um modelo de CNNs (VAID et al., 2020) que detectasse anormalidades estruturais e categorização de doenças a fim de descobrir padrões ocultos. Além disso, para realizar tal abordagem foi utilizada a abordagem de *transfer learning*. Isso trouxe ótimos resultados, oferecendo uma precisão bastante alta de mais de 96,3%. A matriz de confusão da Figura 2.4 exibe os resultados obtidos do modelo (VAID et al., 2020).

Figura 2.4: Matriz de confusão do modelo de predição de testes.



Fonte: (VAID et al., 2020)

2.7 XAI

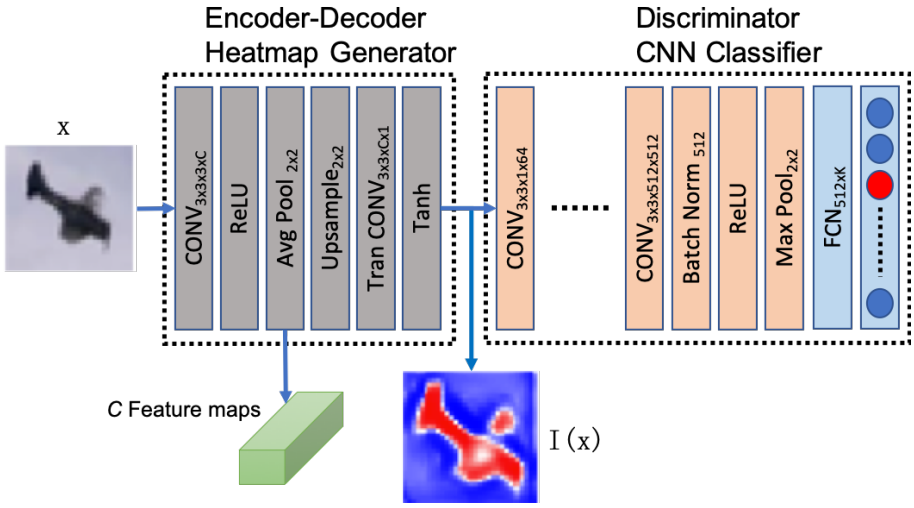
A compreensão das camadas intermediárias que fazem parte de um modelo de *deep learning* e a descoberta dos recursos de condução de estímulos têm ganhado cada vez mais interesse nos últimos anos. Assim, a fim de fornecer uma nova maneira de abrir a caixa preta da IA, tem-se o uso do XAI (*Explainable artificial intelligence*), que traz uma decisão transparente e interpretável do modelo (TAVANAEI, 2020).

Além do fato das CNNs terem demonstrado um desempenho notável em diferentes áreas de reconhecimento de padrões, especialmente na visão computacional e na análise de imagens, o entendimento de como elas realizam a extração de características discriminativas de dados não estruturados é fundamental. Estudos recentes permitiram a visualização dos campos receptivos e dos mapas de características das camadas neurais, proporcionando uma melhor compreensão do fluxo de informações nas hierarquias convolucionais.

Porém, compreender as camadas intermediárias de um modelo de *deep learning* e detectar os recursos de condução de estímulo é uma tarefa desafiadora. Logo, o uso da inteligência artificial explicável XAI ajuda a realizar esta tarefa, descrevendo detalhes sobre os recursos extraídos nas camadas intermediárias. Além disso, o XAI possibilita a compreensão do motivo da previsão, detectando características que impactam positiva e negativamente a ativação da camada neural final (TAVANAEI, 2020).

Um exemplo do funcionamento de uma CNN explicável (TAVANAEI, 2020) é na extração e representação das características espaciais de uma imagem, enquanto classifica ou prevê a imagem. A arquitetura consiste em um componente *encoder-decoder* anexado ao início de uma CNN, onde a saída do *encoder-decoder* é a entrada da CNN. A estrutura dessa CNN explicável pode ser vista na Figura 2.5.

Figura 2.5: Ilustração da arquitetura da rede neural convolucional explicável.



Fonte: (TAVANAEI et al., 2020)

2.8 Imagens histológicas

Segundo estatísticas da Agência Internacional de Pesquisa em Câncer (IARC) da Organização Mundial de Saúde (OMS), e o estudo Carga Global de Doenças (GBD), os casos de câncer aumentaram cerca de 28% entre 2006 e 2016, além de haver 2,7 milhões de novos casos de câncer até 2030 (BOYLE et al., 2008; MORAGA-SERRANO, 2018). Para seu diagnóstico são utilizadas diversas técnicas de biópsia, esses processos envolvem a coleta de amostras de células ou tecidos, fixação na lâmina do microscópio e em seguida, corando-os. Assim, as imagens histopatológicas são analisadas e o diagnóstico é efetuado por patologistas (XIE et al., 2020).

Porém, a análise de imagens histopatológicas é algo difícil e demorado, além do resultado da análise poder ser afetado pelo nível de experiência dos patologistas envolvidos. Logo, a análise de imagens histopatológica com o auxílio de computadores desempenham um papel significativo no diagnóstico do câncer de mama e em prognóstico. Porém, as imagens de câncer são normalmente imagens de alta resolução e de granulação fina que retratam estruturas geométricas ricas e texturas complexas. Assim, a variabilidade em uma classe e a consistência entre elas po-

dem tornar a classificação extremamente difícil, especialmente quando se trata de múltiplas classes.

Outro ponto a ser visto são as limitações de métodos de extração de características em imagens histopatológicas de câncer, pois os métodos tradicionais de extração de recursos dependem de informações supervisionadas. Outra questão é a necessidade de conhecimento prévio para a seleção de recursos úteis, tornando a eficiência do recurso de extração muito baixa e a carga computacional muito pesada. No fim, as características finais extraídas são apenas algumas características de baixo nível e não uma representação dos recursos das imagens histopatológicas, podendo gerar um modelo final que produza resultados de classificações ruins (XIE et al., 2020).

Por isso, o uso de técnicas de *deep learning* é eficiente para o trabalho de imagens histológicas, pois através dessas técnicas é possível fazer a extração de recursos e recuperação de informações de dados de maneira automática, além de aprender representações abstratas avançadas de dados. Além disso, eles conseguem resolver os problemas da extração tradicional de características aplicada em visão computacional.

2.9 Aprendizado por reforço

O aprendizado por reforço (*Reinforcement Learning*) aborda problemas de tomada de decisão sequencial em que um agente descobrirá a política ideal via interações de tentativa e erro com o ambiente. Observações visuais, como imagens, são intuitivas para um agente perceber seu ambiente, assim o aprendizado visual de aprendizado por reforço a partir de observações visuais é aplicado em vários domínios, como na direção autônoma, por exemplo.

Porém, a implantação de técnicas visual de *Reinforcement Learning* no mundo real continua sendo desafiador devido à sua baixa eficiência de amostragem e de grandes lacunas de generalização. Assim, para enfrentar esses obstáculos é utilizado a técnica de aumento de dados (*Data Augmentation*) para adquirir amostras

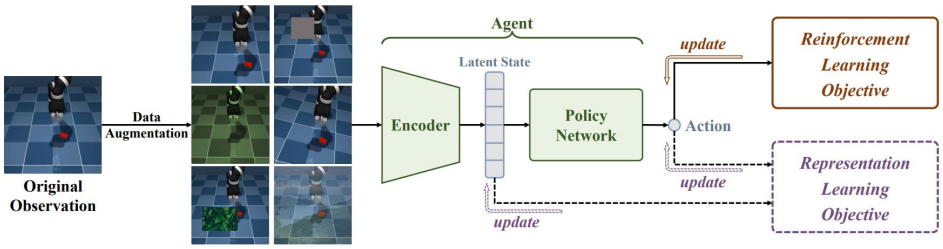
eficientes e generalização de políticas por diversificação de dados de treinamento (MA et al., 2022).

Para aprender de forma eficiente com as amostras e melhorar a generalização visual dos agentes de aprendizado por reforço, tem sido dedicado um esforço considerável ao desenvolvimento de diversas abordagens. Essas incluem o uso de técnicas de regularização de entropia para restringir os pesos do modelo, a realização de aprendizado conjunto combinando a loss de RL com tarefas auxiliares para fornecer supervisão adicional, e a construção de modelos mundiais de Reinforcement Learning que permitem o aprendizado de comportamentos a partir de resultados simulados.

Embora essas abordagens tenham alcançado um sucesso notável, elas ainda sofrem por dados de interação limitados e diversidade amostral pobre. Assim, para aumentar a quantidade e diversidade de dados de treinamento, o *Data Augmentation* tem recebido atenção crescente da comunidade de *Reinforcement Learning* visual nos últimos anos. Como um método baseado em dados, o *Data Augmentation* é ortogonal aos métodos mencionados anteriormente e pode ser combinado com eles para melhorar ainda mais seu desempenho.

Além disso, o *Data Augmentation* é essencial para pré-treinar uma representação de tarefa cruzada e várias técnicas de DA, como o corte aleatório usada em quase todos algoritmos de RL visual em uma etapa de pré-processamento de dados (MA et al., 2022). A Figura 2.6 mostra como esse *Data Augmentation* é feito e aplicado no exemplo citado.

Figura 2.6: Exemplo de um diagrama de fluxo de um trabalho genérico que alavanca *Data Augmentation* em *Reinforcement Learning*.



Fonte: (MA et al., 2022)

2.10 Trabalhos relacionados

A análise das imagens histopatológicas é uma tarefa difícil e demorada, além de necessitar do conhecimento de profissionais (XIE et al., 2020). Outra questão é que o resultado da análise pode ser afetado pelo nível de experiência dos patologistas envolvidos. Por isso as CNNs são amplamente utilizadas para a classificação de imagens histológicas, visto que auxiliado por computadores, a análise de imagens histopatológicas desempenha um papel significativo no diagnóstico de doenças.

Um exemplo de como as redes neurais convolucionais são utilizadas no diagnóstico de câncer de mama é apresentado no modelo de trabalho (XIE et al., 2020) que usa as redes neurais convolucionais profundas Inception V3 e Inception ResNet V2 treinado com técnicas de *transfer learning*. Essas duas redes são pré-treinadas no grande conjunto de dados de imagens do ImageNet. Então, sua estrutura e parâmetros são congelados.

O número de neurônios na última camada totalmente conectada é definido de acordo com nossa tarefa específica, e os parâmetros da camada totalmente conectada são treinados novamente. Logo, o modelo pode ser usado para executar operações binárias ou na classificação de múltiplas classes das imagens histopatológicas do câncer de mama.

Outro exemplo que traz o uso de técnicas e estratégias e de treinamento, como

o *Transfer Learning* e *Reinforcement Learning*, sendo usados comumente juntos à classificação de imagens para se obter resultados mais otimizados e precisos, é o modelo (KATTENBORN et al., 2021) que visa localizar e caracterizar plantas vasculares através da análise de imagens e visão computacional com técnicas de aprendizado profundo. Trazendo o uso da estratégia de treinamento *Data Augmentation*, que visa compensar poucas observações de referência, aumentando o número de dados de referência introduzindo pequenas manipulações nos dados existentes, ou criando dados sintéticos.

Outro exemplo é o modelo (DEEPAK et al., 2019) que aplica conceitos de *deep transfer learning* para a extração de recursos a partir de imagens de ressonância magnética do cérebro, onde esses recursos foram usados junto a modelos de classificadores para um melhor desempenho. Obtendo assim um sistema com melhor precisão de classificação, além de obter também um bom desempenho com um número menor de amostras de treinamento.

Por fim, um exemplo que traz o uso de *Fine-Tuning* é o modelo (GUO et al. 2019) que especializa a estratégia de *fine-tuning* para cada exemplo de treinamento do conjunto de dados. Mostrando que ao se comparar com o treinamento do zero, ao trazer o uso do *fine-tuning* em uma rede neural convolucional pré-treinada em um conjunto de dados consegue melhorar significativamente o seu desempenho, reduzindo os requisitos de dados rotulados de destino.

A explicabilidade de modelos de aprendizado profundo de alto desempenho é um problema desafiador em diferentes áreas de pesquisa, especialmente visão computacional. Por isso, há projetos que buscam a proposta da criação de uma rede neural convolucional explicável (XCNN) para representar os recursos visuais de condução de estímulos em uma arquitetura de rede de ponta a ponta.

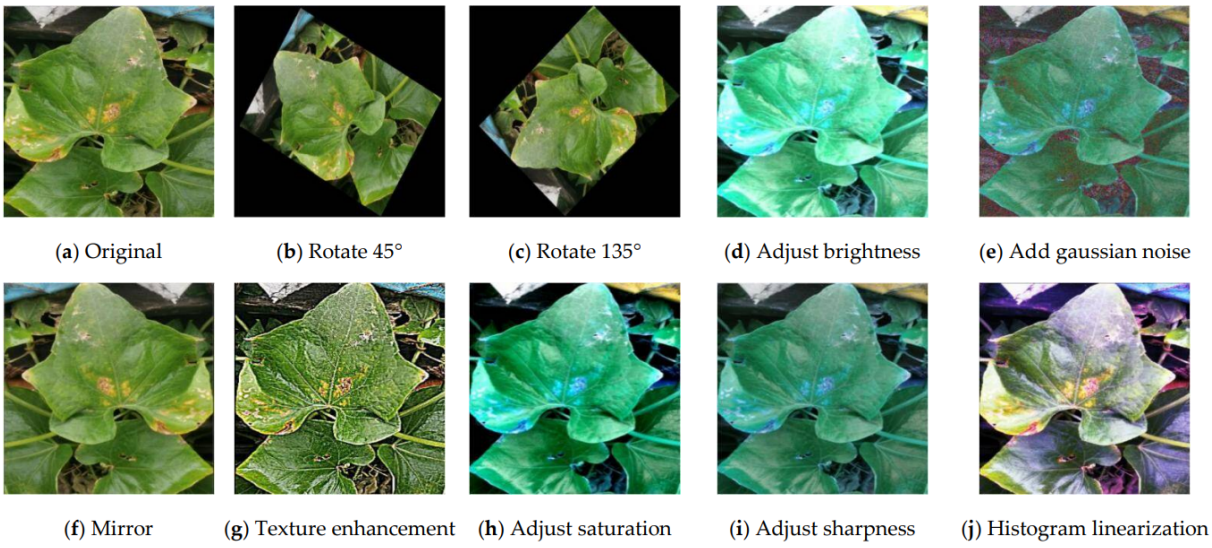
O modelo (TAVANAEI, 2020), que consiste em dois componentes, sendo eles um mapa de calor gerador construído por camadas neurais codificador-decodificador e um classificador CNN, traz resultados experimentais mostrando mapas de calor interpretáveis que visualmente superam as redes explicáveis de última geração e

os geradores de mapas de saliência, oferecendo uma arquitetura simples que pode ser reaplicada a qualquer classificador CNN. Isso mostra o sucesso da rede neural convolucional explicável em descobrir os principais recursos visuais.

Por fim, o modelo (LU et al., 2021) utiliza abordagens de *Deep Learning*, utilizando imagens de doenças em plantas como referência. Primeiramente é feito a preparação de dados e pré-processamento de imagens utilizado técnicas de *Deep Learning*, depois é feita a construção do modelo de arquitetura (treinamento e validação).

Neste ponto, para trabalhar com mais números de datasets além do disponível, é utilizada a técnica de aprendizado por reforço através do *Data Augmentantion*, métodos como rotação, simetria-espelho e ajuste de saturação são alguns dos exemplos de transformação de imagens feita nesta etapa. A figura 2.7 exibe exemplos de imagens resultantes após a execução do Data Augmentation no modelo (LU et al., 2021).

Figura 2.7: Exemplo de imagens resultantes do *Data Augmentation* nas imagens originais.



Fonte: (LU et al., 2021)

O mesmo ocorre no trabalho proposto a fim de trazer uma melhor precisão dos resultados, mas este vai além ao utilizar imagens resultantes de XAI como mais um reforço para o processamento de classificação das imagens. Além disso, diferentes técnicas e arquiteturas de CNNs são utilizadas conforme as características das imagens. A Tabela 2.1 apresenta os trabalhos relacionados citados, comparando seus conceitos aos do objetivo deste trabalho.

Tabela 2.1 – Comparação entre os estudos

Referência	**Imagens Histológicas**	**Técnicas de deep learning**	**Aprendizado por reforço com Data Augmentation**	**Utilização de XAI**	**Classificação de Imagens com XAI**
DEEPAK et al., 2019	SIM	SIM	SIM	NÃO	NÃO
DENIZ et al., 2019	SIM	SIM	NÃO	NÃO	NÃO
TAVANAEI, 2020	NÃO	SIM	NÃO	SIM	SIM
LU et al., 2021	NÃO	SIM	SIM	NÃO	NÃO

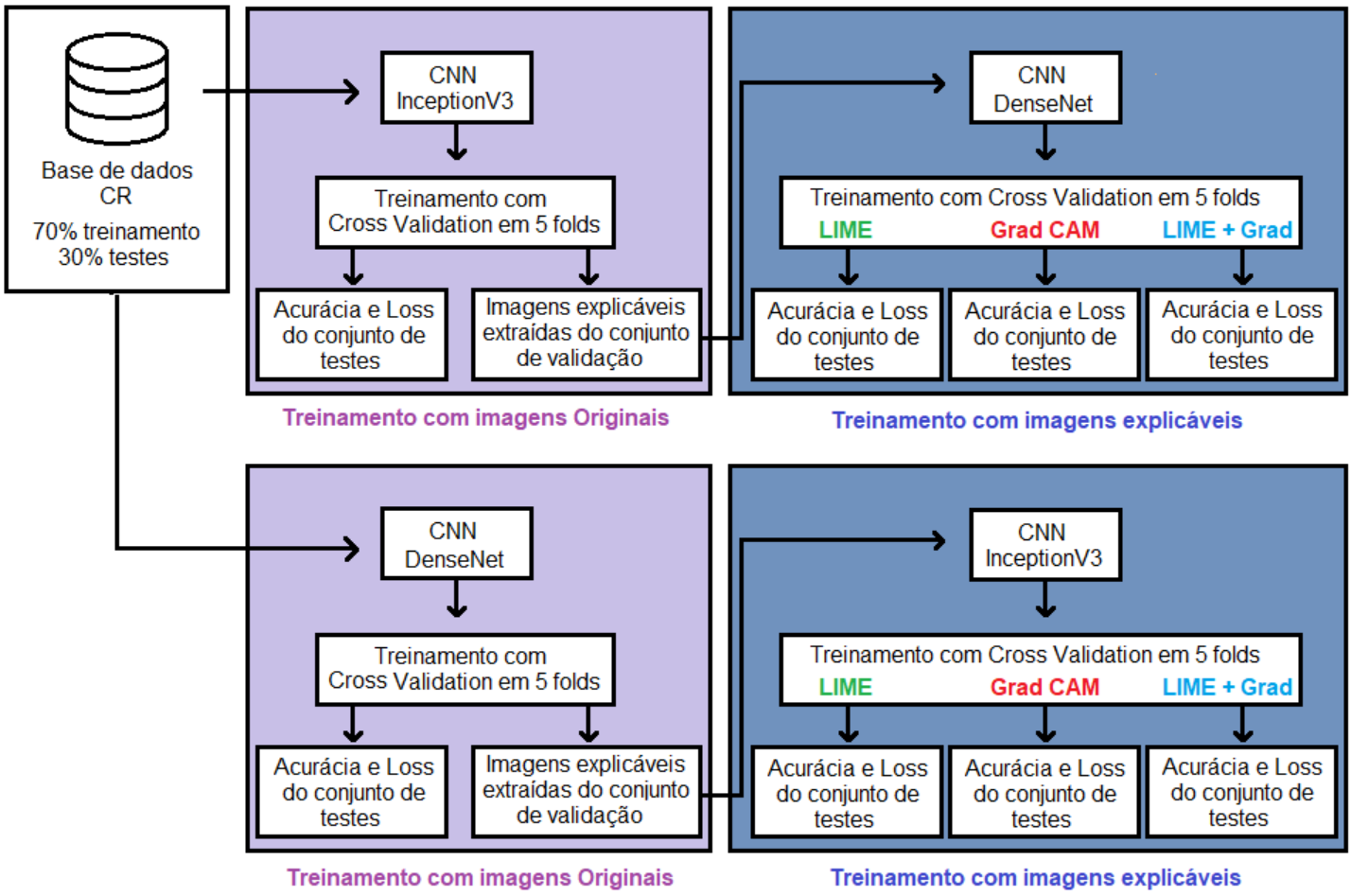
3
Metodologia

Neste capítulo é descrito a metodologia utilizada para efetuar o modelo proposto para o trabalho. São descritos todos os procedimentos adotados para realizar a investigação proposta, que consiste em utilizar dois classificadores de imagens, o Inception V3 e o DenseNet, para realizar classificações em um conjunto de imagens histopatológicas H&E. São feitas 4 classificações de imagens, sendo a classificação por Inceptionv3 e DenseNet pré-treinadas, e a classificação após o transfer learning nestas mesmas CNNs.

O trabalho consiste em diversas etapas, sendo a seleção de imagens H&E, o pré-processamento das imagens, geração dos split de dados, treinamento inicial das CNNs, geração das regiões explicáveis, aplicação do *transfer learning* e *fine tuning*, e por fim as comparações dos resultados obtidos pelas CNNs. Além disso, foi definido que as CNNs trabalharão com *cross validation* em 5 *folds*, onde cada *fold* trabalha com 10 epochs e há uma proporção de split de dados de 70% para treinamento e 30% para testes.

O diagrama da Figura 3.1 mostra resumidamente os processos utilizados na metodologia do projeto proposto. Inicialmente, há a entrada das imagens nos seus devidos classificadores, para que, após a classificação, recebam seus respectivos valores de perda (*loss*) e acurácia, assim como as imagens das regiões explicáveis. Por fim, são feitas novas classificações, utilizando as imagens geradas anteriormente. Cada etapa é descrita mais detalhadamente nas respectivas seções.

Figura 3.1: Ilustração da metodologia proposta.

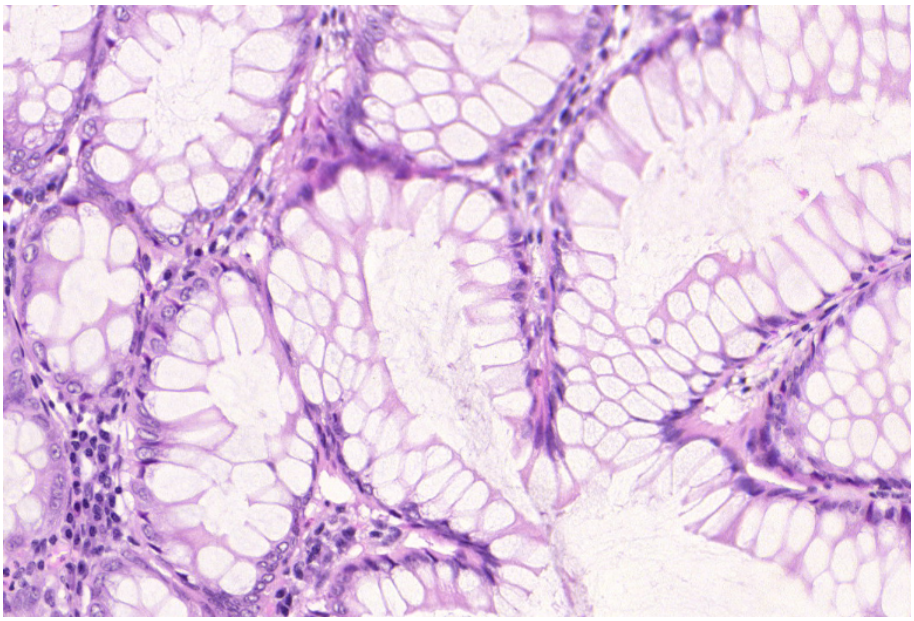


Fonte: Elaborado pelo autor.

3.1 Etapa 1: Base de dados

Para realizar este trabalho, utiliza-se uma base de dados pública (CR) composta por imagens H&E de câncer colorretal. Estas imagens foram obtidas por um conjunto de dados médicos e selecionadas para identificação de padrões específicos. Essa base de dados consiste em um conjunto de imagens de tecidos histológicos, contendo diferentes tipos de células e estruturas relevantes para análise. A Figura 3.2 mostra um exemplo de imagem H&E retirada do dataset fornecido para o trabalho.

Figura 3.2: Imagem histopatológica de carcinoma benigno. Coloração hematoxilina e eosina (H&E).



Fonte: Obtida do dataset público CR fornecido para este trabalho.

3.2 Etapa 2: Pré-processamento dos dados

Antes de treinar as CNNs, é necessário realizar algumas etapas de pré-processamento nos dados. Inicialmente é feito o redimensionamento das imagens para um tamanho fixo, isso irá garantir que haja consistência durante o treinamento. Em seguida, aplica-se um processo de normalização para ajustar os valores de intensidade dos píxeis, para melhorar a convergência da rede durante o treinamento.

Além disso, foram aplicadas também técnicas de aumento de dados, como a rotação, o espelhamento e a translação, para aumentar a variabilidade dos dados de treinamento e evitar *overfitting*.

3.3 Etapa 3: Geração de split dos dados

Como visto anteriormente, através do split de dados é possível fazer uma melhor avaliação do modelo (ISLAM et al., 2021). Logo, para avaliar o desempenho dos classificadores e comparar os resultados, o conjunto de dados será dividido em

conjuntos de treinamento, validação e teste. A divisão será realizada estratificadamente para preservar a proporção de cada classe presente nas imagens. O conjunto de treinamento será utilizado para treinar os modelos, o conjunto de validação para ajustar os hiper parâmetros durante o treinamento, e também para o treinamento da segunda rede após o processo das técnicas XAI, e o conjunto de teste para avaliar o desempenho final dos modelos.

3.4 Etapa 4: Treinamento inicial das redes

Os classificadores escolhidos para este trabalho são o Inceptionv3 e o Densenet. Ambos são arquiteturas de redes neurais convolucionais (CNNs) amplamente utilizadas em tarefas de classificação de imagens.

Inicialmente, será realizado o treinamento dos classificadores utilizando as imagens originais do conjunto de treinamento. Esse processo de treinamento consistirá em alimentar as redes neurais com as imagens e seus rótulos correspondentes, ajustando os pesos das camadas internas por meio de algoritmos de otimização (por exemplo, Gradiente Descendente Estocástico), visando minimizar uma função de perda (como a Entropia Cruzada).

3.5 Etapa 5: Geração de Regiões Explicáveis

Outro ponto é na geração de regiões explicáveis utilizando técnicas de XAI. A partir dessas regiões será possível trazer interpretabilidade ao modelo (TAVANAEI, 2020) e também realizar o *transfer learning*. Essas regiões explicáveis são obtidas a partir das imagens histopatológicas H&E, visando compreender quais áreas das imagens são mais relevantes para a classificação realizada pelos modelos de CNN. Para essa etapa, então, foram utilizadas duas técnicas populares de XAI, o LIME e o CAM.

O LIME é uma técnica que permite gerar explicações locais para as predições dos modelos de aprendizado de máquina, independentemente da sua arquitetura. O LIME será aplicado para gerar regiões explicáveis que indicam quais regiões da imagem contribuíram significativamente para a classificação realizada pelos classi-

ficadores. O LIME opera gerando amostras perturbadas da imagem original e avaliando como a perturbação afeta a predição do modelo. Assim, o LIME fornece uma explicação local, destacando as regiões mais importantes da imagem para a classificação.

Já o CAM é uma técnica específica para CNNs que permite identificar as regiões mais ativadas durante o processo de classificação. Essas regiões são mapeadas na imagem original, gerando um mapa de ativação da classe de interesse. Logo, o CAM foi utilizado para gerar mapas de ativação de classe para as classes de interesse do problema. Esses mapas de ativação ajudaram a identificar as regiões que mais influenciam as predições dos classificadores, permitindo uma interpretação visual das decisões tomadas pelos modelos.

3.6 Etapa 6: *Transfer Learning* e *Fine-Tuning*

Após o treinamento inicial dos classificadores com as imagens originais e a geração das regiões explicáveis, foi aplicado o *Transfer Learning* junto a técnica de *Fine-Tuning*. Essa técnica consiste em ajustar os pesos da rede neural pré-treinada (neste caso, Inception V3 e Densenet) utilizando um conjunto de dados diferente, mas relacionado à tarefa original.

São utilizadas também as regiões explicáveis obtidas por técnicas de XAI, especificamente o *Local Interpretable Model-Agnostic Explanations* (LIME) e o *Class Activation Maps* (CAM), para gerar um novo conjunto de treinamento. Essas regiões são utilizadas como entradas adicionais juntamente com as imagens originais. Assim, o modelo aprenderá a considerar essas regiões ao realizar suas predições.

São utilizadas bibliotecas de *deep learning*, como o TensorFlow, e implementada a técnica em Python. O resultado é uma visualização das regiões mais relevantes nas imagens H&E, permitindo uma análise mais aprofundada e uma melhor interpretação das decisões da CNN. Segue-se então a ideia de gerar mapas de calor interpretáveis e de saliência (TAVANAEI, 2020).

3.7 Etapa 7: Aprendizado por reforço

Nesta etapa, é explorado o uso do aprendizado por reforço e reforço de treinamento via técnicas de Inteligência Artificial Explicável (XAI) para melhorar o desempenho e a interpretabilidade dos modelos de CNN desenvolvidos.

O aprendizado por reforço envolve o treinamento de agentes inteligentes capazes de tomar decisões e aprimorar seu desempenho com base em recompensas recebidas. No contexto deste trabalho, o aprendizado por reforço é aplicado para otimizar as redes neurais convolucionais (CNNs) desenvolvidas anteriormente, permitindo que elas se adaptem de forma dinâmica e melhorem sua capacidade de classificação.

Além disso, é utilizado o reforço de treinamento via técnicas de XAI, que consiste em incorporar sinais adicionais durante o processo de treinamento das CNNs. Essas técnicas de XAI, como o LIME e o CAM, são utilizadas para fornecer explicações compreensíveis das decisões tomadas pelo modelo.

3.8 Etapa 8: Avaliação do desempenho dos modelos pelas métricas

Após o treinamento dos classificadores com as diferentes abordagens mencionadas anteriormente, os resultados obtidos são avaliados. É utilizado o conjunto de teste para medir o desempenho final dos modelos. Por fim, é feita uma análise comparativa entre os classificadores treinados com imagens originais e os classificadores treinados por *transfer learning* com as regiões explicáveis obtidas pelo LIME e CAM.

As métricas de desempenho utilizadas para avaliar os modelos incluirão acurácia, precisão, recall e F1-score, dadas pelas equações 3.1, 3.2, 3.3 e 3.4, respectivamente.

$$\text{Acurácia} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3.1}$$

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

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

$$\text{F1-score} = 2 \cdot \frac{\text{Precisão} \cdot \text{Recall}}{\text{Precisão} + \text{Recall}} \tag{3.4}$$

4
Experimentos e resultados

Neste capítulo são apresentados os experimentos realizados, os resultados obtidos dos testes de implementação da IA explicável como reforço de treinamento, e exemplos de imagens LIME e Grad CAM que foram extraídas e usadas.

4.1 Tecnologias e base de dados

Para o desenvolvimento deste trabalho, foi utilizado a linguagem Python na versão 3.10.12, e bibliotecas sendo as principais delas o Mime (0.2.0.1), Matplotlib (3.7.1), Numpy (1.26.4) e Tensorflow (2.17.0).

Através da implementação do código no google colab, foi possível fazer o treinamento de máquina com cerca de 115 imagens, sendo estas divididas por *cross validation*, e feito o treinamento total em 5 *folds*. Com isso, foram geradas 115 imagens para LIME e Grad CAM sendo extraídas a partir da validação de cada *fold*. Por fim, o mesmo conjunto de testes (50 imagens) definidos no início do projeto foram aplicados tanto nos classificadores pré-treinados quanto nos classificadores treinados por XAI, para que ambas recebam os mesmos dados para avaliação.

4.2 Metodologia e desempenho dos modelos

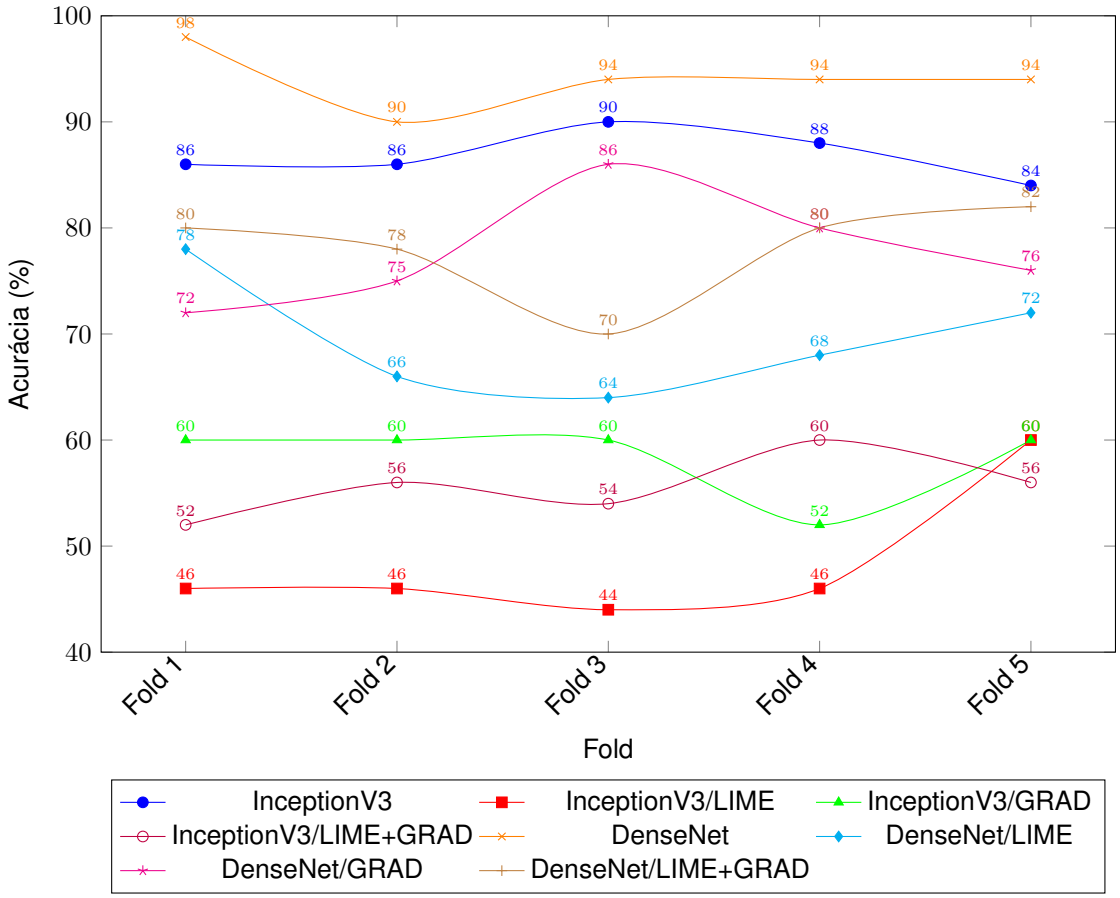
Os modelos do trabalho foram aplicados no conjunto de imagens H&E de cânceres benignos e malignos. Foram feitas classificações com as redes InceptionV3 e Densenet pré-treinadas, e novamente com as mesmas redes, mas utilizando técnicas de transferência de aprendizado com imagens explicáveis LIME e GRAD CAM. A Tabela 4.1 e a ilustração do gráfico na Figura 4.1 mostram o desempenho de cada *fold*, apresentando a porcentagem da acurácia para os diferentes métodos propostos no conjunto de testes. Levando em conta as médias, o maior desempenho obtido foi pelo classificador Densenet com imagens originais no treinamento, enquanto o

menor desempenho foi com o InceptionV3 com imagens LIME no treinamento.

Tabela 4.1 – Resultados dos métodos de classificação para cada *fold* e média geral.

Método	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Média**
InceptionV3	86%	86%	90%	88%	84%	86.8%
InceptionV3/LIME	46%	46%	44%	46%	60%	48.4%
InceptionV3/GRAD	60%	60%	60%	52%	60%	58.4%
InceptionV3/LIME+GRAD	52%	56%	54%	60%	56%	55.6%
DenseNet	98%	90%	94%	94%	94%	94.0%
DenseNet/LIME	78%	66%	64%	68%	72%	69.6%
DenseNet/GRAD	72%	75%	86%	80%	76%	77.8%
DenseNet/LIME+GRAD	80%	78%	70%	80%	82%	78.0%

Figura 4.1: Gráfico com os resultados da acurácia dos métodos de classificação pra cada *fold*.



Fonte: Elaborado pelo autor.

Além dos resultados dos testes realizados em cada *fold*, foi realizada uma execução final após a conclusão de todos os *folds*. A Figura 4.2 apresenta os métodos de classificação, juntamente com suas respectivas matrizes de confusão e acurácias. Observa-se que, inicialmente, durante a fase de pré-treinamento, as acurácias obtidas pelos classificadores foram superiores às obtidas após o treinamento com imagens explicáveis.

Figura 4.2: Ilustração das matrizes de confusão e acurácias resultantes de cada método.

InceptionV3 / Acurácia: 84%

	Pred. Positivo	Pred. Negativo
Real Positivo	22	1
Real Negativo	7	20

Densenet / Acurácia: 94%

	Pred. Positivo	Pred. Negativo
Real Positivo	21	2
Real Negativo	1	26

InceptionV3 LIME / Acurácia: 60%

	Pred. Positivo	Pred. Negativo
Real Positivo	12	11
Real Negativo	9	18

Densenet LIME / Acurácia: 72%

	Pred. Positivo	Pred. Negativo
Real Positivo	9	14
Real Negativo	0	27

InceptionV3 GRAD / Acurácia: 60%

	Pred. Positivo	Pred. Negativo
Real Positivo	23	0
Real Negativo	20	7

Densenet GRAD / Acurácia: 76%

	Pred. Positivo	Pred. Negativo
Real Positivo	23	0
Real Negativo	12	15

InceptionV3 LIME + GRAD / Acurácia: 56%

	Pred. Positivo	Pred. Negativo
Real Positivo	23	0
Real Negativo	22	5

Densenet LIME + GRAD / Acurácia: 82%

	Pred. Positivo	Pred. Negativo
Real Positivo	23	0
Real Negativo	9	18

Fonte: Elaborado pelo autor.

Esse comportamento pode ser atribuído à dificuldade dos classificadores em distinguir de maneira eficaz entre as classes, sugerindo que os modelos não foram capazes de aprender as características discriminativas dos dados adequadamente. A introdução dos métodos explicáveis, como LIME e Grad-CAM, parece ter influenciado negativamente o desempenho dos modelos, possivelmente porque esses métodos não conseguiram capturar de forma robusta as características relevantes para cada classe. Isso pode ter resultado em classificações inconsistentes ou enviesadas, comprometendo a generalização do modelo.

A fim de confirmar os resultados da acurácia obtidos, na tabela 4.2 são apresentados os valores de precisão, sensibilidade (*recall*) e *F1-score* dos métodos testados. No geral, a DenseNet mais uma vez se mostra com um melhor desempenho comparada aos outros métodos, enquanto nos modelos com a incorporação do LIME e Grad-CAM resultou em uma redução nos valores de precisão, de sensibilidade e de *F1-score*.

Isso sugere que, embora as técnicas de XAI possam ser úteis para o entendimento do comportamento de classificadores, elas acabam introduzindo uma complexidade maior, reduzindo a precisão das previsões. Apesar disso, o Grad-Cam combinado aos modelos apresenta melhor desempenho e estabilidade do que o LIME, indicando que a técnica de visualizações das regiões importantes (Grad-CAM) é mais compatível com os modelos comparado à técnica LIME, que utiliza explicações locais para prever a contribuição de cada característica. Na combinação das duas técnicas XAI, o desempenho ficou próximo aos resultados do Grad-CAM.

Tabela 4.2 – Médias de precisão, sensibilidade e F1-score para cada método de classificação.

Método	**Precisão**	**Sensibilidade**	**F1-score**
InceptionV3	0,855	0,85	0,84
InceptionV3/LIME	0,595	0,595	0,595
InceptionV3/GRAD	0,765	0,63	0,55
InceptionV3/LIME+GRAD	0,75	0,60	0,495
DenseNet	0,94	0,935	0,94
DenseNet/LIME	0,83	0,695	0,675
DenseNet/GRAD	0,83	0,78	0,75
DenseNet/LIME+GRAD	0,86	0,835	0,82

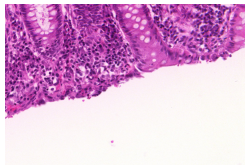
4.3 Análise das técnicas de XAI

A análise geral mostra que a DenseNet se destaca como uma arquitetura mais robusta para a classificação de imagens, mesmo com a adição de técnicas XAI. Estas mesmas técnicas, apesar de serem uma forma eficiente de compreender os modelos CNNs, ainda requerem um maior cuidado na implementação, pois podem comprometer o desempenho por conta de sua complexidade. A figura 4.3 mostra alguns exemplos de imagens explicáveis geradas pelos modelos InceptionV3 e Densenet. Além das extrações das regiões e explicações locais, é feito um overlay com as imagens originais a fim de obter um melhor resultado na classificação.

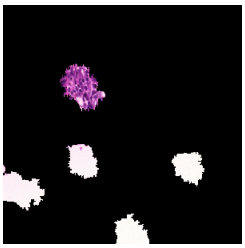
Através da análise das imagens (b) e (c) da Figura 4.3, mostra um exemplo do porquê a utilização da técnica LIME no treinamento não obteve um desempenho significativo comparado ao Grad CAM. Através do LIME é possível fazer a extração dos locais que contribui para a característica da imagem, sendo que trabalho foi construído a partir de que o LIME fosse capaz de retornar uma imagem com até 5 locais.

Porém, analisando a imagem resultante, nota-se a presença de 1 local que foi realmente significativo para classificação da imagem, enquanto há 4 locais em branco. Com o *transfer learning* para um novo classificador (DenseNet), a imagem acabou não se mostrando suficiente para ajudar no treinamento. Já com a técnica do Grad CAM foi possível destacar regiões importantes para a classificação da imagem sem desconsiderar grandes partes da imagem que ajudam no processo de treinamento.

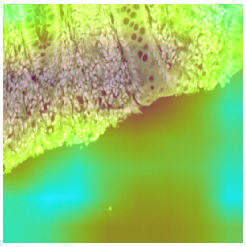
Figura 4.3: Visualização de imagens de câncer usando InceptionV3 e DenseNet com técnicas LIME e XAI.



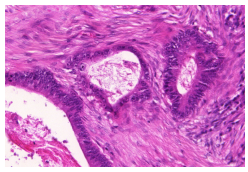
(a) Imagem Original Cancer Benigno



(b) Imagem LIME InceptionV3

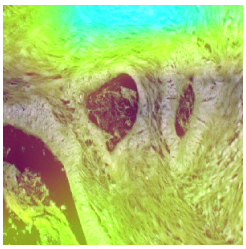


(c) Imagem Grad CAM InceptionV3

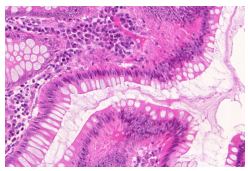


(d) Imagem Original Cancer Maligno

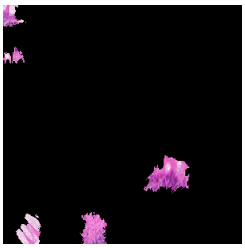
(e) Imagem LIME InceptionV3



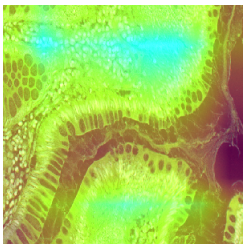
(f) Imagem Grad CAM InceptionV3



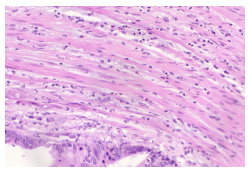
(g) Imagem Original Cancer Benigno



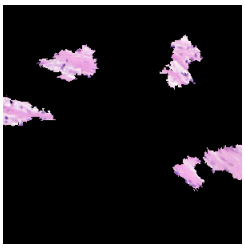
(h) Imagem LIME DenseNet



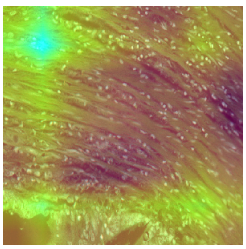
(i) Imagem Grad CAM DenseNet



(j) Imagem Original Cancer Maligno



(k) Imagem LIME DenseNet



(l) Imagem Grad CAM DenseNet

Fonte: Obtida do dataset CR e extraídas pelo autor.

5
Conclusão

Este estudo explorou técnicas avançadas de classificação, reconhecimento de padrões e geração de explicações em imagens H&E de câncer colorretal (CR), proporcionando uma análise comparativa entre abordagens tradicionais e técnicas de explicabilidade. Com base nos experimentos realizados, observou-se que a arquitetura DenseNet apresentou um desempenho superior em termos de precisão, mesmo quando comparada às abordagens que incorporaram técnicas de explicabilidade, como LIME e Grad-CAM. Esses resultados sugerem que a DenseNet é uma arquitetura mais robusta e eficaz para a classificação de imagens H&E de câncer, independentemente da aplicação de técnicas de explicabilidade, em comparação ao InceptionV3.

Apesar do potencial das técnicas de inteligência artificial explicável (XAI) em auxiliar na interpretação do comportamento dos modelos, os experimentos evidenciam que sua utilização pode introduzir uma complexidade adicional, impactando negativamente o desempenho dos classificadores. Isso foi particularmente evidente com o uso do LIME, cuja abordagem de explicações locais não conseguiu abranger adequadamente as regiões mais relevantes das imagens, comprometendo a eficácia das previsões.

Por outro lado, o Grad-CAM demonstrou-se uma técnica mais apropriada para integração com os classificadores, proporcionando explicações visuais mais consistentes e com menor impacto negativo sobre o desempenho, em comparação ao LIME. Essa diferença ressalta a importância de uma escolha criteriosa das técnicas de explicabilidade conforme a arquitetura da rede neural e os objetivos da aplicação. No contexto da classificação de imagens médicas, como as de câncer colorretal, técnicas que focam em regiões mais amplas e relevantes, como o Grad-CAM, podem oferecer vantagens significativas tanto em termos de interpretabilidade quanto de

desempenho.

Embora as técnicas de explicabilidade não tenham contribuído diretamente para a melhoria do desempenho dos classificadores, elas proporcionaram uma compreensão mais profunda do comportamento dos modelos. Essa compreensão é particularmente crítica em aplicações sensíveis como a saúde, onde a confiabilidade das decisões automatizadas é de suma importância. A transparência fornecida pelas técnicas de XAI pode aumentar a confiança dos profissionais de saúde na utilização de modelos de IA.

Pesquisas futuras poderiam se concentrar em otimizar as técnicas de explicabilidade de modo a minimizar o impacto negativo sobre o desempenho dos classificadores, especialmente em redes neurais convolucionais. Além disso, estratégias inovadoras de pré-processamento de imagens explicáveis ou a combinação de diferentes métodos de XAI podem ser investigadas, visando equilibrar a interpretabilidade dos modelos e sua precisão. A longo prazo, a integração efetiva de técnicas de IA explicável com modelos robustos de classificação tem o potencial de fornecer não apenas maior transparência, mas também maior facilidade e confiança na adoção dessas tecnologias.

Referências bibliográficas

Boyle, P., and Levin, B. (2008). World Cancer report 2008: IARC Press. International Agency for Research on Cancer.

Deepak, S., and Ameer, P. M. (2019). Brain tumor classification using deep CNN features via transfer learning. *Computers in Biology and Medicine*.

Deniz, E., Şengür, A., Kadiroğlu, Z., Guo, Y., Bajaj, V., and Budak, Ü. (2018). Transfer learning based histopathologic image classification for breast cancer detection. *Health Information Science and Systems*, 6(1).

Guo, Y., Shi, H., Kumar, A., Grauman, K., Rosing, T., and Feris, R. (2019). SpotTune: Transfer Learning through Adaptive Fine-tuning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Islam, M. M., Karray, F., Alhajj, R., and Zeng, J. (2021). A Review on Deep Learning Techniques for the Diagnosis of Novel Coronavirus (COVID-19). *IEEE Access*, 9, 30551-30572.

Kattenborn, T., Leitloff, J., Schiefer, F., and Hinz, S. (2021). Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, Pages 24-49.

Lu, J., Tan, L., and Jiang, H. (2021). Review on convolutional neural network (CNN)

applied to plant leaf disease classification. Agriculture, 11(8), 707.
Ma, G., Wang, Z., Yuan, Z., Wang, X., Yuan, B., and Tao, D. (2022). A Comprehensive Survey of Data Augmentation in Visual Reinforcement Learning. *arXiv:2210.04561*.

Moraga-Serrano, P. E. (2018). Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability adjusted life-years for 29 cancer groups, 1990 to 2016: a systematic analysis for the global burden of disease study. JAMA Oncol.

Swati, Z. N. K., Zhao, Q., Kabir, M., Ali, F., Ali, Z., Ahmed, S., and Lu, J. (2019). Brain tumor classification for MR images using transfer learning and fine-tuning. *Computerized Medical Imaging and Graphics*, 75, 34-46.

Tavanaei, A. (2020). Embedded Encoder-Decoder in Convolutional Networks Towards Explainable AI. *arXiv:2007.06712*.

Tomar, N. K., Jha, D., Riegler, M. A., Johansen, H. D., Johansen, D., Rittscher, J., Halvorsen, P., and Ali, S. (2022). FANet: A Feedback Attention Network for Improved Biomedical Image Segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 1-14.

Vaid, S., Kalantar, R. and Bhandari, M. (2020). Deep learning COVID-19 detection bias: accuracy through artificial intelligence. *International Orthopaedics (SICOT) 44*, 1539–1542.

Xie, J., Liu, R., Luttrell IV, J., and Zhang, C. (2019). Deep Learning Based Analysis

of Histopathological Images of Breast Cancer. *Frontiers in Genetics*, 10.

Zhou, L., Zhang, Z., Chen, Y.-C., Zhao, Z.-Y., Yin, X.-D., and Jiang, H.-B. (2019). A Deep Learning-Based Radiomics Model for Differentiating Benign and Malignant Renal Tumors. *Translational Oncology*, Pages 292-300.