UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
UNIDADE ACADÊMICA DE SISTEMAS E COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MATHEUS GOMES MAIA

INTERPRETABILIDADE DE REDES NEURAIS CONVOLUCIONAIS COM ESTUDO DE CASO EM DIAGNÓSTICO POR IMAGEM

CAMPINA GRANDE – PB
2022

Matheus Gomes Maia

Interpretabilidade de Redes Neurais Convolucionais com Estudo de Caso em Diagnóstico por Imagem

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Universidade Federal de Campina Grande

Centro de Engenharia Elétrica e Informática

Coordenação de Pós-Graduação em Ciência da Computação

Orientador: Herman Matins Gomes

Brasil

Junho de 2022

M217i Maia, Matheus Gomes.
Interpretabilidade de redes neurais convolucionais com estudo de caso em diagnóstico por imagem / Matheus Gomes Maia. - Campina Grande, 2022.
107 f. il. color.

Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2022.
"Orientação: Prof. Dr. Herman Martins Gomes."
Referências.

1. Inteligência Artificial. 2. Redes Neurais Convolucionais. 3. Interpretabilidade de Redes Neurais Convolucionais. 4. Diagnóstico por Imagem. I. Gomes, Herman Martins. II. Título.

CDU 004.8(043)

CATALOGRÁFICA ELABORADA PELA BIBLIOTECÁRIA ITAPUANA SOARES DIAS GONÇALVES CRB-15/93

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
POS-GRADUACAO CIENCIAS DA COMPUTACAO
Rua Aprigio Veloso, 882, - Bairro Universitario, Campina Grande/PB, CEP 58429-900

FOLHA DE ASSINATURA PARA TESES E DISSERTAÇÕES

MATHEUS GOMES MAIA

INTERPRETABILIDADE DE REDES NEURAIS CONVOLUCIONAIS COM ESTUDO DE CASO EM DIAGNÓSTICO POR IMAGEM

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação como pré-requisito para obtenção do título de Mestre em Ciência da Computação.

Aprovada em: 14/06/2022

Prof. Dr. HERMAN MARTINS GOMES, Orientador, UFCG

Prof. Dr. EANES TORRES PEREIRA, Examinador Interno, UFCG

Prof. Dr. KELSON ROMULO TEIXEIRA AIRES, Examinador Externo, UFPI

Documento assinado eletronicamente por **HERMAN MARTINS GOMES**, **PROFESSOR 3 GRAU**, em 14/06/2022, às 23:11, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **EANES TORRES PEREIRA**, **PROFESSOR 3 GRAU**, em 15/06/2022, às 09:20, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **KELSON RÔMULO TEIXEIRA AIRES**, **Usuário Externo**, em 15/06/2022, às 10:44, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

A autenticidade deste documento pode ser conferida no site https://sei.ufcg.edu.br/autenticidade, informando o código verificador **2481820** e o código CRC **207810F7**.

Agradecimentos

Os primeiros agradecimentos são direcionados ao professor Herman pelos ensinamentos, orientação e suporte durante todo o mestrado. Os maiores agradecimentos aos meus pais, Carolina e Fernando, pelo apoio e incentivo de sempre, pois se não fosse pela dedicação de vocês eu não concluiria mais esta etapa da minha vida. Agradeço ao meu irmão, Pedro Arthur; às minhas avós, Ladjane e Arabela, e ao meu avô, Hugo, que estiveram sempre ao meu lado, me apoiando e torcendo por mim. Agradecimento à minha namorada Carmem pelo companheirismo, carinho e compreensão em todos os momentos do mestrado. Amo amo você. Agradeço a todos do Laboratório de Percepção Computacional (LPC) pelo suporte e amizade. Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo incentivo financeiro e oportunidade de aperfeiçoamento. Finalmente, agradeço aos professores que aceitaram fazer parte da banca.

Resumo

As redes neurais profundas viabilizaram notáveis avanços em aplicações de processamento e análise de imagens. A complexidade decorrente da adoção desses modelos de aprendizado de máquina, caracterizados por um número crescente de parâmetros, induz representações de conhecimento e fluxos de decisão que ultrapassam a compreensão humana, principalmente quando se trata de modelos utilizados em tarefas visuais, como as Redes Neurais Convolucionais. A presente pesquisa revisa e estrutura abordagens e técnicas de interpretabilidade, que objetivam expor, de maneira compreensível, o funcionamento ou conhecimento interno desses modelos convolucionais. Nesta pesquisa, estão apresentados e organizados os objetivos das técnicas de interpretabilidade e as características de modelos interpretáveis revisados. O recente debate científico sobre a avaliação de novas técnicas de interpretabilidade foi apresentado e explorado de maneira prática em um estudo de caso. O estudo foi realizado em um nicho de imagens não usual para pesquisas que avaliam técnicas de interpretabilidade, o de diagnóstico médico, que se mostrou desafiador e rico em aprendizados. O estudo de caso considera, de ponta a ponta, as etapas de treinamento, interpretabilidade e avaliação das técnicas, algo que pode ser facilmente reproduzido em outros conjuntos de dados. Para a etapa de treinamento foram utilizadas duas arquiteturas convolucionais treinadas em dois conjuntos compostos por imagens de exames médicos, sendo esses, raio-X torácico e tomografia de coerência óptica (OCT). Para a etapa de explicação, dez técnicas de interpretabilidade foram utilizadas para produzir explicações para os modelos treinados. Por fim, para a etapa de avaliação das técnicas, as explicações foram submetidas a três avaliações: Guia para Pertubações, Randomização dos Rótulos e Jogo de Apontar. Cada avaliação trouxe desafios e aprendizados sobre os seus enfoques e recursos necessários. Por exemplo, a avaliação Guia para Pertubações favorece técnicas concisas, mas é prejudicada por um efeito conhecido como “entradas fora da distribuição”, além de possuir elevado custo de processamento. A presente dissertação tem como principais contribuições a produção de uma revisão bibliográfica narrativa e propositiva sobre a área de interpretabilidade de redes neurais, incluindo a discussão de temas tais como taxonomia das abordagens para interpretabilidade e categorização das avaliações quantitativas de técnicas de atribuição, além de um estudo de caso sobre avaliações de técnicas que interpretam Redes Neurais Convolucionais.

Palavras-chave: inteligência artificial. redes neurais convolucionais. interpretabilidade. diagnóstico por imagem.

Abstract

Deep neural networks have enabled remarkable advances in image processing and analysis applications. The complexity arising from the adoption of such machine learning models, characterized by an ever-increasing number of parameters, induces knowledge representations and decision flows that surpass human comprehension, especially when considering models used in visual tasks, such as Convolutional Neural Networks. The present research reviews and structures interpretability approaches and techniques, with ambition to expose, in an understandable way, the functioning or the internal underpinnings of these convolutional models. Throughout this study, the objectives of reviewed interpretability techniques and the characteristics of interpretable models are presented and organized. The recent scientific debate on the evaluation of novel techniques was presented and explored practically in the form of a study case. The study case was carried out in a niche of images unsual for researches that avaluate interpretality techniques, the niche of medical diagnosis, which proved to be challenging but learning-rich. The study case carries from end to end the stages of training, interpretability, and evaluation of the techniques, something that can be easily reproduced in other data sets. For the training stage, two convolutional architectures were used, a VGG16 architecture and a custom architecture, that is simpler, but still share the same principles as the VGG16. Both networks were trained with two sets of medical images, a thoracic X-ray dataset and optical coherence tomography (OCT) dataset. For the explanation stage, ten interpretation techniques been used to produce explanations for the trained models. Finally, for the evaluation stage, three known evaluations were carried out: Perturbation Guide, Labels Randomization and Pointing Game. Each evaluation brought challenges and lessons learned about its approaches and requirements. For example, the Perturbation Guide assessment benefit conciseness, but is harmed by an effect known as "out-of-distribution" entries, as well as being computational costly. The main contributions of this dissertation are the production of a of a narrative and propositional bibliographic review on the area of interpretability of neural networks, including the discussion of topics, such as taxonomy of approaches to interpretability and categorization of quantitative assessments of attribution techniques, in addition to a study case on evaluations of techniques that interpret Convolutional Neural Networks.

Keywords: artificial intelligence. convolutional neural networks. interpretability. medical imaging.

Sumário

Lista de ilustrações

Lista de quadros

Lista de tabelas

Lista de abreviaturas e siglas

ACE	*Automated Concept-based Explanation*
ALE	*Accumulated Local Effects*
AM	Aprendizado de Máquina
ATV	*Average Total Variation*
AUC	*Area Under the Curve*
BAM	*Benchmarking Attribution Methods*
CAD	*Computer-aided Diagnosis*
CAM	*Class Activation Mapping*
CAV	*Concept Activation Vectors*
CNN	*Convolutional Neural Network*
CNV	*Choroidal Neovascularization*
COCO	*Common Objects in Context*
CV	*Computer Vision*
DCNN	*Deep Convolutional Neural Network*
DL	*Deep Learning*
DME	*Diabetic Macular Edema*
GAN	*Generative Adversarial Network*
GDPR	*General Data Protection Regulation*
IA	Inteligência Artificial
ICE	*Individual Conditional Expectation*
ILSVRC	*ImageNet Large Scale Visual Recognition Challenge*
LIME	*Local Interpretable Modelagnostic Explanations*
ML	*Machine Learning*

MNIST	*Modified National Institute of Standards and Technology*
OCT	*Optical Coherence Tomography*
OoD	*Out of Distribution*
PDP	*Partial Dependence Plot*
RISE	*Randomized Input Sampling for Explanation of Black-box Models*
SEC	*Seed, Expand and Constrain*
SHAP	*SHapley Additive exPlanations*
TCAV	*Testing with Concept Activation Vectors*
VAE	*Variational Autoencoder*
VGG	*Visual Geometry Group*
XAI	*Explainable Artificial Intelligence*

1 Introdução

Existe um grande interesse por Aprendizado de Máquina como ferramenta para extrair padrões e apoiar decisões em cenários com grande volume de dados disponíveis, como classificação de imagens (SZEGEDY et al., 2015) e processamento de linguagem natural (YOUNG et al., 2018). Entre os ramos existentes no contexto de Aprendizado de Máquina, o Aprendizado Profundo (*Deep Learning* ou DL) é o mais proeminente, que pode ser caracterizado como uma forma de aprendizagem composta por várias camadas de representações internas hierárquicas. A criação de modelos com um número crescente de camadas é o que explica o termo "Profundo", que só foi possibilitado pelo grande volume de dados digitais disponibilizados nos últimos anos, avanços no poder computacional e novos algoritmos de aprendizado de máquina. (BRAAMS, 2008)

Alguns dos maiores sucessos do aprendizado profundo ocorreram no campo de Visão Computacional (*Computer Vision* ou CV) (DENG et al., 2009). A CV se concentra na compreensão de imagens e vídeos, e lida com tarefas como classificação, detecção e segmentação de objetos — úteis para determinar se uma radiografia de um paciente contém tumores malignos, por exemplo (ESTEVA et al., 2019). O foco deste trabalho está no domínio da compreensão de imagens e nas Redes Neurais Convolucionais (*Convolutional Neural Networks* ou CNNs), que são modelos biologicamente inspirados, muito utilizados para o aprendizado no domínio das imagens. Uma rede neural convolucional pode, em poucas horas, ser exposta a milhões de exemplos, algo que um médico pode nunca ter acesso em sua carreira.

O Aprendizado Profundo pode ser efetivo em uma variedade de tarefas de diagnóstico médico, inclusive superando humanos em algumas dessas tarefas. No entanto, a falta de transparência que esses modelos oferecem é um obstáculo para o uso clínico desses algoritmos (SINGH; SENGUPTA; LAKSHMINARAYANAN, 2020). A complexidade proveniente da adoção de modelos neurais com número crescente de parâmetros e hierarquias torna o mecanismo interno desses modelos superior ao que a compreensão humana consegue assimilar. Existe uma variedade de abordagens e técnicas presentes na literatura que objetivam expor, de maneira compreensível, o funcionamento ou conhecimento interno desses modelos profundos, em um campo denominado interpretabilidade ou Inteligência Artificial Explicável (*Explainable Artificial Intelligence* ou XAI).

A presente dissertação inclui uma revisão bibliográfica com foco em interpretabilidade de Redes Neurais Convolucionais. São discutidos objetivos, abordagens, técnicas pertinentes, questões científicas ainda em aberto e tendências identificadas na literatura. O campo da interpretabilidade, por ser recente, possui muitos debates científicos em aberto,

e um ponto de evolução perceptível é a questão da validação das técnicas. Muitas das publicações iniciais não validam de maneira robusta as explicações propostas e utilizam apenas exemplos ilustrativos para tentar convencer o leitor da sua utilidade. A presente dissertação apresenta diferentes avaliações que permitem a comparação entre técnicas existentes na literatura, além de realizar um estudo de caso sobre como se comparam diversas técnicas de interpretabilidade no nicho de diagnóstico por imagem.

1.1 Motivações

O aprendizado profundo é o método de IA líder para uma ampla gama de tarefas, incluindo diagnóstico de imagens médicas. É o estado da arte para várias tarefas de visão computacional e tem sido utilizado em tarefas de diagnóstico por imagem como na classificação de Alzheimer (JO; NHO; SAYKIN, 2019), detecção de câncer de pulmão (HUA et al., 2015), detecção de doenças da retina (SENGUPTA et al., 2020; LEOPOLD et al., 2021), etc. Apesar de alcançar resultados notáveis no domínio médico, os métodos baseados em IA não alcançaram uma implantação significativa nas clínicas (SINGH; SENGUPTA; LAKSHMINARAYANAN, 2020). Diante desta dificuldade, é possível acrescentar interpretabilidade aos métodos de IA para fornecer explicações visuais, exemplos similares ou legendas automáticas, permitindo uma maior simbiose e confiança entre humanos e máquinas na análise de imagens médicas, promovendo a utilização de IA em um formato de Diagnóstico Assistido por Computador (*Computer-aided Diagnosis ou CAD*).

A interpretabilidade se faz necessária não apenas para aumentar a confiança e o dinamismo entre o usuário e o sistema computadorizado, mas também atender conformidade às leis que asseguram direito de explicação, uma das barreiras para a ampla utilização de aprendizado profundo. Um exemplo de barreira legal é a diretiva GDPR (*General Data Protection Regulation*) (GREENE et al., 2019), imposta em maio de 2018 pela União Europeia, a qual concede direito à explicação para todas as decisões tomadas por sistemas automatizados ou sistemas algorítmicos artificialmente inteligentes.

A interpretabilidade promove o desenvolvimento de sistemas mais transparentes, explicáveis e interativos, de modo a ganhar a confiança de profissionais, reguladores e usuários. Logo, um maior entendimento das abordagens e técnicas de interpretabilidade permite a difusão de IA mais segura e ética, principalmente em aplicações de alto risco. A revisão bibliográfica presente nesta dissertação tem como motivação auxiliar o entendimento da área de interpretabilidade.

A fartura de técnicas disponíveis pode intimidar um interessado em interpretabilidade. A biblioteca *Saliency* (RESEARCH, 2021), por exemplo, disponibiliza 9 técnicas diferentes de interpretabilidade para redes convolucionais. O estudo de caso conduzido para este trabalho tem como motivação integrar conhecimentos sobre diferentes técnicas

e comparação entre técnicas. Tal estudo foi conduzido em um nicho particular e pode exemplificar a um interessado o processo para definição de técnicas a serem consideradas no desenvolvimento de um sistema interpretável.

1.2 Objetivos

O objetivo geral deste trabalho é contribuir com o entendimento da área de interpretabilidade de Redes Neurais Convolucionais, ao integrar conhecimentos presentes na literatura na forma de revisão bibliográfica e estudo de caso. Para isso, fez-se necessário atingir os seguintes objetivos específicos:

- Revisar a literatura sobre técnicas de interpretabilidade de Redes Neurais Convolucionais;
- Revisar a literatura sobre avaliações de técnicas de interpretabilidade;
- Treinar modelos que realizam diagnóstico por imagem;
- Utilizar técnicas de interpretabilidade para explicar as predições dos modelos treinados;
- Comparar explicações, replicando avaliações quantitativas presentes na literatura;
- Analisar os resultados do estudo de caso, de modo a levantar recomendações de treinamento, explicação e comparação entre avaliações.

1.3 Contribuições

O presente trabalho tem um caráter integrativo, organizando conhecimentos na literatura e realizando um estudo de caso completo. Em resumo, as principais contribuições deste trabalho são:

- Revisão bibliográfica sobre interpretabilidade de Redes Neurais Convolucionais;
- Compilação das avaliações para validação e comparação de técnicas de interpretabilidade;
- Estudo de caso com treinamento, explicação e comparação entre técnicas de explicação, no nicho de diagnóstico por imagens.

1.4 Estrutura do Documento

Este documento está organizado da seguinte forma: no Capítulo 2 são introduzidos os conceitos fundamentais e o arcabouço teórico necessário à compreensão desta pesquisa. O Capítulo 3 apresenta uma revisão bibliográfica sobre aprendizado de máquina e interpretabilidade de Redes Neurais Convolucionais. No Capítulo 4 estão apresentados os materiais utilizados e a metodologia relacionada ao estudo de caso conduzido. O Capítulo 5 apresenta os resultados obtidos a partir do estudo de caso. No Capítulo 6, as conclusões da pesquisa e as perspectivas para trabalhos futuros são discutidas.

2 Fundamentação

Nas seções seguintes estão apresentados conceitos gerais sobre aprendizado de máquina, redes neurais e interpretabilidade. Na Seção 2.1 estão apresentados os conceitos de aprendizado de máquina e redes neurais. Na Seção 2.2 está introduzido o campo de estudo de interpretabilidade que será aprofundado nos capítulos posteriores.

2.1 Aprendizado de Máquina e Redes Neurais

No contexto de Inteligência Artificial, o interesse por Aprendizado de Máquina (*Machine Learning* ou ML) explodiu na última década. *Softwares* de detecção de *spam*, sistemas de recomendação, marcação em fotos de redes sociais, assistentes pessoais ativados por voz, carros autônomos, reconhecimento facial e muitas outras aplicações têm sido desenvolvidas e estão atingindo níveis de estado da arte devido às recentes inovações. Existem três principais abordagens para o Aprendizado de Máquina, sendo elas o aprendizado supervisionado, não supervisionado e por reforço, cujas diferenças estão na categoria de dados que utilizam. Na abordagem supervisionada a máquina aprende a partir de exemplos rotulados, ao contrário da abordagem não supervisionada, que não necessita de rótulos; por sua vez, na abordagem por reforço a máquina aprende por meio de um mecanismo de recompensa e punição de suas decisões com base no resultado de ações sobre o ambiente do problema. Na Figura 1 estão apresentadas as abordagens de aprendizado e aplicações habitualmente associadas a cada um dos tipos.

Entre os ramos existentes no contexto de Aprendizado de Máquina, o Aprendizado Profundo (*Deep Learning* ou DL) é o mais proeminente. Soluções que utilizam DL têm sido desenvolvidas com grande sucesso para resolver problemas em uma variedade de campos, como preprocessamento digital de imagem, visão computacional, processamento de linguagem natural, segurança e tantos outros.

Deep Learning é uma forma de aprendizado de representação — em que uma máquina é alimentada com dados brutos e desenvolve as representações necessárias para o reconhecimento de padrões. O conhecimento aprendido é organizado em várias camadas de representações. Essas camadas são normalmente organizadas sequencialmente e compostas por um grande número de operações primitivas não lineares, de modo que o conhecimento de uma camada alimenta a próxima camada, transformando essa representação em uma representação mais abstrata. O fluxo de representações transforma a entrada até que as informações se tornem distinguíveis, permitindo o aprendizado de funções altamente complexas (ESTEVA et al., 2019).

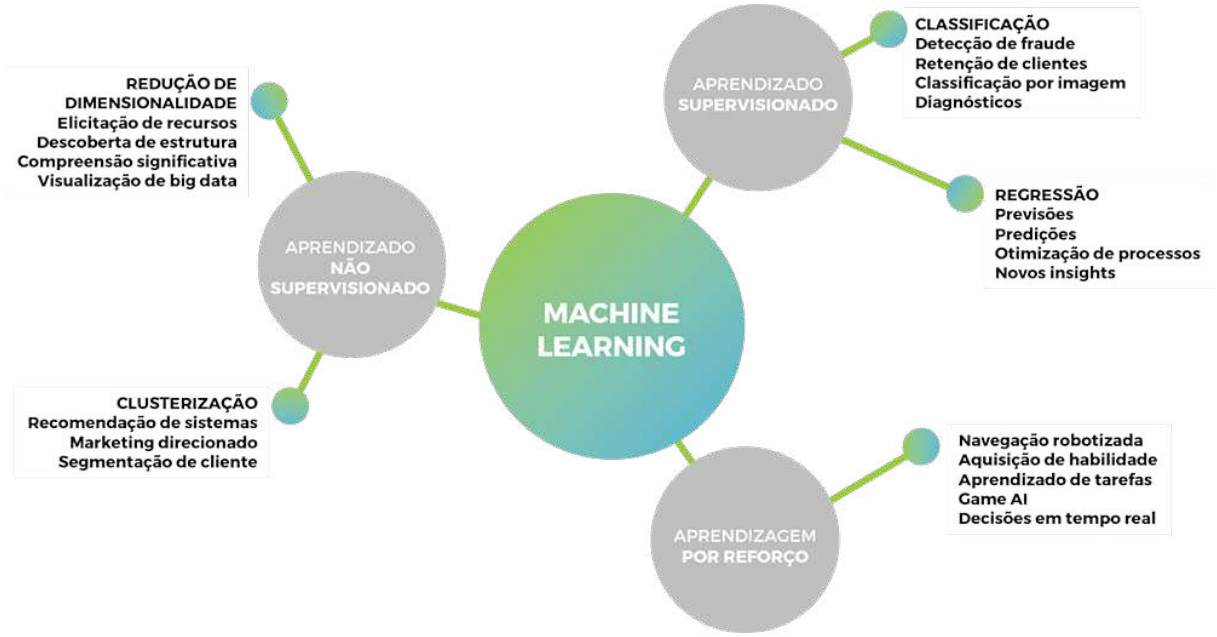


Figura 1 – Diferentes abordagens para o Aprendizado de Máquina e aplicações. Fonte: Deal Labs (LABS, 2018)

Deep Learning incorpora e expande o conceito dos modelos neurais, modelos biologicamente inspirados compostos por neurônios e camadas. O que caracteriza DL é a quantidade de camadas e a capacidade de recuperar os padrões necessários para a decisão a partir dos dados brutos, sem a necessidade de um processamento de dados manualmente ajustado. Existem diferentes categorias de modelos neurais profundos, incluindo modelos convolucionais, recorrentes e generativos adversariais. Os modelos de aprendizado profundo escalam bem para grandes conjuntos de dados, em parte devido à sua capacidade de execução em *hardware* de computação especializado, como placas de vídeo, e continuam a melhorar com mais dados, permitindo-lhes superar muitas abordagens clássicas de ML.

Ao treinar uma rede neural, algumas técnicas podem ser utilizadas para acelerar o treinamento e obter uma melhoria na acurácia ou desempenho. Técnicas como transferência de conhecimento (*transfer learning*) (PAN; YANG, 2009), ajuste fino (*fine tuning*) e aumento de dados (*data augmentation*) (DYK; MENG, 2001) são comumente utilizadas no treinamento de redes neurais profundas. A transferência de conhecimento se dá quando um modelo já treinado é utilizado em um contexto diferente, mas similar ao contexto original, como ponto de partida para o treinamento ou como extrator de características, permitindo um reuso do resultado de treinamento. Uma etapa comum após a transferência de conhecimento é o ajuste fino dos pesos das camadas mais profundas da rede, deixando as camadas iniciais, comumente associadas a extratores de características, intactas.

O aumento de dados é uma técnica utilizada para expandir artificialmente o conjunto de treinamento, ao reinserir no conjunto exemplos modificados. No contexto de

imagens, mudanças de brilho, contraste e orientação são algumas das muitas modificações possíveis ao realizar um aumento de dados. Um maior volume de dados, mesmo que incrementado de maneira artificial, tende a promover uma maior robustez do modelo na fase de uso.

2.2 Interpretabilidade em Aprendizado de Máquina

Para compreender as seções seguintes é relevante introduzir conceitos gerais, objetivos e categorias de interpretabilidade no domínio do aprendizado de máquina supervisionado.

2.2.1 Introdução

O procedimento padrão do aprendizado de máquina supervisionado é utilizar um conjunto de treinamento para produzir modelos de classificação ou regressão e um conjunto de teste para gerar métricas, com as quais é possível estimar o poder de generalização do modelo treinado. Quando apenas predições e métricas não são suficientes para entender como o modelo se comporta, recorre-se a técnicas de interpretabilidade para a produção de artefatos que expliquem o funcionamento do modelo ou seu conhecimento interno, de forma compreensível para humanos.

Quando a tarefa de aprendizado de máquina envolve utilizar dados tabulares, aplicam-se diversos modelos intrinsecamente interpretáveis, tais como Modelos Lineares, Árvores de Decisão e K Vizinhos mais Próximos. Esses modelos contêm em seus parâmetros e organização as explicações que tornam a tomada de decisão rastreável. Além de modelos interpretáveis existem técnicas que ajudam a entender a importância e possíveis relações entre as características das entradas, como PDP (*Partial Dependence Plot*) (FRIEDMAN, 2001), ICE (*Individual Conditional Expectation*) (GOLDSTEIN et al., 2015) e ALE (*Accumulated Local Effects*) (APLEY; ZHU, 2020). Muitas dessas técnicas envolvem a geração de gráficos que informam a importância e a interdependência entre as características da entrada. Em Molnar (2019), diversas técnicas de interpretabilidade de modelos treinados a partir de dados tabulares estão apresentadas.

Existe um grande desafio em explicar a importância e os possíveis relacionamentos entre as características utilizadas por modelos treinados a partir de dados tabulares, mas existe um desafio maior em explicar modelos que extraem dos dados as características que dão suporte à decisão, como os modelos neurais profundos com aplicação em imagem ou texto. Extratores de características projetados manualmente têm sido substituídos por modelos neurais em aplicações que utilizam sinais e imagens (em contraposição a dados tabulares) — neste caso, os extratores de características passam a fazer parte do processo de treinamento.

Existe uma tendência em que modelos mais complexos, com mais camadas, são aqueles que atingem as maiores taxas de acerto nos desafios de aprendizado de máquina. Como exemplo, o modelo vencedor do desafio de classificação de imagens *Image Net* (ILSVRC (RUSSAKOVSKY et al., 2015)), no ano de 2012, AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) utiliza 8 camadas, já o modelo vencedor em 2015, ResNet (HE et al., 2016), conta com 152 camadas. A maior complexidade e a inclusão dos extratores de características no treinamento são fatores que afetam a interpretabilidade destes modelos. Técnicas de interpretabilidade são ferramentas que podem mitigar essa desvantagem e tornar modelos neurais que alcançam acurácias satisfatórias também interpretáveis.

A Seção 2.2.2 organiza objetivos comuns para as técnicas de interpretabilidade e características comuns encontradas em modelos interpretáveis. A Seção 2.2.3 trata da categorização das técnicas conforme o funcionamento.

2.2.2 Classificação dos Objetivos e Características

"O termo interpretabilidade não tem significado amplamente aceito pela comunidade especializada, no entanto, muitos autores recorrem ao termo de maneira descuidada" (LIPTON, 2018). Para determinar qual é o sentido pretendido na utilização do termo, deve-se questionar o objetivo da interpretabilidade em cada uso. Termos como transparência, confiança, explicabilidade e auditabilidade têm significados mais estritos no contexto de interpretabilidade, por aludirem a objetivos ou características específicas, explicadas a seguir.

A Figura 2 apresenta categorizações de objetivos relacionados ao uso do termo interpretabilidade conforme descrito em dois artigos (LIPTON, 2018; DOSHI-VELEZ; KIM, 2017). Os objetivos descritos pelos autores estão reclassificados em três categorias abstratas, detalhadas a seguir:

- **Ética e Regulação:** Corresponde à utilização de interpretabilidade para atingir conformidade com regulação e padrões éticos. Utilizar interpretabilidade para adequação às leis que asseguram direito de explicação é um exemplo de uso de interpretabilidade com objetivos éticos ou de regulação. A interpretabilidade está presente no debate acadêmico e em questões regulatórias. O debate sobre a utilização de modelos em tarefas críticas é recorrente na literatura, como no artigo de Rudin (2019), que desencoraja o uso de técnicas de interpretabilidade para explicar modelos complexos e propõe a utilização de modelos intrinsecamente interpretáveis para cenários de risco. A diretiva GDPR (*General Data Protection Regulation*), imposta em maio de 2018 pela União Europeia, concede direito à explicação para todas as decisões tomadas por sistemas automatizados ou sistemas algorítmicos artificialmente inteli-

gentes podendo forçar a conformidade por empresas que atuam na União Europeia (GOODMAN; FLAXMAN, 2017).

- **Apoio à Decisão:** Condiz à utilização de interpretabilidade na fase de uso do modelo de aprendizado de máquina. Em cenários de uso críticos ou nebulosos, técnicas de interpretabilidade podem retornar informações para que o usuário entenda a razão por trás da tomada de decisão, o que torna dinâmico o uso do modelo por parte do usuário. Técnicas que atendem a esse objetivo são úteis na fase de uso do modelo. O campo de diagnóstico auxiliado por computador (*Computer Aided Diagnosis* ou CAD) é um campo frequentemente conectado com o campo de interpretabilidade, como no artigo de Kim et al. (2018b), que apresenta um protótipo *web* de diagnóstico de glaucoma com explicação fornecida por uma técnica de interpretabilidade chamada Grad-CAM (SELVARAJU et al., 2017).

- **Confiança e Entendimento:** Expor o funcionamento e conhecimento internos utilizados por um modelo estabelece confiança na capacidade preditiva do modelo e ajuda o entendimento do problema e de sua solução. Utilização da interpretabilidade primordialmente na fase de modelagem e entendimento do problema. Técnicas que atendem a esse objetivo são úteis na fase de validação e depuração do modelo. Por exemplo, no artigo de Ribeiro, Singh e Guestrin (2016), é apresentado um viés em um modelo que classifica lobos entre outras classes, em que a neve presente no fundo da imagem é considerada na decisão de classificação do modelo treinado. O desenvolvimento de bibliotecas como *CheckList* (RIBEIRO et al., 2020) e *Innvestigate* (ALBER et al., 2019) torna acessível realizar testes de sanidade em modelos utilizando interpretabilidade e não apenas levantar as métricas de acurácia. No artigo de Wu et al. (2018) foram identificadas e rotuladas diversas representações internas utilizadas por redes neurais profundas treinadas para classificação de tecido doente em mamografias 2D que se alinham com o BI-RADS, uma forma padronizada de relatar os achados radiológicos da mamografia. Tais representações internas foram associadas com conceitos conhecidamente importantes (como massa de nódulos, calcificação e composição da mama) para a classificação, dando assim respaldo para o modelo neural treinado.

Termos comumente relacionados à interpretabilidade, como transparência, explicabilidade e auditabilidade se referem a características que um modelo pode possuir. A presença dessas características viabilizam os objetivos anteriormente apresentados. Nos tópicos a seguir, esses termos são melhor contextualizados e relacionados com os objetivos apresentados:

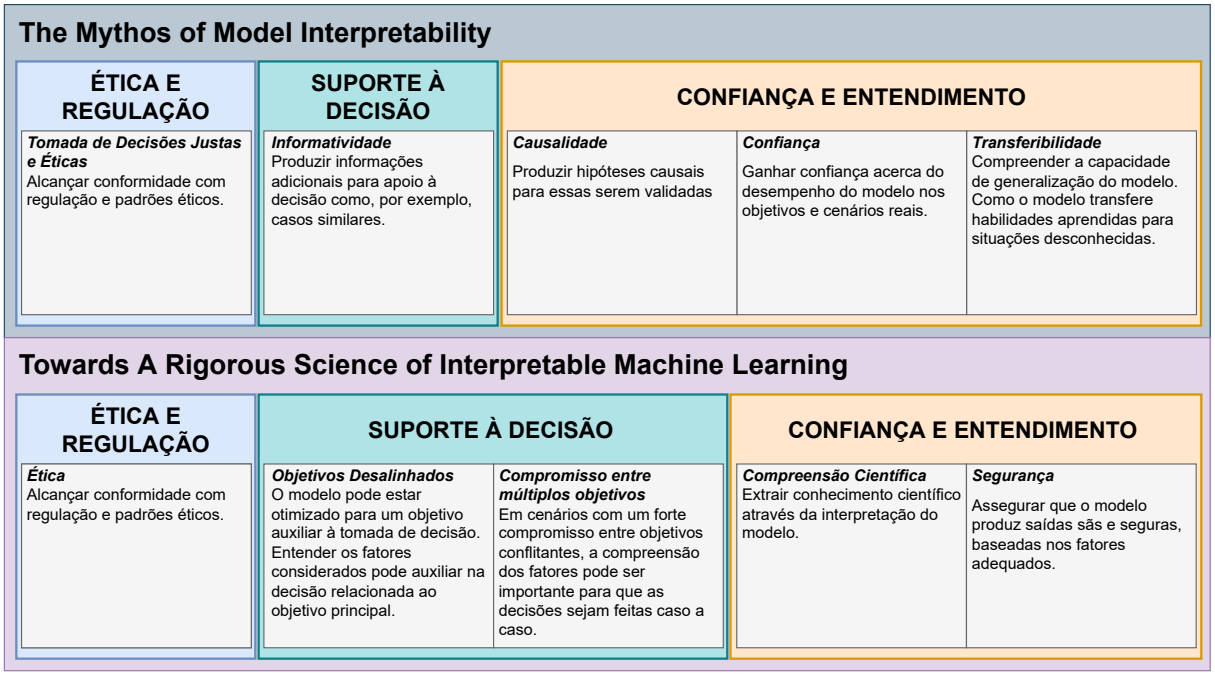


Figura 2 – Objetivos relacionados ao uso da interpretabilidade segundo os artigos *Towards A Rigorous Science of Interpretable Machine Learning* (DOSHI-VELEZ; KIM, 2017) e *The Mythos of Model Interpretability* (LIPTON, 2018) com agrupamento proposto. Fonte: Elaboração própria do autor.

- **Auditabilidade:** Modelos auditáveis são necessários para atender agentes reguladores e serem utilizados em aplicações críticas, como medicina, direito e aeronáutica. A auditabilidade está relacionada com os objetivos categorizados por ética e regulação.
- **Explicabilidade:** Modelos explicáveis têm a capacidade de fornecer explicações ao usuário e estão relacionados com os objetivos categorizados por apoio à decisão.
- **Transparência:** Modelos transparentes podem ser apresentados em termos compreensíveis por humanos e estão relacionados com os objetivos categorizados por confiança e entendimento.

A finalidade das classificações propostas é enquadrar os objetivos e características para facilitar a compreensão do campo da interpretabilidade. As classificações não são absolutas, um modelo pode estabelecer confiança ao fornecer boas explicações, mesmo que suas representações internas continuem complexas e incompreensíveis.

2.2.3 Classificação Geral das Técnicas de Interpretabilidade

Existem na literatura categorizações para técnicas de interpretabilidade que informam o escopo e o funcionamento da técnica, discutidas a seguir.

- **Intrínsecas ou Pós Treinamento:** Técnicas de interpretabilidade podem ser classificadas como intrínsecas ou pós treinamento (*Intrinsic* ou *Post-hoc*). As técnicas classificadas como intrínsecas se utilizam do fato de que alguns modelos são interpretáveis por si só. Modelos intrinsecamente interpretáveis abrigam em seus parâmetros e organização elementos necessários para a compreensão. Apesar de existirem Redes Neurais que tentam retornar uma explicação junto à sua predição, métodos de interpretabilidade de redes neurais são usualmente executados após o treinamento.

- **Local ou Global:** O escopo das técnicas de interpretabilidade pode ser local ou global. Técnicas globais inspecionam o comportamento mais amplo do modelo e estão mais relacionadas à transparência. Técnicas locais esclarecem uma decisão individual e estão relacionadas à explicabilidade.

- **Agnóstica a Modelo ou Específica a Modelo:** A abrangência da técnica pode ser classificada como agnóstica a modelo ou específica a modelo (*Model Agnostic)* ou *Model Specific*). Técnicas específicas são limitadas a uma determinada classe de modelos. Técnicas agnósticas independem da arquitetura.

2.3 Considerações Finais

A interpretabilidade, portanto, é uma área de pesquisa que atende diversos objetivos relacionados ao aumento do entendimento e interação com modelos complexos. Por ser um campo recente, é importante apresentar e esclarecer os termos utilizados na literatura e construir um vocabulário para facilitar o aprofundamento das discussões.

O termo interpretabilidade é tido como pouco elucidativo. É importante esclarecer o objetivo do uso da interpretabilidade em cada caso para que modelos ou técnicas que possuam ou promovam características específicas sejam levantados para alcançar tais objetivos. A Figura 3 apresenta as categorias para o funcionamento das técnicas de interpretabilidade e também a relação entre objetivos e características específicas, como por exemplos, auditabilidade e confiança e entendimento estão conectados.

Os avanços na área de interpretabilidade são cruciais para a disseminação do uso do aprendizado de máquina em novas aplicações. Os maiores ganhos e oportunidades residem em áreas críticas, como sistema judicial (BIBAL et al., 2021) e medicina (TJOA; GUAN, 2019). Uma fundamentação da área está disponível em Molnar (2019) e uma discussão sobre as oportunidades de pesquisa pode ser encontrada em Das e Rad (2020).

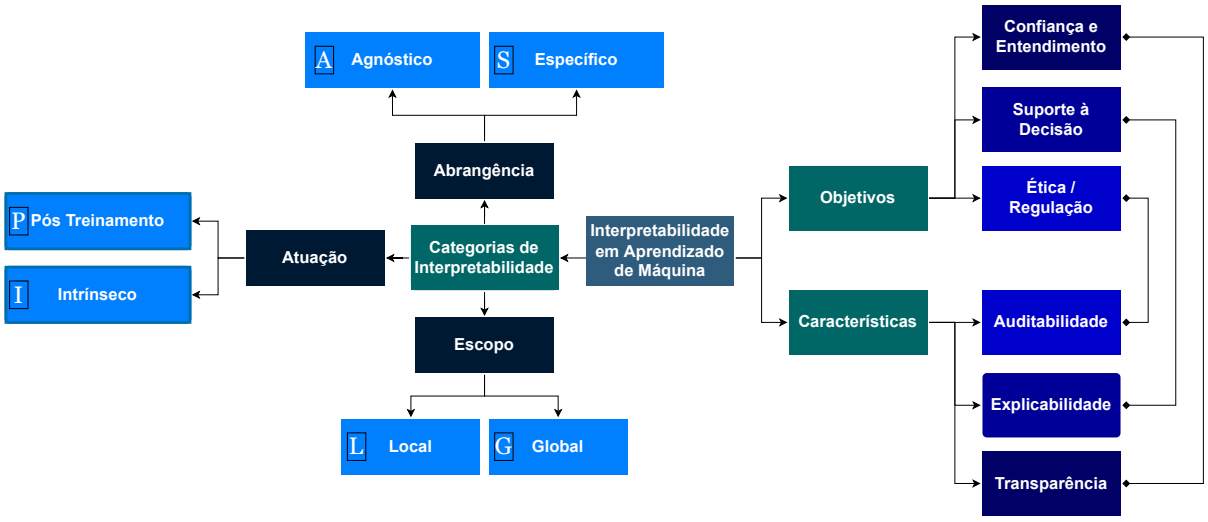


Figura 3 – Diagrama apresentando taxonomia proposta para Interpretabilidade em Aprendizado de Máquina. **Esquerda:** Categorias para técnicas de interpretabilidade. **Direita:** Objetivos e características relacionados à interpretabilidade. Fonte: próprio autor.

3 Revisão Bibliográfica

A complexidade proveniente do aprendizado cada vez mais profundo e hierárquico (WU; SHEN; HENGEL, 2019) é um obstáculo à maior aceitação de DL em áreas críticas como medicina, mercado financeiro e sistema de justiça criminal (RUDIN, 2019), onde os resultados das inferências podem trazer grande impacto na vida das pessoas. A interpretabilidade é uma tendência em voga que promete ajudar na difusão de DL tornando os modelos de aprendizado de máquina capazes de apresentar as informações úteis para explicar predições individuais ou expor, em termos compreensíveis por humanos, o seu funcionamento e conhecimento internos.

A interpretabilidade pode ajudar a esclarecer o conhecimento adquirido por modelos de *Deep Learning* em aplicações envolvendo diferentes domínios de dados, como dados tabulares, texto e imagem. O foco dessa revisão bibliográfica está no domínio das imagens e em Redes Neurais Convolucionais Profundas (*Deep Convolutional Neural Network* DCNN), que são as redes mais utilizadas para aprendizado a partir de imagens.

O estudo de fatores que influenciam os modelos utilizados em tarefas visuais tem sido abordado por várias pesquisas na literatura a partir de técnicas que fornecem explicações também visuais, como mapas de calor, utilizados como explicação por técnicas como Grad-CAM (SELVARAJU et al., 2017) e RISE (PETSIUK; DAS; SAENKO, 2018). O caráter visual torna a interpretabilidade de redes convolucionais especialmente desafiadora, já que existe uma diferença entre a compreensão visual emulada por modelos convolucionais e a compreensão visual humana. Durante a fase de treinamento, modelos convolucionais podem utilizar características incompreensíveis por humanos (ILYAS et al., 2019).

Analisando a literatura recente, percebe-se que o campo de interpretabilidade de redes neurais convolucionais possui muitos debates científicos em aberto e muitos artigos alimentam o debate ao apontar falhas em algumas das abordagens. Esta revisão, além de reunir referências bibliográficas relevantes e propor uma visão e taxonomia próprias para a área de interpretabilidade, apresenta uma discussão dos artigos que direcionam críticas a abordagens existentes.

3.1 Visão Geral das Abordagens para Interpretabilidade de Redes Neurais Convolucionais

É comum entender as camadas convolucionais de uma CNN como camadas poderosas na tarefa de extração de características, e as camadas finais como camadas que

tomam decisões embasadas pelas características extraídas. Em vista disso, os métodos de interpretabilidade de CNN mais comuns seguem duas abordagens principais, quais sejam: visualização de características e atribuição de características. Ambas buscam expor quais foram as características aprendidas pelo modelo treinado. Existem aplicações que tentam unir ambas as abordagens em uma única interface, a exemplo da pesquisa de Carter et al. (2019), para produzir uma visualização global voltada ao entendimento da rede. Existem outras abordagens que não tentam expor características aprendidas, mas sim interpretar o modelo através de outras abordagens, como resgatar os exemplos de treinamento mais influentes para cada predição. A divisão comentada nesta seção, assim como onde se encontra o subcampo de interpretabilidade de CNN, está ilustrada na Figura 4.

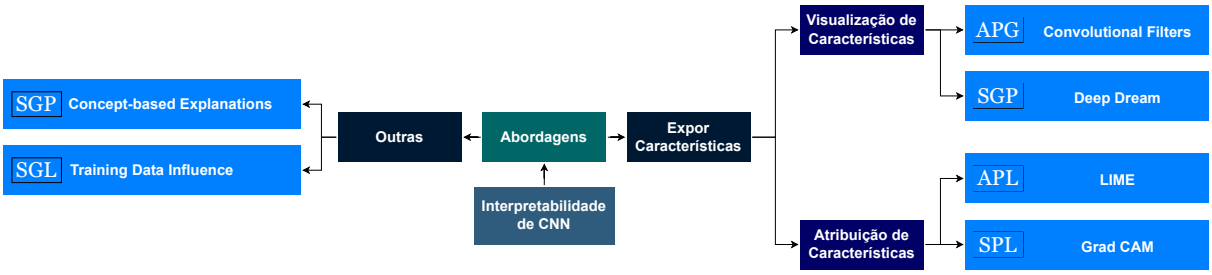


Figura 4 – Diagrama apresentando taxonomia proposta para Interpretabilidade em Aprendizado de Máquina. Cada técnica presente no diagrama foi classificada utilizando a classificação apresentada na Figura 3. Legenda para os símbolos utilizados para categorizar as técnicas: S=Específico, A=Agnóstico, P=Pós-Treinamento, I=Intrínseco, L=Local, G=Global. Fonte: Elaboração própria do autor.

Utilizar as informações contidas nas camadas internas de uma rede para produzir imagens que induzem erros nessas redes neurais é uma maneira de produzir um ataque adversarial (NGUYEN; YOSINSKI; CLUNE, 2014a). Através de ataques adversariais, imagens podem ser perturbadas de maneira imperceptível por humanos, no entanto, as redes neurais classificam essas imagens erroneamente com alta probabilidade. Esses chamados exemplos adversariais provocam grande incerteza sobre as redes neurais convolucionais e ainda são debatidos no meio científico. Inspiradas pelos ataques adversariais, as técnicas de interpretabilidade também se valem de informações como pesos e gradientes, mas com um propósito diferente, expor padrões ou regiões que mais influenciam na classificação, produzindo, desta forma, explicações visuais.

Nas subseções a seguir, além de uma discussão sobre as duas abordagens dominantes para interpretabilidade da CNN (visualização e atribuição de características), também incluímos outras abordagens, como projetar modelos intrinsecamente interpretáveis.

3.1.1 Atribuição de Características

Existem diversas técnicas que permitem extrair explicações locais de redes neurais ao se atribuir responsabilidade às características da entrada. Os métodos focados em

fornecer explicações sob a ótica de atribuição de características relacionam regiões das imagens com a sua importância para o modelo. Uma visualização comum dessa importância é realizada por mapas de calor, nesse contexto nominados mapas de atribuição, que indicam as regiões que contribuem positiva e negativamente para a classificação.

Existem duas abordagens principais para gerar mapas de atribuição de características, a abordagem utilizando perturbação e a abordagem utilizando retropropagação de gradientes. Na abordagem que utiliza perturbação, para cada entrada do modelo, alteram-se características individuais (pixels, super pixels, etc.) na forma de remoção ou perturbação e então verifica-se o impacto de cada característica individual na predição do modelo. Essas técnicas não consideram os pesos internos da rede neural e podem ser aplicadas a uma grande variedade de redes neurais. Técnicas como LIME (RIBEIRO; SINGH; GUESTRIN, 2016), *Occlusion* (ZEILER; FERGUS, 2014) e RISE (PETSIUK; DAS; SAENKO, 2018) são técnicas que seguem a abordagem de perturbação e diferem na maneira que alteram a entrada e na maneira como calculam a importância da perturbação.

As abordagens baseadas na retropropagação de gradientes calculam a responsabilidade de cada *pixel* na predição ao propagar um sinal de importância partindo das camadas finais até atingir a camada de entrada. As técnicas que seguem a abordagem da retropropagação precisam considerar os pesos internos e são mais eficientes por não precisar lidar com a ponderação entre diversas perturbações, no entanto, existem grandes desafios para a conservação do sinal de importância que percorre vários neurônios até alcançar a entrada da rede. Técnicas baseadas em gradientes são específicas a modelos, no entanto, são apoiadas por um grande conjunto de arquiteturas.

A principal diferença entre as várias técnicas baseadas em atribuição está na maneira em como propagar um sinal de importância das camadas finais até a imagem de entrada, formando um mapa de calor na imagem, atribuindo importância para determinadas regiões que contenham os padrões que se mostraram importantes para a predição de acordo com os pesos internos da rede. Os empecilhos para a conservação dessa propagação são de ordem matemática. Uma rede neural é composta por muitas funções de ativação encadeadas e navegar por essas funções pode ser um entrave para análises de sensibilidade. As funções de ativação comumente utilizadas nas redes neurais são não lineares, apresentando muitas vezes formatos irregulares onde a função é ativada apenas a partir de determinado valor limite, a exemplo da função ReLU, comumente encontrada em redes neurais. A depender da faixa de valor, essas funções podem retornar um valor de saída que pode não se alterar de maneira proporcional ao valor recebido na entrada. Os desafios matemáticos de se analisar a sensibilidade através de muitas dessas funções encadeadas são conhecidos e estão detalhados de maneira formal nos artigos de Sundararajan, Taly e Yan (2017) e Shrikumar, Greenside e Kundaje (2017), nos quais as técnicas *Integrated Gradients* e *DeepLIFT* são respectivamente apresentadas. A Figura 5 ilustra o resultado

de diferentes técnicas baseadas em atribuição em que cada coluna apresenta uma técnica diferente, onde é possível perceber que, enquanto algumas técnicas utilizam mapas de calor para apresentar a atribuição, outras utilizam regiões segmentadas.

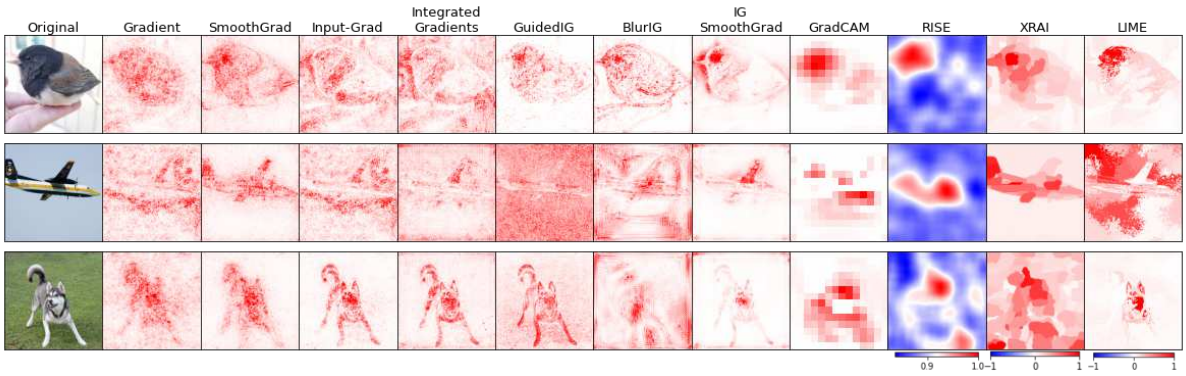


Figura 5 – Evolução de diferentes técnicas baseadas em gradientes e em pertubações. Técnicas de atribuição adotam diferentes estratégias para atribuir importância a regiões da entrada. As técnicas LIME (RIBEIRO; SINGH; GUESTRIN, 2016) e XRAI (KAPISHNIKOV et al., 2019) utilizam *superpixels*. Gradients (SIMONYAN; VEDALDI; ZISSERMAN, 2014), Input-Grad (SHRIKUMAR; GREENSIDE; KUNDAJE, 2017) Integrated Gradients (SUNDARARAJAN; TALY; YAN, 2017) e Guided Integrated Gradients (SUNDARARAJAN; TALY; YAN, 2017) atribuem importância para cada *pixel* da imagem. Grad-CAM (SELVARAJU et al., 2017) e RISE (PETSIUK; DAS; SAENKO, 2018) atribuem importância para macro regiões suavizadas. Fonte: (DAS; RAD, 2020)

Experimentos que tentam verificar a validade dessas explicações evidenciaram limitações na construção e utilidade de determinadas técnicas que seguem a abordagem de atribuição de características. Na pesquisa de Ghorbani, Abid e Zou (2019) discute-se que muitas dessas abordagens podem ser frágeis a ataques adversariais, ou seja, imagens perturbadas de maneira imperceptível por humanos geram mapas de atribuição muito diferentes. Outro artigo (ADEBAYO et al., 2018) conduziu testes com diferentes técnicas de atribuição de características nos quais algumas técnicas retornaram explicações qualitativamente e quantitativamente semelhantes tanto para modelos treinados quanto para modelos com pesos e dados de treinamento randomizados. Os resultados desses experimentos funcionam como testes de sanidade e põem em xeque algumas técnicas que produzem mapas de atribuição que são insensíveis aos pesos do modelo e sensíveis a detalhes que, para os humanos, são imperceptíveis e algumas vezes irrelevantes para o problema.

3.1.2 Visualização de Características

"Se quisermos descobrir que tipo de entrada causaria um certo comportamento, seja um disparo de neurônio interno ou o comportamento final, podemos usar derivadas para ajustar iterativamente a entrada em direção a esse objetivo" (OLAH; MORDVINTSEV; SCHUBERT, 2017). Isso pode ser feito como um problema de otimização, calcu-

lando o gradiente da função de ativação dada uma imagem inicializada aleatoriamente e mudando a imagem na direção do gradiente (ERHAN et al., 2009).

Otimizar imagens para que elas maximizem uma saída alvo na rede (ou objetivo de otimização) é a abordagem mais comum e foi introduzida como *Activation Maximization* por Erhan et al. (2009). Diferentes objetivos de otimização têm como alvo partes diferentes de uma rede neural, como neurônios individuais ou camadas, e sintetizam imagens que maximizam a ativação dessas partes.

A abordagem ingênua de ajustar iterativamente uma imagem gerada aleatoriamente para maximizar a ativação de um neurônio resulta em imagens com muito ruído de alta frequência que são pouco inteligíveis (NGUYEN; YOSINSKI; CLUNE, 2014b). Produzir imagens sem ruídos de alta frequência, inteligíveis por humanos, é o principal desafio no campo de visualização de características. Para orientar a otimização e obter visualizações úteis é necessário impor uma estrutura mais natural para as imagens sintetizadas utilizando informações prévias (*prior*), regularizadores (*regularizer*) ou restrições (*constraint*).

Em (OLAH; MORDVINTSEV; SCHUBERT, 2017), as imposições para a produção de imagens com estruturas mais naturais estão categorizadas em três famílias:

- **Penalização de Frequência (Frequency penalization):** "ataca diretamente o ruído de alta frequência que esses métodos sofrem." (OLAH; MORDVINTSEV; SCHUBERT, 2017). Pode-se impor penalidade aos pixels com alta intensidade (MAHENDRAN; VEDALDI, 2015) ou penalizar o ruído de alta frequência implicitamente, utilizando desfoques em cada etapa de otimização (NGUYEN; YOSINSKI; CLUNE, 2014a).

- **Robustez de transformação (Transformation robustness):** Transformar a imagem sendo otimizada na tentativa de encontrar exemplos que ativam de maneira elevada o alvo de otimização, mesmo se transformados levemente. Concretamente, significa aplicar *jitter*, rotação e dimensionamentos estocasticamente na imagem antes de aplicar a etapa de otimização. (OLAH; MORDVINTSEV; SCHUBERT, 2017)

- **Informação prévia via aprendido (Learned priors):** Uma próxima etapa é aprender um modelo com dados reais e utilizar o modelo para impor uma estrutura mais natural. Em algumas abordagens, a imagem não é otimizada diretamente, mas através de uma rede generativa treinada, como um GAN (*Generative Adversarial Network*) ou VAE (*Variational Autoencoder*), onde o espaço vetorial de baixa dimensão é otimizado para gerar imagens mais naturais (NGUYEN et al., 2016).

A visualização de características teve um grande progresso nos últimos anos. Foram desenvolvidas novas técnica para criar visualizações atraentes (OLAH; MORDVINTSEV; SCHUBERT, 2017), no entanto, em algumas técnicas fica pouco claro se a forte imposição para síntese de estruturas naturais através de regularizadores torna a imagem infiel ao conhecimento aprendido pelo modelo e fiel ao regularizador. Esse problema é debatido em (OLAH; MORDVINTSEV; SCHUBERT, 2017) e (GHORBANI; WEXLER; KIM, 2019). Abordagens de visualização de características são discutidas em maiores detalhes em (NGUYEN; YOSINSKI; CLUNE, 2019).

3.1.3 Outras Abordagens

Nesta seção são discutidas outras abordagens que não interpretam modelos pela ótica de características. Além de métodos de interpretabilidade para as redes convolucionais convencionais existe também um esforço para o desenvolvimento de arquiteturas que sejam intrinsecamente interpretáveis. Há também técnicas que combinam outras informações importantes no treinamento das redes neurais e não se apoiam apenas em uma imagem de teste, mas em um conjunto de imagens ou nos exemplos de treinamento. Tais métodos podem ser encontrados em pesquisas que exploram exemplos de treinamento, protótipos e conceitos, conforme comentado a seguir.

Os modelos convolucionais que utilizam protótipos armazenam características através do treinamento em protótipos e então comparam as entradas com os protótipos aprendidos (ARIK; PFISTER, 2020). Apesar de obterem boa acurácia e proverem explicações locais, esses modelos necessitam de grande poder de armazenamento e de processamento nas fases de treinamento e de uso.

As técnicas que utilizam exemplos de treinamento explicam localmente decisões atribuindo responsabilidade a exemplos presentes no conjunto de treinamento. Uma abordagem simples é realizar uma busca no conjunto de treinamento por imagens semelhantes a uma imagem de teste utilizando métricas de similaridade. Na pesquisa de Koh e Liang (2017) o modelo é tratado como uma função dos dados de treinamento e então é calculada a importância de cada imagem de treino com respeito à imagem de teste. A inspiração para Koh e Liang (2017) é uma técnica da estatística robusta, chamada funções de influência (HAMPEL, 1974), utilizada para calcular a influência de observações individuais em distribuições estatísticas, por exemplo, a influência de uma medida individual na média das medidas.

Técnicas que retornam conceitos utilizam conjuntos de imagens de uma mesma classe com o objetivo de encontrar características de alto nível, compreensíveis por humanos, importantes para classificação. O conjunto de conceitos retornado têm o objetivo de fornecer uma explicação global do modelo. Conceitos são segmentos da informação visual (grupos de *pixels*) sobre os quais humanos possuem familiaridade, como, por exemplo,

numa tarefa de reconhecimento de carros, um recorte da imagem contendo uma das rodas do carro. O artigo (GHORBANI; WEXLER; KIM, 2019) discute essa abordagem e propõe uma técnica para buscar conceitos de forma automática.

3.2 Detalhamento de Técnicas para Interpretabilidade de Redes Neurais Convolucionais

Seguindo as classificações apresentadas nas Seções 2.2.2 e 3.1, o Quadro 1 organiza diferentes técnicas de interpretabilidade pós-treinamento e esforços para produzir modelos intrinsecamente interpretáveis. As técnicas presentes na tabela são detalhadas nas próximas subseções.

Quadro 1 – Apanhado e classificação de técnicas de interpretabilidade de Redes Neurais Convolucionais.

Técnica de Interpretabilidade	Referência	Local	Global	Agnóstico	Específico	Intrínseco	Pós Treinamento	Atribuição	Visualização	Outra Abordagem
Activation Maximization	(ERHAN; COURVILLE; BENGIO, 2010)	✓		✓			✓		✓	
Gradient-based Saliency Maps	(SIMONYAN; VEDALDI; ZISSERMAN, 2014)	✓		✓			✓	✓		
Guided Backprop	(SPRINGENBERG et al., 2014)	✓		✓			✓	✓		
Integrated Gradients	(SUNDARARAJAN; TALY; YAN, 2017)	✓		✓			✓	✓		
DeepLIFT	(SHRIKUMAR; GREENSIDE; KUNDAJE, 2017)	✓		✓			✓	✓		
SmoothGrad	(SMILKOV et al., 2017)	✓		✓			✓	✓		
CAM	(ZHOU et al., 2016)	✓		✓			✓	✓		
Grad-CAM	(SELVARAJU et al., 2017)	✓		✓			✓	✓		
Grad-CAM++	(CHATTOPADHAY et al., 2018)	✓		✓			✓	✓		
LIME	(RIBEIRO; SINGH; GUESTRIN, 2016)	✓		✓			✓	✓		
RISE	(PETSIUK; DAS; SAENKO, 2018)	✓		✓			✓	✓		
SHAP	(LUNDBERG; LEE, 2017)	✓		✓			✓	✓		
Influence Functions	(KOH; LIANG, 2017)	✓		✓			✓			✓
Testing with Concept Activation Vectors	(KIM et al., 2018a)		✓	✓			✓			✓
Automatic Concept-based Explanations	(GHORBANI; WEXLER; KIM, 2019)		✓	✓			✓			✓
CaCE	(GOYAL; SHALIT; KIM, 2019)		✓	✓			✓			✓
ConceptSHAP	(YEH et al., 2019)		✓	✓			✓			✓
Zhang Explanatory Graph for CNN	(ZHANG et al., 2018)		✓	✓			✓			✓
Zhang Decision Tree for CNN	(ZHANG et al., 2019)		✓	✓			✓			✓
Activation Atlas	(CARTER et al., 2019)		✓	✓			✓	✓	✓	
Zhang interpretable CNN	(ZHANG; WU; ZHU, 2018)		✓		✓	✓				✓
Feedforward interpretable CNN	(KUO et al., 2019)		✓		✓	✓				✓
ProtoAttend	(ARIK; PFISTER, 2020)		✓		✓	✓				✓

3.2.1 Atribuição de Características Baseada em Gradientes

As técnicas de atribuição baseadas em gradientes calculam um mapa de saliência dada uma imagem de teste e a classe correspondente. A técnica se baseia no cálculo do gradiente da saída da rede em relação a cada *pixel*, conectando os *pixels* com o *score* de classificação através da rede. Mapas de saliência baseados em gradiente (SIMONYAN; VEDALDI; ZISSERMAN, 2014) utilizam retropropagação para proliferar o sinal de importância até a entrada e construir uma imagem como um mapa de calor, chamado mapa

de saliência, que expressa topograficamente a importância de cada *pixel*. Uma interpretação para a utilização da derivada do *score* de classificação é a de que a magnitude da derivada indica os *pixels* que, mediante mínima alteração, mais afetem o *score* da classe (SIMONYAN; VEDALDI; ZISSERMAN, 2014).

A maneira de como lidar com a retropropagação dos gradientes até os *pixels* da entrada foi evoluída por técnicas posteriores de atribuição, como GuidedBackProp (SPRINGENBERG et al., 2014), Integrated Gradients (SUNDARARAJAN; TALY; YAN, 2017), DeepLIFT (SHRIKUMAR; GREENSIDE; KUNDAJE, 2017) e SmoothGrad (SMILKOV et al., 2017). Essas técnicas almejam maior robustez ao conservar melhor os gradientes que percorrem as funções não lineares que compõem os neurônios intermediários da rede.

3.2.2 Mapas de Ativação de Classes (*Class Activation Maps* ou CAM)

A técnica CAM (ZHOU et al., 2016) (*Class Activation Maps*), produz mapas de calor para regiões discriminativas para a classificação. A CAM tira proveito da conservação espacial dos filtros convolucionais em uma CNN, ou seja, a localização dos pesos de um filtro convolucional está associada a uma região com posicionamento correspondente na imagem de entrada. A técnica CAM utiliza os filtros convolucionais mais próximos da última camada da rede para construir mapas de calor sobre a área que explica uma determinada classificação, produzindo assim mapas de calor que discriminam bem a classe. No artigo, os autores utilizaram as regiões identificadas como importantes para realizar localização além da classificação, tendo, assim, como subproduto do treinamento do classificador um localizador de objetos treinado de maneira fracamente supervisionada, ou seja, sem a necessidade das delimitações precisas dos objetos de interesse.

A técnica Grad-CAM (SELVARAJU et al., 2017) pode ser vista como uma abordagem mais geral da técnica CAM, aplicável a uma variedade maior de arquiteturas. A técnica CAM exige a substituição de camadas densamente conectadas por camadas GAP (*Global Average Pooling*), já a técnica Grad-CAM evita o retreino que ocorre quando se substitui camadas além de evitar uma eventual perda de acurácia em favor da interpretabilidade. Embora as visualizações produzidas por Grad-CAM sejam discriminativas para a classificação, elas não mostram a atribuição de importância das regiões de maneira refinada como nos métodos baseados em gradiente, por se utilizar de filtros com baixa resolução. As técnicas Grad-CAM e Guided Backpropagation podem ser combinadas, no que se denomina Guided Grad-CAM, para obter mapas de calor mais discriminativos e detalhados. A técnica Grad-CAM++ (CHATTOPADHAY et al., 2018) aprimora a técnica Grad-CAM ao refinar as visualizações produzidas e estender a técnica para o domínio de vídeos digitais.

3.2.3 Atribuição de Características Baseada em Perturbação

Dada uma imagem e um classificador, a técnica LIME (RIBEIRO; SINGH; GUESTRIN, 2016) segmenta a imagem em diversas regiões e otimiza um modelo neural para aprender quais regiões da imagem influenciam positivamente ou negativamente a classificação final. A técnica LIME foi expandida em diversos outros trabalhos, como nas técnicas *Sound*-LIME (SLIME) (MISHRA; STURM; DIXON, 2017), *Quadratic* LIME (QLIME) (BRAMHALL et al., 2020), *Modified Pertubated Sampling* LIME (MPS-LIME) (SHI; ZHANG; FAN, 2020) e NormLime (AHERN et al., 2019).

RISE (PETSIUK; DAS; SAENKO, 2018) é outra técnica baseada em perturbação que utiliza máscaras aleatórias aplicadas às imagens de entrada para então verificar a saída produzida. De maneira semelhante à técnica LIME, RISE não utiliza os pesos internos da rede, considera apenas as máscaras utilizadas e a perturbação gerada pelas máscaras na saída do modelo. O artigo ainda propõe duas avaliações automáticas para técnicas de atribuição, chamadas *inserção* e *deleção*. A intuição por trás da avaliação deleção é que a remoção de pixels importantes forçará o modelo a mudar sua decisão. Especificamente, essa avaliação mede uma diminuição na probabilidade da classe correta ser predita à medida que mais e mais *pixels* importantes sejam removidos, onde a importância de cada pixel é obtida a partir do mapa de atribuição. A avaliação de adição funciona de maneira complementar.

3.2.4 SHAP

Lundberg e Lee (2017) propuseram SHAP *(SHapley Additive exPlanations)* como medida unificada para comunicar a contribuição de características para uma predição de um modelo de aprendizado de máquina. Os valores SHAP utilizam *Shapley values* (ROTH, 1988), uma solução utilizada na teoria dos jogos para aferir a contribuição de cada jogador em um jogo cooperativo.

Os valores SHAP podem ser utilizados como medida para outras técnicas de atribuição de características como LIME e DeepLift, gerando as variações *Kernel SHAP (Linear LIME combinado com Shapley values)* (AAS; JULLUM; LØLAND, 2021) e *Deep SHAP (DeepLIFT combinado com Shapley values)* (GARCÍA; AZNARTE, 2020). Em algoritmos de aprendizado de máquina recentes, como CatBoost (PROKHORENKOVA et al., 2018), SHAP é disponibilizado de maneira nativa para informar a contribuição de cada característica da entrada.

3.2.5 Funções de Influência

A técnica de *Influence Functions* (KOH; LIANG, 2017) trata um modelo de aprendizado de máquina como produto dos exemplos de treinamento correspondentes e, em

seguida, calcula a importância de cada exemplo de treinamento para uma previsão individual. Resgatar exemplos importantes para uma determinada previsão é uma explicação intuitiva que difere dos mapas de calor e visualizações.

É possível encontrar com exatidão a influência de um determinado exemplo de treinamento ao retreinar o modelo com todo o conjunto de treino menos com o exemplo em questão. É possível encontrar o exemplo mais influente para uma determinada predição ao se testar qual o exemplo que, se removido do conjunto de treinamento, mais impacta na predição do modelo para determinada imagem de teste. Koh e Liang (2017) realizaram uma validação do tipo *leave-one-out cross validation*, em que a comparação entre a técnica proposta e a técnica de retreino produziu um bom alinhamento entre as técnicas.

Além da validação da acurácia da técnica o artigo também valida as otimizações e cálculos desenvolvidos para que o cálculo das derivadas de segunda ordem, requeridas para o algoritmo, seja possível dentro de um intervalo de tempo factível.

3.2.6 Conceitos como Explicação

Uma linha de pesquisa é a utilização de conceitos visuais chave para expressar de maneira mais abstrata predições de redes convolucionais. Um exemplo de explicação que utiliza conceitos visuais seria o de um modelo treinado para classificar zebras que recorre aos seguintes conceitos visuais em ordem decrescente de importância: listras pretas e brancas, forma de cavalo e savana ao fundo.

Esse formato de explicação foi introduzido por Kim et al. (2018a), na pesquisa que apresentou a técnica TCAV (*Testing with Concept Activation Vectors*). Com TCAV é possível testar a importância de um conceito para a classificação de determinada classe. São necessários um conjunto de imagens pertencentes a uma classe e um conjunto de imagens que contenham apenas um conceito pertinente. Do conjunto de conceitos é extraído um CAV (*Concept Activation Vector* ou Vetor de Ativação de Conceito), TCAV utiliza o CAV e derivadas direcionais para avaliar a sensibilidade do modelo a perturbações nas ativações internas do modelo, perturbações essas realizadas de maneira alinhada com o vetor CAV, indicando assim a importância do conceito selecionado para os exemplos pertencentes a uma dada classe.

O TCAV é uma ferramenta interessante para avaliar viés em redes convolucionais. É possível ver qual a importância do conceito "ser homem" para a classificação de um médico, ou a importância do conceito "presença de pessoas asiáticas" para a classificação de uma cena como sendo tênis de mesa. Decidir quais conceitos serão testados é visto como uma desvantagem da técnica ante outras que expandiram essa linha de pesquisa e realizam a busca de conceitos importantes de forma automática, como ACE (*Automatic Concept Based Explanations*) (GHORBANI; WEXLER; KIM, 2019). Outros métodos expandiram

o TCAV, como o CaCE (GOYAL; SHALIT; KIM, 2019), que busca o efeito causal da presença do conceito e o ConceptSHAP (YEH et al., 2019), que utiliza valores SHAP para atribuir importância para os conceitos.

3.2.7 Destrinchar Representações Internas

Existem esforços para desembaraçar as representações internas de uma rede neural e organizar os padrões internos identificados em visualizações como grafos explanatórios interpretáveis ou árvores de decisão. Para os autores Zhang et al. (2016), destrinchar ou desembaraçar as representações de características internas de uma rede neural convolucional, ou treinar desde o princípio redes com representações internas compreensíveis, são grandes desafios para os algoritmos de interpretabilidade que formam o estado da arte.

Um mesmo filtro convolucional pode ser ativado por diversos padrões contidos na mesma imagem. No artigo *Network Dissection* (BAU et al., 2017) o autor conduz um experimento para quantificar a interpretabilidade das representações internas de uma rede ao contar a quantidade de unidades internas alinhadas com um padrão único e interpretável, como cor, textura ou formato de objeto, quantificando assim o embaraçamento das representações internas das redes testadas. O artigo analisa o efeito de fatores de treinamento, como quantidade de épocas, profundidade e largura de uma rede convolucional na interpretabilidade das representações internas de uma rede convolucional ao aferir a quantidade de filtros alinhados com esses padrões únicos.

Em (ZHANG et al., 2018) os autores propuseram uma abordagem para destrinchar a hierarquia de conceitos escondida nos filtros convolucionais de uma CNN pré-treinada, provendo uma visão global dos conceitos em uma estrutura de grafo. O grafo proposto funciona como uma representação grosseira e interativa do conhecimento interno da rede convolucional, onde cada nó, aresta e nível ajuda no entendimento da hierarquia dos padrões que existem em uma rede convolucional.

Em um segundo trabalho por Zhang et al. (2019), os autores desenvolveram uma técnica de explicação local que utiliza uma árvore de decisão para organizar de maneira visual e hierárquica as predições individuais realizadas por uma rede convolucional. A árvore de decisão indica quais representações foram utilizadas pela rede convolucional e qual a contribuição de cada uma para a predição.

Outro trabalho que permite a expansão do entendimento sobre redes convolucionais e suas representações internas é o *Activation Atlas* (CARTER et al., 2019). Nesse trabalho foram utilizados um milhão de recortes de imagens em que, para cada recorte, as ativações de todas as camadas da rede foram registradas, ou seja, os valores numéricos que quantificam a intensidade e o perfil da ativação de cada camada com respeito a cada recorte. O *Activation Atlas* se destaca por conseguir exibir de maneira inteligível e inte-

rativa um milhão de ativações, valendo-se de uma combinação de técnicas de redução de dimensionalidade, visualização e atribuição de características.

As técnicas apresentadas nesta seção ajudam a expandir e consolidar o entendimento científico sobre as representações internas escondidas nos parâmetros de um modelo neural convolucional profundo, utilizadas por pesquisadores e entusiastas para entender a fundo as redes neurais convolucionais.

3.2.8 Modelos Explicáveis

A maioria das técnicas de interpretabilidade atua de maneira pós treinamento, no entanto, existem alguns trabalhos que propõem mudanças no algoritmo tradicional de treinamento para que os modelos treinados sejam mais facilmente interpretáveis.

Zhang, Wu e Zhu (2018) propuseram um método que transforma uma CNN tradicional em uma CNN mais interpretável ao forçar representações internas mais claras e concisas. Para tal, o autor utiliza funções de perda em cada filtro das camadas convolucionais.

Os parâmetros de redes neurais profundas são normalmente alterados por um processo de retropropagação (*backpropagation*). Na pesquisa de (KUO et al., 2019), é proposto um método de treinamento que não utiliza retropropagação para determinar os parâmetros aprendidos. A rede proposta segue um *design feedforward* interpretável sem a utilização de nenhuma retropropagação como referência. O *design* proposto adota uma abordagem centrada em dados em que os parâmetros de cada camada são determinados a partir do uso de estatísticas sobre a saída da camada anterior. O *design* sem retropropagação se mostrou superior ao *design* tradicional na robustez contra ataques adversariais, sendo também mais transparente por ser fácil acompanhar a determinação dos pesos ao longo do processo de treinamento.

Arık e Pfister (2020) apresentam um *design* de modelos convolucionais que utilizam protótipos denominados *ProtoAttend*, nos quais os modelos armazenam características dos dados de treinamento através do treinamento de protótipos para que na fase de uso as entradas sejam comparadas contra os protótipos aprendidos. A explicação fornecida com a predição é intuitiva por utilizar uma lógica simples, “isso se parece com aquilo” (*This looks like that*), ou seja, a explicação para a classificação é a semelhança entre a imagem sendo testada e o que já foi aprendido. Apesar de obterem boa acurácia e proverem explicações locais, esses modelos necessitam de grande poder de armazenamento e processamento nas fases de treinamento e uso.

3.3 Avaliação de Técnicas de Atribuição

O campo da interpretabilidade, por ser recente, possui muitos debates científicos em aberto. Muitos artigos alimentam o debate técnico ao apontar falhas em algumas abordagens. Os artigos (KINDERMANS et al., 2019), (ADEBAYO et al., 2018), (GHORBANI; ABID; ZOU, 2019) e (GHORBANI; WEXLER; KIM, 2019), apontam falhas fundamentais relacionadas à robustez de técnicas que utilizam mapas de atribuição como saída para a interpretabilidade. Por exemplo, espera-se que uma técnica retorne explicações semelhantes para imagens de entrada semelhantes, entretanto, o oposto ocorre para um conjunto de técnicas avaliadas em um experimento conduzido por Ghorbani, Abid e Zou (2019).

Um ponto de evolução perceptível nas pesquisas dessa área é a questão da validação das técnicas. Muitas das publicações iniciais não validam a utilidade das explicações fornecidas e utilizam apenas exemplos ilustrativos para tentar convencer o leitor sobre a utilidade das explicações. Publicações mais recentes tendem a apresentar experimentos mais robustos para validar as técnicas de interpretabilidade propostas, contrastando a técnica em questão com técnicas anteriores. As publicações que propõem as técnicas Grad-Cam (SELVARAJU et al., 2017) e ACE (GHORBANI; WEXLER; KIM, 2019) apresentam validações quantitativas e qualitativas que vão além de exemplos ilustrativos.

A tarefa de validar e comparar técnicas de interpretabilidade é desafiadora por não existir um gabarito definitivo sobre o que é uma explicação ideal. A abordagem qualitativa é utilizar questionários para confrontar as explicações com a expectativa de usuários ou especialistas. No entanto, avaliações qualitativas são difíceis de uniformizar e escalar devido a eventuais questões de custos (e.g. tempo de especialistas) e à forte dependência de avaliadores humanos. Avaliações quantitativas são mais utilizadas para validações e comparações (LI et al., 2020).

A partir dos artigos revisados neste capítulo foi possível observar que técnicas de atribuição geralmente produzem um cenário com três atores: a imagem sendo explicada, o modelo treinado e a explicação resultante. Muitas das avaliações quantitativas acompanham causa e efeito ao se alterar algum dos atores no cenário relatado e então registrar o efeito nos demais. Por exemplo, uma das avaliações propostas por Adebayo et al. (2018) altera o modelo e registra o efeito nas explicações, representando uma tentativa de capturar a sensibilidade das explicações resultantes mediante alterações no modelo. Dessa forma podemos avaliar propriedades, como sensibilidade ao modelo, que indicam uma maior qualidade das técnicas.

Em Li et al. (2020) as avaliações quantitativas utilizadas para avaliar técnicas de atribuição são categorizadas conforme a dimensão de avaliação: fidelidade, capacidade de localização, formação de falsos positivos, sensibilidade e estabilidade. As seções a seguir

apresentam uma proposta de categorização para as avaliações quantitativas baseada no princípio de funcionamento.

3.3.1 Sondagem das Saídas do Modelo

Existe um conjunto de avaliações que monitoram as predições produzidas pelo modelo conforme modificações nas entradas do modelo são realizadas. As modificações nas entradas seguem o mapa de atribuição produzido pelas técnicas, que indica quais as regiões devem ser importantes para a predição. A Figura 6 ilustra o processo de sondagem das saídas conforme as entradas são perturbadas.

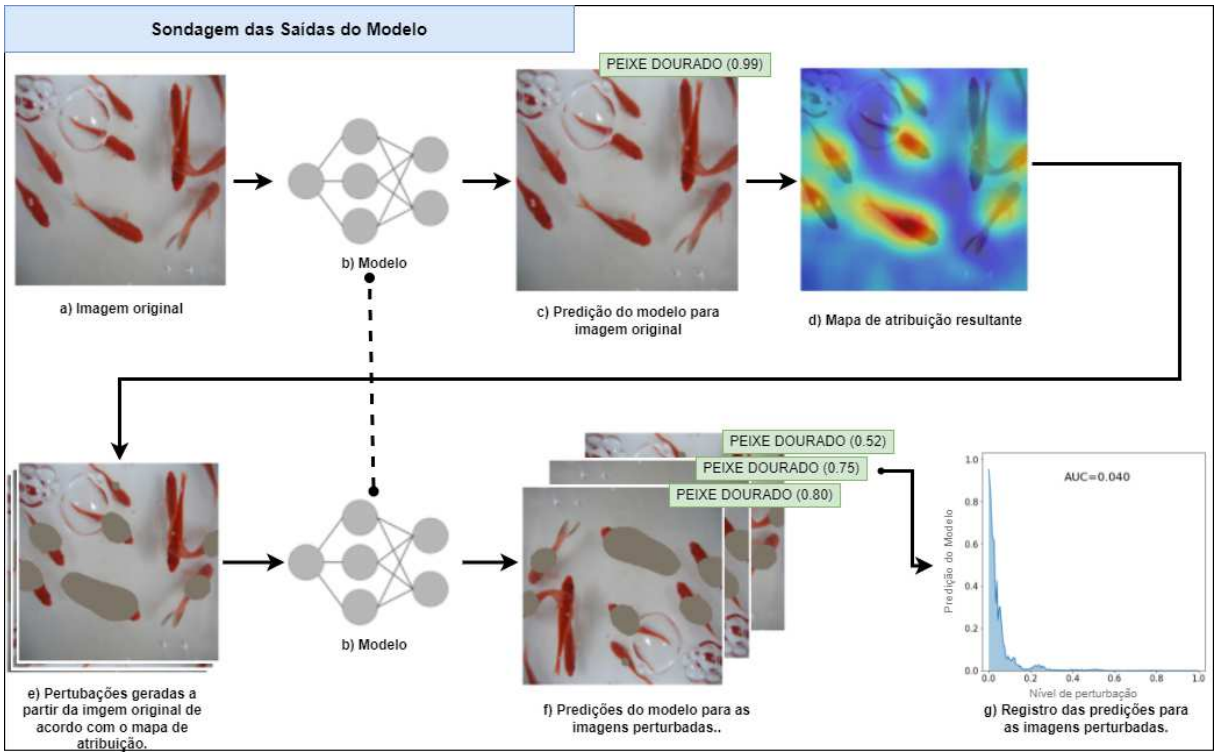


Figura 6 – Processo para avaliar técnica de interpretabilidade ao sondar as saídas do modelo. Para uma determinada imagem (a) o modelo a ser explicado (b) realiza uma predição (c) e a técnica de atribuição produz um mapa de atribuição (d). A partir do mapa de atribuição as pertubações são geradas (e), conforme a ordem de importância do mapa. As novas imagens são então passadas para o modelo, que realiza novas predições (f), que podem ser visualizadas em formato de gráfico (g). A área abaixo da curva (*Area Under The Curve* ou AUC) pode ser utilizada para sumarizar o resultado da técnica avaliada e comparar diferentes técnicas. Fonte: Elaboração própria do autor.

3.3.1.1 Área Sob a Curva

No artigo (PETSIUK; DAS; SAENKO, 2018) duas avaliações complementares são propostas: deleção (*deletion*) e inserção (*insertion*). A deleção segue o processo ilustrado na Figura 6: a partir dos mapas de atribuição os *pixels* com maior importância são sucessivamente substituídos por ruído, prejudicando a confiança do modelo naquela predição.

Uma queda mais abrupta na probabilidade de predição resulta em uma menor área sob a curva, o que é interpretado como indicativo de uma boa técnica de explicação. As linhas de queda de probabilidade são sumarizadas e comparadas utilizando área abaixo da curva (*Area Under The Curve* ou AUC). A avaliação de inserção é medida pelo aumento da probabilidade da classe de interesse conforme os *pixels* são adicionados seguindo a importância mapeada pelo método de interpretabilidade. A imagem original é o ponto de partida para a avaliação de deleção e o ponto de chegada para a avaliação de adição. A Figura 7 ilustra as avaliações deleção e inserção.

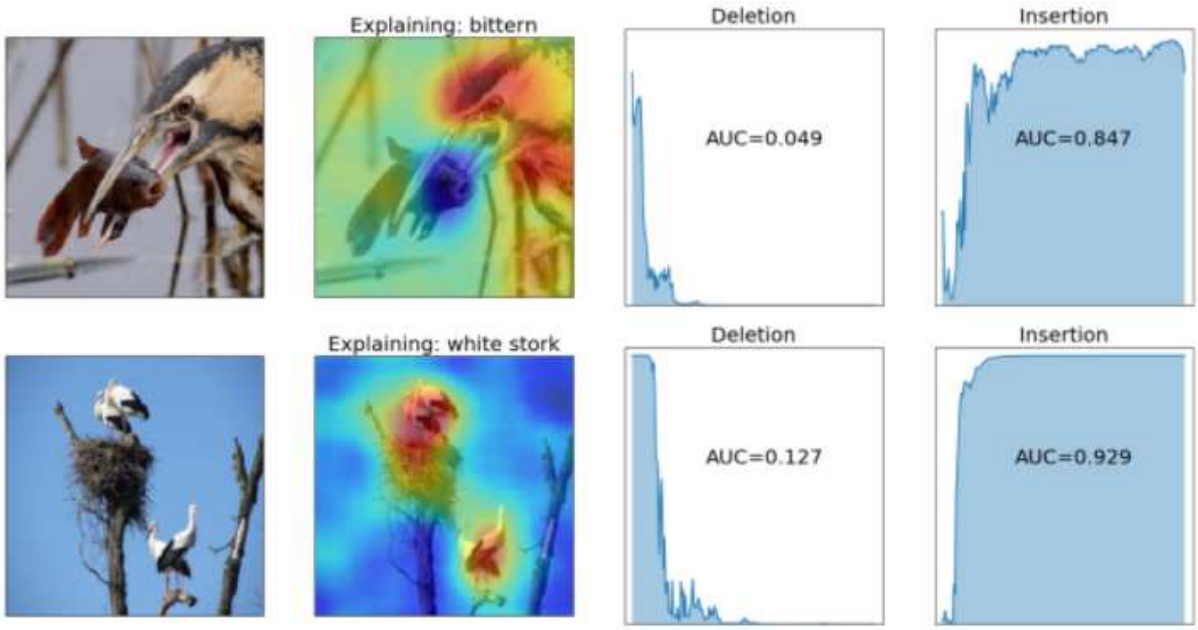


Figura 7 – Avaliações inserção e deleção, como apresentadas por Petsiuk, Das e Saenko (2018). Exemplo de mapa de atribuição conforme a técnica RISE (segunda coluna) para imagens representativas (primeira coluna). A avaliação deleção está ilustrada na terceira coluna. A avaliação inserção está ilustrada na quarta coluna. Fonte: (PETSIUK; DAS; SAENKO, 2018)

No artigo (GOH et al., 2021), os autores realizaram um procedimento diferente para computar a avaliação deleção, na qual regiões quadradas não sobrepostas são ordenadas conforme a importância mapeada pelo método de atribuição sendo avaliado e, então, essas regiões são perturbadas em etapas. A cada etapa, a região mais importante é perturbada de diferentes maneiras e a probabilidade média da classe de interesse é registrada. Dentre as pertubações realizadas, aquela que causou uma mudança na probabilidade da classe de interesse mais próxima da mediana é a escolhida para iniciar o próximo passo de perturbação. Da mesma forma que na pesquisa discutida no parágrafo anterior, as técnicas também são comparadas utilizando AUC.

As avaliações apresentadas se valem da inserção de ruído e artefatos nas imagens originais, o que pode causar um problema conhecido em redes neurais profundas, o problema de entradas fora da distribuição (*out-of-distribution data* ou OoD) — para entradas

inconsistentes com as entradas vistas em treinamento, em que a rede tende a predizer uma classe aleatória com alta confiança, assim como em exemplos adversariais. O procedimento alternativo para calcular a avaliação de deleção realizado por Goh et al. (2021) mitiga o problema OoD, a depender do modelo.

3.3.2 Comparação Entre Explicações

Para um segundo conjunto de avaliações quantitativas, as alterações nos mapas de atribuição são monitoradas e as modificações que causam essas alterações são realizadas no modelo ou nas entradas. Essas avaliações servem como testes de sanidade. Para modificações grosseiras é esperado que as explicações sejam alteradas. Já para modificações discretas é esperado que as explicações se alterem ligeiramente.

3.3.2.1 Avaliações que Alteram o Modelo

No artigo (ADEBAYO et al., 2018) o autor propõe dois testes de sanidade: randomização das camadas e randomização dos dados. É esperado que modelos diferentes tenham explicações diferentes para a mesma imagem, no entanto, não é o que acontece com muitas explicações, que permanecem idênticas mesmo com a randomização de parte dos modelos. Ainda no mesmo artigo o autor compara as explicações produzidas por dois modelos, treinados a partir do mesmo conjunto de dados: um modelo é treinado com rótulos corretos e o outro modelo é treinado com os rótulos randomizados. Para os dois modelos treinados as explicações produzidas são similares para diversas técnicas de atribuição, ou seja, para algumas técnicas as explicações são similares independente do modelo ser despropositado ou coerente, o que é um comportamento não desejável.

Em (ARUN et al., 2021), o autor busca a propriedade de repetibilidade das explicações. O autor quantifica a repetibilidade dos mapas de saliência comparando as explicações produzidas por modelos de mesma arquitetura, mas provenientes de diferentes treinamentos. O autor também quantifica a reprodutibilidade das explicações geradas por modelos de diferentes arquiteturas.

3.3.2.2 Avaliações que Alteram as Entradas

Imagens idênticas produzindo explicações destoantes podem causar problemas em aplicações. Nos artigos (GHORBANI; ABID; ZOU, 2019; DOMBROWSKI et al., 2019; KINDERMANS et al., 2019) os autores provocam mudanças injustificadas nas explicações a partir de pertubações imperceptíveis nas entradas. Ataques que afetam as técnicas de interpretabilidade assim como defesas para esses ataques (RIEGER; HANSEN, 2020) estão presentes na literatura de maneira similar aos ataques adversariais, que alteram as predições dos modelos com modificações mínimas. As variações nas explicações podem ser medidas, quantificando-se a robustez das técnicas de interpretabilidade. O resultado

de alguns testes estão apresentados na Figura 8, em que exemplos modificados de maneira imperceptível produziram explicações destoantes de acordo com três técnicas diferentes.

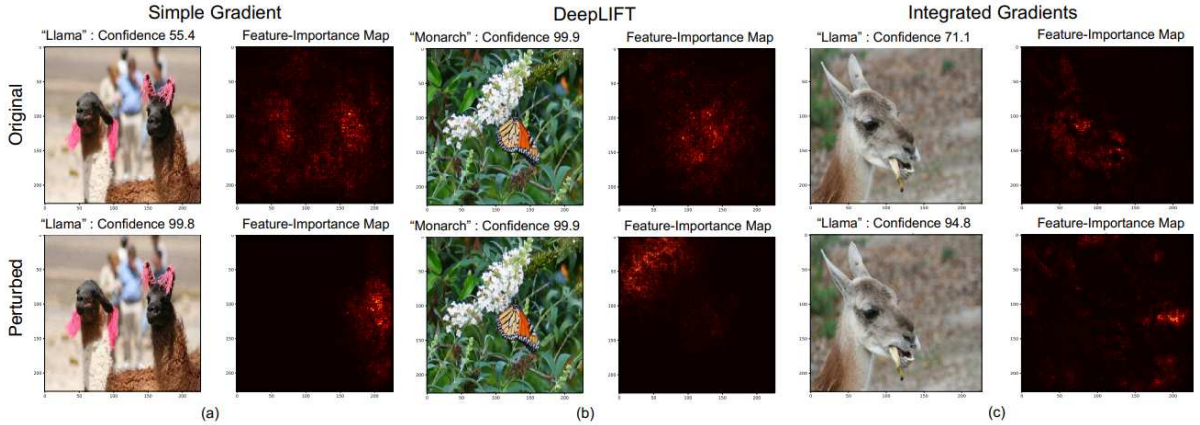


Figura 8 – Exemplos nos quais as entradas foram perturbadas de maneira imperceptível, no entanto, as pertubações geraram modificações notáveis nas explicações. Por exemplo, para a técnica *Simple Gradient*) (a), uma imagem de uma lhama foi perturbada de forma imperceptível (primeira coluna), no entanto, as explicações geradas pela técnica (segunda coluna) foram drasticamente diferentes.As técnicas avaliadas foram: Gradiente Simples (*Simple Gradient*) (a) (SIMONYAN; VEDALDI; ZISSERMAN, 2014), DeepLIFT (b) (SHRIKUMAR; GREENSIDE; KUNDAJE, 2017) e Integrated Gradients (c) (SUNDARARAJAN; TALY; YAN, 2017). Fonte: (GHORBANI; ABID; ZOU, 2019)

3.3.2.3 Outras Comparações

Li et al. (2020) conduziram o seguinte experimento: duas explicações são produzidas para cada imagem do conjunto de teste, uma explicação para a classe de maior probabilidade, a classe predita, e uma explicação para a classe de menor probabilidade. É esperado que as duas explicações sejam diferentes, pois essas duas classes geralmente não têm estratégias de discriminação semelhantes. Para quantificar a semelhança é utilizada uma medida de similaridade entre imagens como correlação de Pearson. Nos experimentos realizados por Li et al. (2020), as técnicas *Grad-CAM*, *RISE* e *Oclussion* obtiveram excelentes resultados (menor similaridade entre explicações), enquanto as técnicas baseadas nos gradientes da rede neural, *Gradient* e *IntegratedGradients* produziram explicações semelhantes, independente da classe sendo explicada.

Goh et al. (2021) apresentam uma avaliação cujo objetivo é quantificar o ruído existente em um mapa de atribuição. Para tal, calcula-se o que foi chamado de variação média total (*Average Total Variation* ou ATV). O procedimento envolve aferir as variações de intensidade na vizinhança de cada *pixel* do mapa de atribuição. Os autores fazem o mesmo cálculo em diversas escalas diferentes, diminuindo a resolução a cada etapa até que o mapa seja encolhido para uma resolução menor que 30×30 *pixels*, como que subindo nos níveis de uma pirâmide (Pyramid Steps). Com os valores ATV para cada etapa de redução de resolução, pode-se desenhar um gráfico como na Figura 9, em que a área

abaixo da curva é utilizada como comparação entre técnicas. A avaliação é de simples implementação e quantifica a média da quantidade de ruído local nas explicações, sendo útil para comparar técnicas que seguem a mesma abordagem.

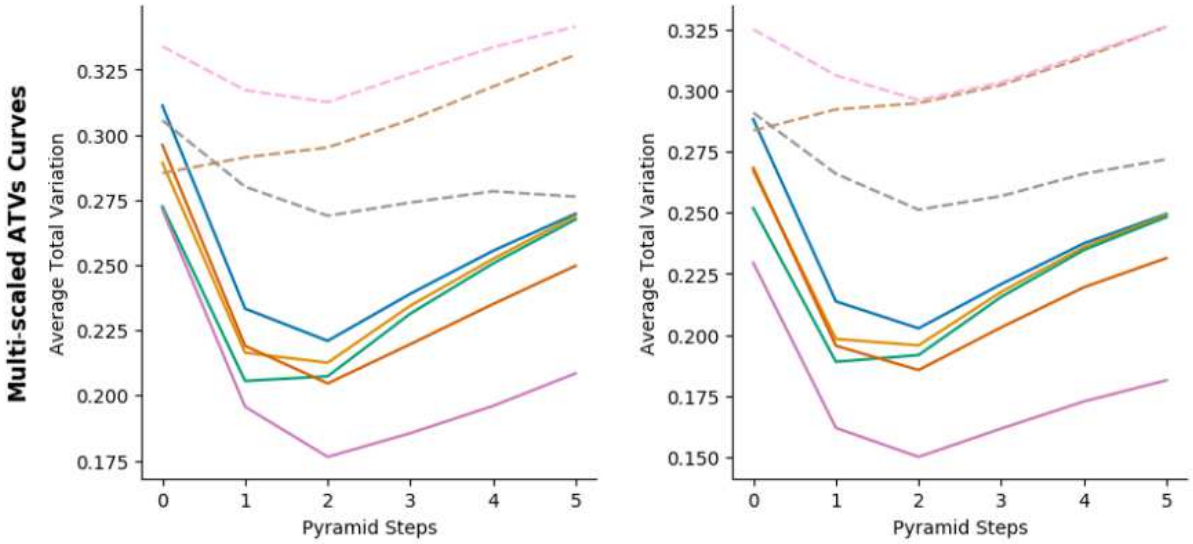


Figura 9 – Variação média total (*Average Total Variation* ou ATV). Curvas em várias escalas de resolução (*Pyramid Steps*). Os resultados comparam as arquiteturas DenseNet121 (esquerda) e ResNet152 (direita). Fonte: (GOH et al., 2021)

3.3.3 Utilização de Conjuntos de Dados Anotados

Outro conjunto de avaliações é dependente de um tipo específico de conjunto de dados, como os comumente utilizados na tarefa de segmentação que podem servir também para classificação. Avaliações que dependem de conjuntos específicos servem principalmente para comparar técnicas diferentes e não são comumente utilizadas para avaliar a aplicação de técnicas de interpretabilidade em cenários específicos, exceto se meta informações sobre as imagens também sejam fornecidas.

3.3.3.1 Localização e Classificação

Existem conjuntos de dados que podem ser utilizados para classificação, mas que também possuem anotações sobre a localização dos objetos de interesse contidos nas imagens, a exemplo dos conjuntos COCO (LIN et al., 2014) e Pascal VOC 2012 (EVERINGHAM et al., 2012). A meta informação das regiões segmentadas nas imagens possibilita o experimento proposto por Zhang et al. (2016) e utilizado por Selvaraju et al. (2017), por exemplo, como validação para a técnica GradCam. No experimento é extraído o ponto de máxima importância no mapa de calor gerado. Em seguida, é avaliado se o ponto está na região segmentada anotada para categoria do objeto alvo da classificação, contando assim um acerto ou erro. Em outras palavras, a capacidade de localização de

uma técnica de interpretabilidade pode ser vista como um indicativo da qualidade dessa técnica.

Na Figura 10 estão alguns exemplos da utilização da técnica GradCam para a tarefa de segmentação de imagens do conjunto de dados Pascal VOC 2012 (EVERINGHAM et al., 2012). Os exemplos apresentados utilizam as explicações da técnica GradCam como semente para uma abordagem intitulada SEC (*Seed, Expand and Constrain* ou Semear, Expandir e Restringir) (KOLESNIKOV; LAMPERT, 2016).

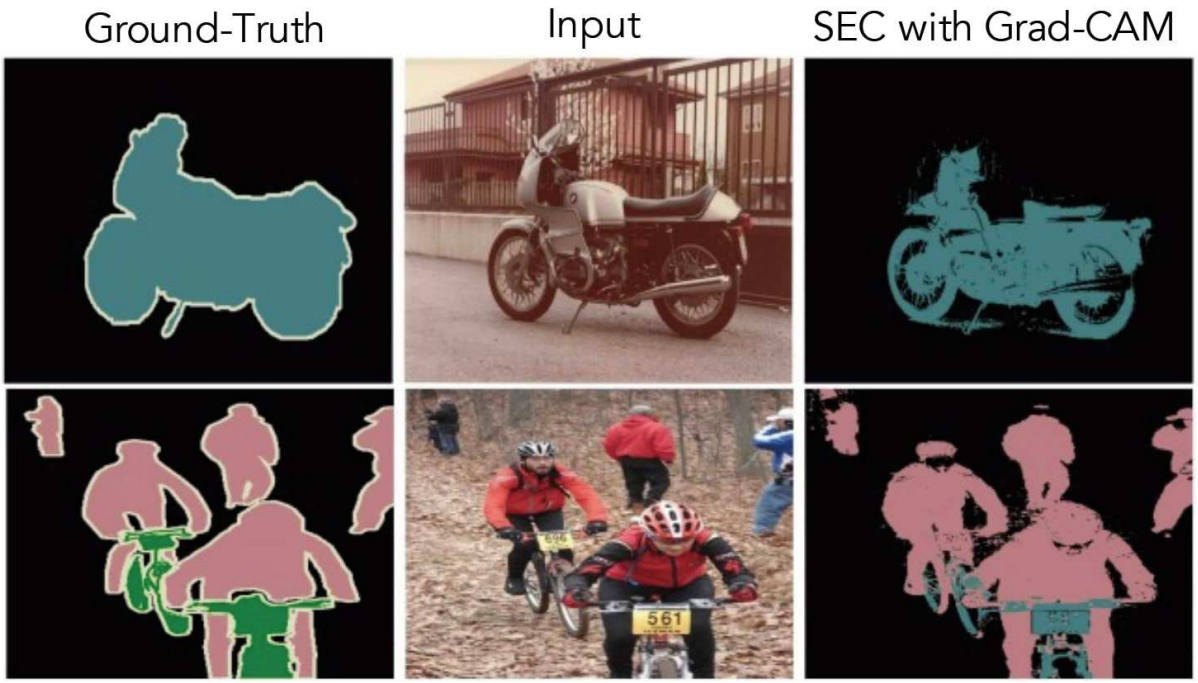


Figura 10 – Resultados qualitativos da utilização das explicações da técnica GradCam como semente para a tarefa de segmentação de imagens do conjunto de dados Pascal VOC 2012 (EVERINGHAM et al., 2012). Fonte: (SELVARAJU et al., 2017)

3.3.3.2 BAM

Yang e Kim (2019) produziram um conjunto de dados específico para comparação de métodos de atribuição (*Benchmarking Attribution Methods* ou BAM). O conjunto de dados é o produto de duas outras bases de imagens rotuladas: uma base composta por cenas e uma base composta por objetos segmentados. No conjunto BAM os objetos segmentados são inseridos sobre as cenas, permitindo o treinamento de dois modelos a partir das mesmas imagens: um classificador de cenas e um classificador de objetos. Um mesmo objeto é inserido em diversos planos de fundo diferentes, portanto, é esperado que os planos de fundo sejam menos importantes para o classificador de objetos e mais importantes para o classificador de cenas. A síntese da base de dados a partir da união de outras duas bases possibilita uma avaliação de técnicas de atribuição mais controlada e amparada por meta informações, como localização e classe dos objetos e dos planos

de fundo. A Figura 11 ilustra a síntese do conjunto BAM, na qual estão apresentados exemplos de imagens sintetizadas e os dois modelos possíveis de serem treinados.

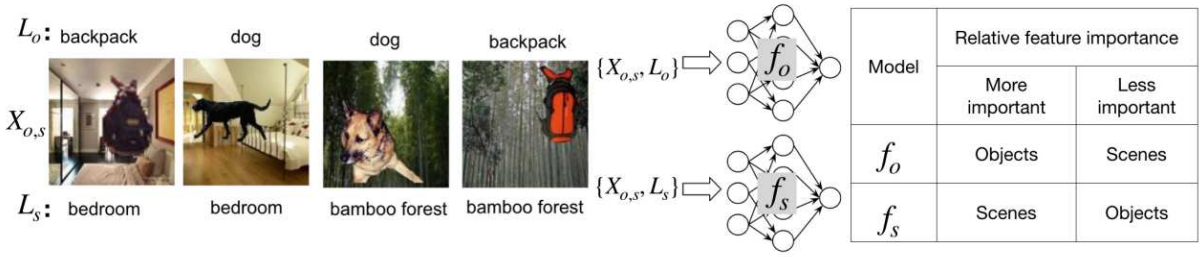


Figura 11 – *Benchmarking Attribution Methods* ou BAM. Duas bases de imagens são unidas para produzir um conjunto de dados específico para comparação de métodos de atribuição. Nesse conjunto dois modelos podem ser treinados, um específico para detectar a classe plano de fundo fs e um específico para detectar a classe do objeto inserido fo. Fonte: (YANG; KIM, 2019)

3.3.3.3 *Revel Cancel*

Para a validação da técnica Deep Lift, os autores Shrikumar, Greenside e Kundaje (2017) utilizaram o conjunto de imagens de dígitos manuscritos MNIST (DENG, 2012). A partir de uma imagem de um caractere, os autores utilizaram a técnica de atribuição de características para explicar outro caractere "escondido" no caractere original, por exemplo, a partir de uma imagem de um 8, a técnica tenta explicar um 3, revelando então o 3 que está normalmente contido no traçado do caractere 8, como que apagando algumas regiões para transformar um número em outro. A confiança do modelo em rotular o 3 revelado de dentro da imagem original é então quantificada. A Figura 12, retirada do artigo original, exemplifica como o 3 e o 6 são revelados a partir do traçado do 8 utilizando técnicas de atribuição.

3.4 Considerações Finais

Ao final desta revisão bibliográfica observou-se uma variedade de esforços para explicar ou expor o funcionamento dos modelos utilizados em tarefas visuais. As técnicas de atribuição explicam predições individuais atribuindo responsabilidade a regiões da entrada e podem utilizar ou não os parâmetros internos dos modelos. As técnicas de visualização produzem imagens para as quais determinadas regiões dos modelos respondem fortemente, na tentativa de explicar o funcionamento interno do modelo ao expor a organização interna, construída via treinamento, para fins de extração de características e identificação de padrões.

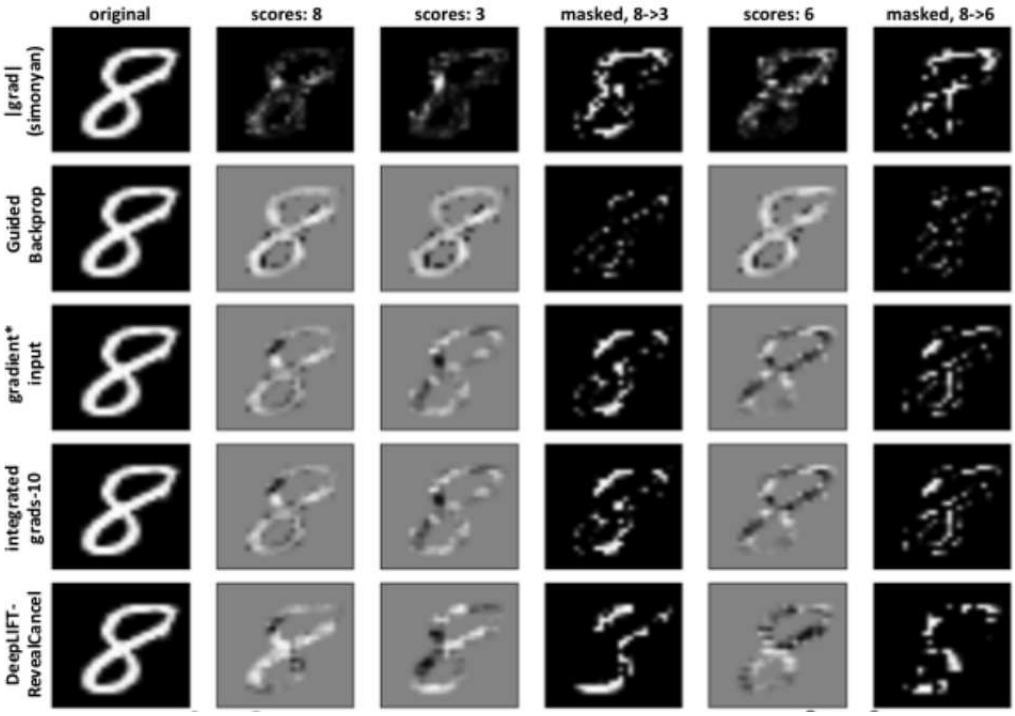


Figura 12 – Avaliação *Revel Cancel*. A partir de uma imagem de um dígito manuscrito 8, diferentes técnicas revelam o caractere escondido no caractere original. Na quarta e sexta colunas os caracteres alvos são o 3 e o 6, respectivamente. Fonte: (SHRIKUMAR; GREENSIDE; KUNDAJE, 2017)

Técnicas que não se encaixam em atribuição ou visualização de características comunicam as suas explicações de outras maneiras, tais como, apresentando exemplos de treinamento influentes ou elicitando conceitos importantes para determinadas classes.

Entre as pesquisas sobre atribuição de características existe uma clara tendência de aprimorar a validação e a comparação de técnicas. Tal tendência é consequência de pesquisas que criticam técnicas de atribuição (KINDERMANS et al., 2019), (ADEBAYO et al., 2018), (GHORBANI; ABID; ZOU, 2019) e (GHORBANI; WEXLER; KIM, 2019) e de técnicas recentes que comparam os resultados com técnicas prévias (SELVARAJU et al., 2017; SHRIKUMAR; GREENSIDE; KUNDAJE, 2017). Nesta revisão, as avaliações existentes foram organizadas nas seguintes categorias: Sondagem das Saídas do Modelo, Comparação Entre Explicações e Utilização de Conjuntos de Dados Anotados.

4 Materiais e Métodos

Um estudo de caso foi conduzido objetivando integrar os conhecimentos levantados sobre técnicas de interpretabilidade de redes convolucionais e entender como avaliar essas técnicas. Foram realizadas, em sequência, as etapas de treinamento, interpretabilidade e avaliação dos resultados, integrando os conhecimentos adquiridos na revisão bibliográfica, no Capítulo 3.

Para o estudo de caso foram escolhidos conjuntos de dados de diagnóstico médico por imagem, por ser um domínio de aplicação de extrema relevância por lidar com a saúde humana, além de ser diferente daquilo que é habitualmente utilizado em avaliações de técnicas de interpretabilidade. Os conjuntos de imagens normalmente utilizados em comparações de técnicas de interpretabilidade são provenientes de desafios de classificação de imagens e possuem diversas categorias, além de uma maior variedade entre exemplos, como nos conjuntos *ImageNet* (DENG et al., 2009), COCO (LIN et al., 2014) e Pascal (EVERINGHAM et al., 2012), que possuem 1000, 80 e 20 categorias respectivamente. Diferentemente dos conjuntos habituais, os conjuntos de dados constituídos por imagens médicas seguem um forte padrão, definido pela categoria do exame em questão, com baixa quantidade de classes e pouca variedade entre exemplos se comparado com os conjuntos anteriormente citados.

As técnicas de interpretabilidade avaliadas nesta pesquisa seguem a abordagem de atribuição, ou seja, explicam predições individuais de um modelo ao distribuir a responsabilidade da predição entre as características da imagem (por exemplo, *pixels* individuais ou regiões segmentadas), detalhadas com maior profundidade no Capítulo 3. As técnicas de atribuição são especialmente interessantes para aplicação de diagnóstico assistido por computador, por possibilitarem apoio à decisão ao explicar predições individuais. A região que importa para a predição pode ser utilizada para auxiliar um médico em um diagnóstico e ajudar o profissional a entender os porquês de uma possível divergência entre interpretações.

O presente capítulo apresenta os materiais e métodos utilizados na condução do estudo de caso. Nas Seções 4.1, 4.2 e 4.3, estão apresentados, respectivamente, os conjuntos de dados, as arquiteturas neurais e as técnicas de interpretabilidade utilizadas no estudo. Na Seção 4.4 estão descritas as avaliações utilizadas para analisar as diferentes técnicas de explicabilidade. Por fim, alguns detalhes de implementação estão descritos na Seção 4.5.

4.1 Conjuntos de Dados

Os conjuntos de dados rotulados utilizados foram disponibilizados por Kermany, Zhang e Goldbaum (2018) e explorados na pesquisa de Kermany et al. (2018). Ambos os conjuntos estão restritos à escala de cinza e estão detalhados a seguir.

4.1.1 Raio-X Torácico

Um dos conjuntos de dados escolhidos é constituído por imagens de raio-X torácico. O conjunto é composto por radiografias da região do tórax de pacientes acometidos por pneumonia viral ou bacteriana, além de imagens provenientes de pacientes sãos. A resolução das imagens varia entre 200dpi e 400dpi. Para os treinamentos realizados, as imagens foram divididas em duas classes: normal e pneumonia (abrangendo pneumonia viral e bacteriana). O diagnóstico de pneumonia a partir de um exame de raio-x tem como base a detecção de padrões, como regiões opacas, algo que não foge do escopo de possibilidades de uma rede neural convolucional treinada. Exemplos de diagnóstico a partir de algumas características presentes nas imagens estão apresentados na Figura 13.

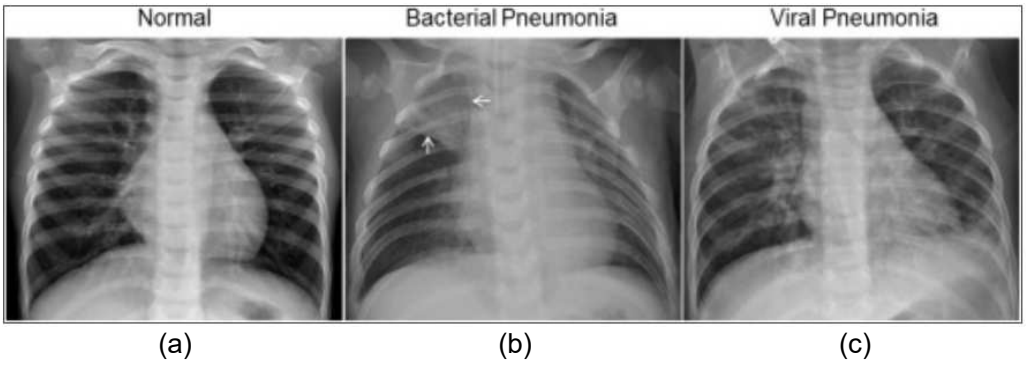


Figura 13 – (a) Radiografia de tórax de uma pessoa sadia mostra pulmões limpos sem nenhuma área opaca anormal na imagem. (b) A pneumonia bacteriana tipicamente exibe uma consolidação focal em um lobo pulmonar, neste caso, no lobo superior direito (setas brancas). (c) A pneumonia viral se manifesta com um padrão mais difuso em ambos os pulmões. Fonte: (KERMANY et al., 2018)

4.1.2 OCT

O segundo conjunto escolhido é constituído por imagens de tomografias de coerência óptica (*Optical Coherence Tomography* ou OCT). O exame OCT é utilizado para diagnosticar doenças na retina e conta com 109.312 imagens provenientes de 5.319 pacientes. Cada imagem pode ser rotulada como: neovascularização coroidal (*Choroidal Neovascularization* ou CNV), edema macular diabético (*Diabetic Macular Edema* ou DME), drusen ou normal. O conjunto de dados foi adquirido em cinco hospitais e centros oftalmológicos entre as datas de 1 ° de julho de 2013 e 1 ° de março de 2017. O formato e a presença

de anomalias são características discriminantes para o diagnóstico de doenças na retina a partir de um exame de OCT, como apresentado na Figura 14, algo que é passível de solução por intermédio do treinamento de uma rede neural convolucional treinada.

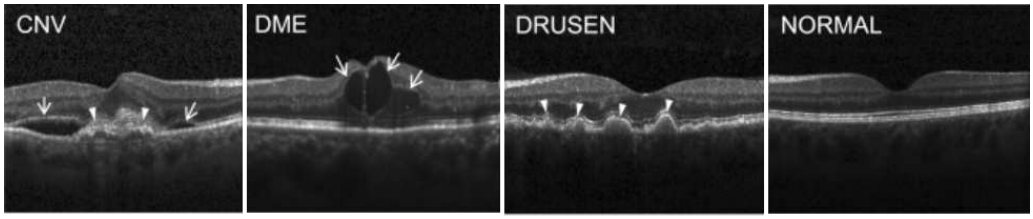


Figura 14 – (a) Neovascularização coroidal apresenta membrana neovascular (setas ao centro) e fluido sub-retiniano associado (setas laterais). (b) Edema macular diabético apresenta líquido intrarretiniano associado ao espessamento da retina (setas). (c) Múltiplas drusas presentes em decorrência de uma degeneração macular relacionada à idade inicial. (d) Retina normal com contorno preservado e ausência de qualquer fluido/edema retiniano. Fonte: (KERMANY et al., 2018)

4.1.3 Configuração dos Conjuntos de Dados

Na Tabela 1, estão apresentadas as configurações dos conjuntos de dados utilizados. Os conjuntos de teste foram utilizados para avaliar os respectivos treinamentos e alguns dos exemplos de teste foram utilizados para a fase de interpretabilidade do estudo. As configurações dos conjuntos foram herdadas dos desafios de dados originais para facilitar comparações e reutilização de código de treinamento.

Tabela 1 – Configurações dos conjuntos de dados utilizados. Estão descritas as quantidades de exemplos para cada classe e segmentação das bases de dados Raio-X e OCT.

Configuração Conjunto Raio-X		
Classe	**Treino**	**Teste**
Normal	1349	234
Pneumonia	3883	390
Configuração Conjunto OCT		
Classe	**Treino**	**Teste**
CNV	37205	250
DME	11348	250
Drusen	8616	250
Normal	8616	250

4.2 Arquiteturas Neurais

Os modelos treinados seguiram duas arquiteturas convolucionais, uma arquitetura se chama VGG16 (SIMONYAN; ZISSERMAN, 2014). A outra arquitetura utilizada foi

uma arquitetura customizada, mais simples, mas que segue os mesmos princípios da arquitetura VGG16. A rede customizada teve como inspiração a arquitetura utilizada por Adebayo et al. (2018) em uma comparação entre técnicas de interpretabilidade de maneira similar ao estudo aqui proposto. Apenas mudanças relacionadas com as dimensões das imagens de entrada foram realizadas em comparação com a rede treinada por Adebayo et al. (2018). Os diagramas das arquiteturas utilizadas estão presentes no Apêndice A.

- **VGG16** — VGG16 é uma arquitetura convolucional com cerca de 138 milhões de parâmetros e 16 camadas proposta por Simonyan e Zisserman (2014) no artigo "Very Deep Convolutional Networks for Large-Scale Image Recognition". A arquitetura VGG16 ganhou notoriedade no desafio de reconhecimento de imagens ILSVRC (RUSSAKOVSKY et al., 2015) e influenciou arquiteturas subsequentes sendo frequentemente citada e utilizada em experimentos na literatura.
- **CNN Customizada** — A arquitetura customizada utilizada segue os mesmos princípios de construção da VGG16, porém com menos camadas. A rede utilizada contêm apenas 9 camadas e entre 10 e 25 milhões de parâmetros, a depender do conjunto de dados. O propósito da utilização de uma rede de menor capacidade é enriquecer os resultados ao permitir a comparação das técnicas de interpretabilidade entre redes com diferentes capacidades (número de parâmetro ajustáveis).

4.3 Técnicas

As técnicas selecionadas nos experimentos são técnicas de atribuição de características aplicáveis a CNNs. Os critérios utilizados para a seleção das técnicas foram a relevância e a facilidade de reprodução, resguardando-se a pluralidade de abordagens. Diferentes técnicas comunicam de diferentes formas a atribuição, podendo distribuir a responsabilidade em *pixels*, regiões ou regiões segmentadas. Mais detalhes sobre as abordagens e técnicas selecionadas estão apresentados nas Seções 3.2 e 3.1, sendo as selecionadas:

- **Gradiente** — Técnica de atribuição simples apresentada por Simonyan, Vedaldi e Zisserman (2014). O mapa de saliência resultante atribui a cada *pixel* da imagem uma importância sobre a predição.
- **SmoothGrad** — Técnica que elabora a ideia básica da técnica Gradiente ao melhorar a visualização dos mapas baseados em gradientes com suavização de ruídos. (SMILKOV et al., 2017). O mapa de saliência resultante atribui a cada *pixel* da imagem uma importância sobre a predição.

- **Input-Grad** — Consiste em processar a imagem gerada pela técnica Gradiente reutilizando a entrada original. O mapa de saliência resultante atribui a cada *pixel* da imagem uma importância sobre a predição.
- **Integrated Gradients** — Técnica baseada em gradientes em que os gradientes são acumulados ao longo de uma interpolação entre imagens, que tem como início uma imagem base e tem como fim a imagem sendo explicada. O mapa de saliência resultante atribui a cada *pixel* da imagem uma importância sobre a predição (SUNDARARAJAN; TALY; YAN, 2017).
- **Guided Integrated Gradients** — Técnica que expande a técnica *Integrated Gradients* ao mudar o caminho da interpolação visando a redução dos ruídos produzidos. O mapa de saliência resultante atribui a cada *pixel* da imagem uma importância sobre a predição.
- **Blur Integrated Gradients** — Expande a técnica *Integrated Gradients* ao acumular os gradientes ao longo de sucessivos desfoques gaussianos (*Gaussian Blur*) com diferentes resoluções de desfoque (*blur kernel*). O mapa de saliência resultante atribui a cada *pixel* da imagem uma importância sobre a predição.
- **GradCam** — Técnica que utiliza os filtros convolucionais mais próximos da última camada da rede para construir mapas de calor sobre a área que explica uma determinada classificação. O mapa de atribuição resultante atribui a regiões da imagem uma importância sobre a predição.
- **RISE** — Técnica baseada em perturbação que utiliza máscaras aleatórias aplicadas às imagens de entrada para então verificar que regiões impactam na predição quando estas são ocultadas. O mapa de atribuição resultante atribui a regiões da imagem uma importância sobre a predição.
- **XRAI** — Tem como base a técnica *Integrated Gradients*, no entanto, utiliza um segmentador de regiões para produzir um mapa de atribuição que comunica a atribuição através de regiões segmentadas.
- **LIME** — Técnica baseada em pertubações que segmenta a imagem e ajusta um modelo interpretável para informar que regiões influenciam positivamente ou negativamente a classificação final. O mapa de atribuição resultante atribui a regiões segmentadas da imagem uma importância sobre a predição.

4.4 Plano Experimental

O experimento foi subdividido em três etapas empreendidas em sequência: treinamento, interpretabilidade e avaliação das explicações produzidas pelas técnicas de in-

terpretabilidade. As etapas do estudo estão ilustradas em um diagrama apresentado na Figura 15. Na etapa de treinamento foram produzidos quatro modelos, cujas saídas foram explicadas pelas dez técnicas na etapa de interpretabilidade.

Para a etapa de treinamento os conjuntos de dados coletados foram subdivididos em treino e teste, assim, os modelos foram treinados no conjunto de treino e avaliados no conjunto de testes. Para o modelo VGG16 foi possível realizar uma etapa de transferência de conhecimento e ajuste fino (*transfer learning e fine tuning*) (BOZINOVSKI, 2020; ZHANG et al., 2020), que auxiliou a convergência do modelo ao utilizar como ponto de partida para o treinamento os pesos previamente ajustados em outro contexto. Os gráficos que apresentam as curvas de convergências estão presentes no Apêndice A.

Para a etapa de interpretabilidade foram separados 25 exemplos de cada classe do conjunto de teste, não utilizados na fase de treinamento, para serem explicados pelas técnicas a serem avaliadas. As bibliotecas Saliency (RESEARCH, 2021), RISE (PALATNIK, 2022) e LIME (RIBEIRO, 2022) foram utilizadas para a etapa de interpretabilidade.

Três avaliações foram utilizadas para comparar as técnicas de interpretabilidade. A primeira avaliação, intitulada Guia para Pertubações, segue a abordagem de realizar uma sondagem das saídas do modelo conforme pertubações na entrada se acumulam, como descrito na Subseção 3.3.1. A avaliação seguinte, intitulada Randomização dos Rótulos, segue a abordagem de comparar explicações, como descrito na Subseção 3.3.2. A última avaliação, intitulada Jogo de Apontar, segue a abordagem que utiliza conjuntos de dados anotados, como descrito na Subseção 3.3.3. As avaliações conduzidas estão detalhadas nas seções a seguir.

4.4.1 Guia para Pertubações

A avaliação Guia para Pertubações, implementada a partir do código disponibilizado por Goh et al. (2021), tem como princípio de funcionamento realizar pertubações em sequência na imagem de teste. Cada técnica é utilizada como guia para pertubações na imagem a ser predita, sendo as regiões ditas mais importantes perturbadas primeiro. O acúmulo de pertubações adultera a imagem original e forma um decaimento na confiança do modelo na predição da imagem. Um decaimento mais rápido indica que a técnica orienta melhor as regiões importantes para o modelo. Nessa avaliação as explicações servem de roteiro para uma sequência de pertubações, causando uma queda na confiança do modelo nas predições conforme as pertubações vão se acumulando.

O experimento Guia para Pertubações está ilustrado na Figura 16, na qual o mapa de saliência é utilizado como roteiro para várias etapas de pertubação que, quando passadas para o modelo, produzem um gráfico de queda na confiança das predições. Ao se avaliar uma única técnica, são geradas linhas de decaimento para cada imagem testada.

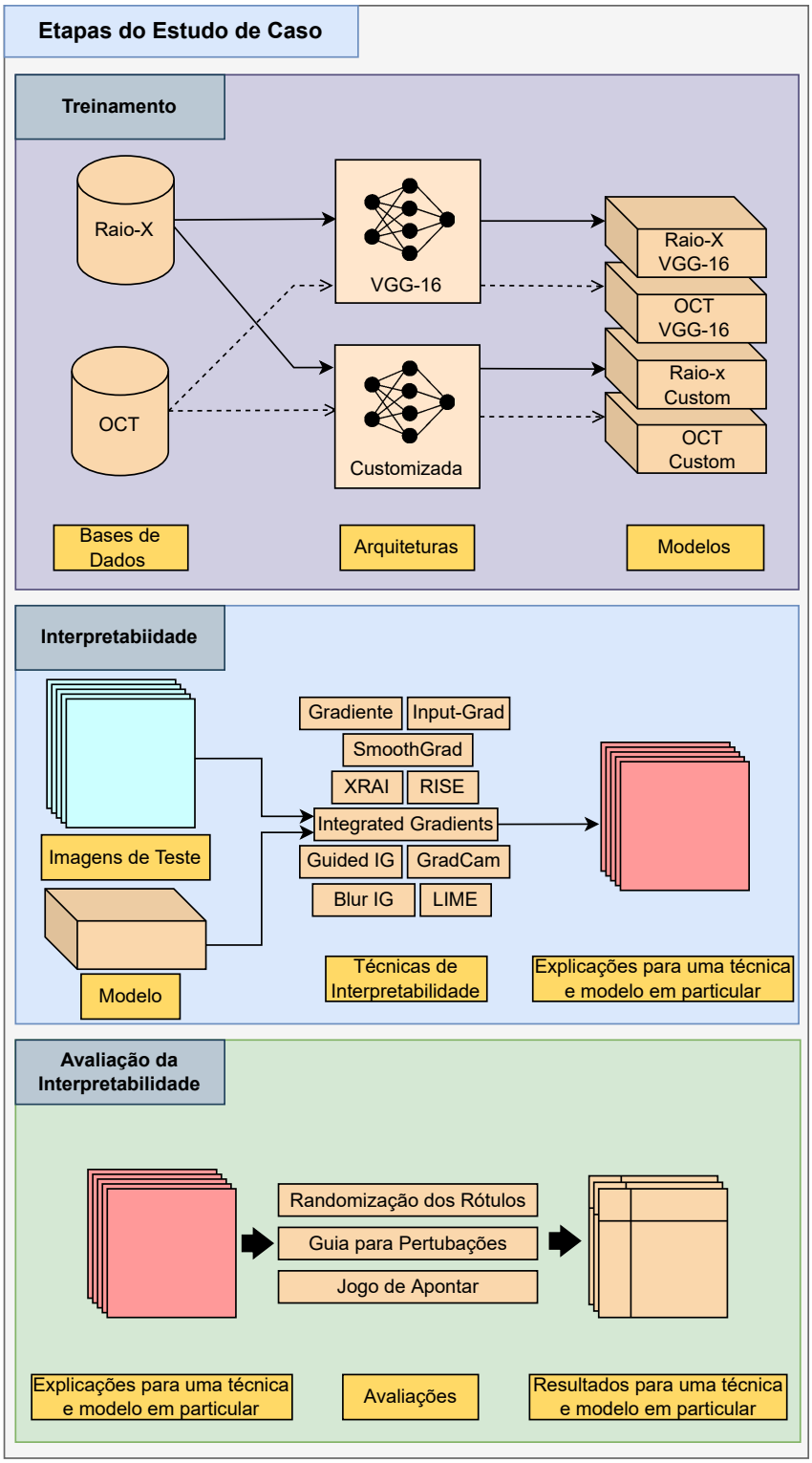


Figura 15 – Diagrama contendo as etapas do estudo de caso. Para a etapa do treinamento, quatro modelos são produzidos. Para a etapa de interpretabilidade, diversas explicações são produzidas para cada uma das técnicas. Para a etapa da avaliação das técnicas de interpretabilidade são produzidos resultados que indicam a qualidade das técnicas de interpretabilidade.

Entretanto, para comparação entre técnicas, as linhas de uma mesma técnica são agrupadas em uma linha média, produzindo uma linha por técnica. A métrica de comparação entre técnicas é a área abaixo da curva média.

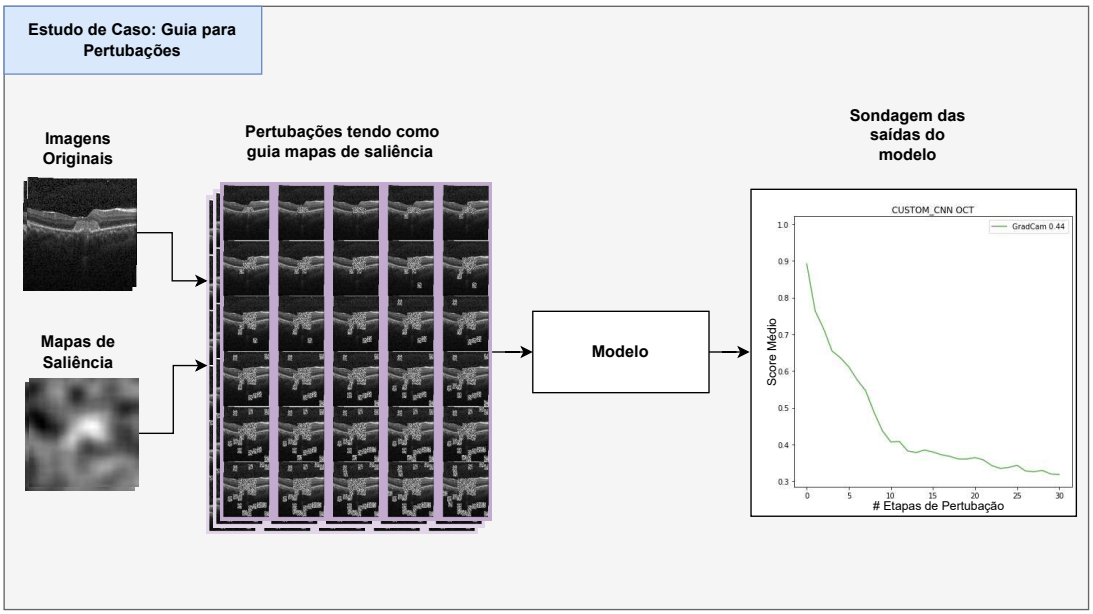


Figura 16 – Diagrama do experimento de guia para pertubações. Os mapas de saliência são utilizados com guias para pertubações. As confianças nas predições são monitoradas conforme as imagens modificadas são classificadas pelo modelo. O eixo X informa a quantidade de etapas de pertubação e o eixo Y a média da queda de confiança.

4.4.2 Randomização dos Rótulos

A avaliação Randomização dos Rótulos, proposta por Adebayo et al. (2018), tem como princípio de funcionamento a comparação entre as explicações fornecidas por dois modelos treinados de maneiras diferentes. Um modelo é treinado com rótulos verdadeiros e o outro modelo é treinado com rótulos aleatórios. A intuição é que os dois modelos aprendem a extrair características discriminativas a partir das imagens, mas apenas o modelo verdadeiro utiliza as características extraídas de maneira alinhada e direcionada com os rótulos verdadeiros.

O experimento Randomização dos Rótulos está ilustrado na Figura 17, na qual o modelo *"True"* e o modelo *"Random"* produzem explicações diferentes seguindo a mesma técnica de explicação. A métrica utilizada para avaliar a técnica é a correlação de Spearman entre imagens. Quanto maior a similaridade entre as explicações pior é a técnica, ou seja, se as explicações são similares, independente do modelo ser despropositado ou coerente, a técnica é indiferente aos rótulos do modelo.

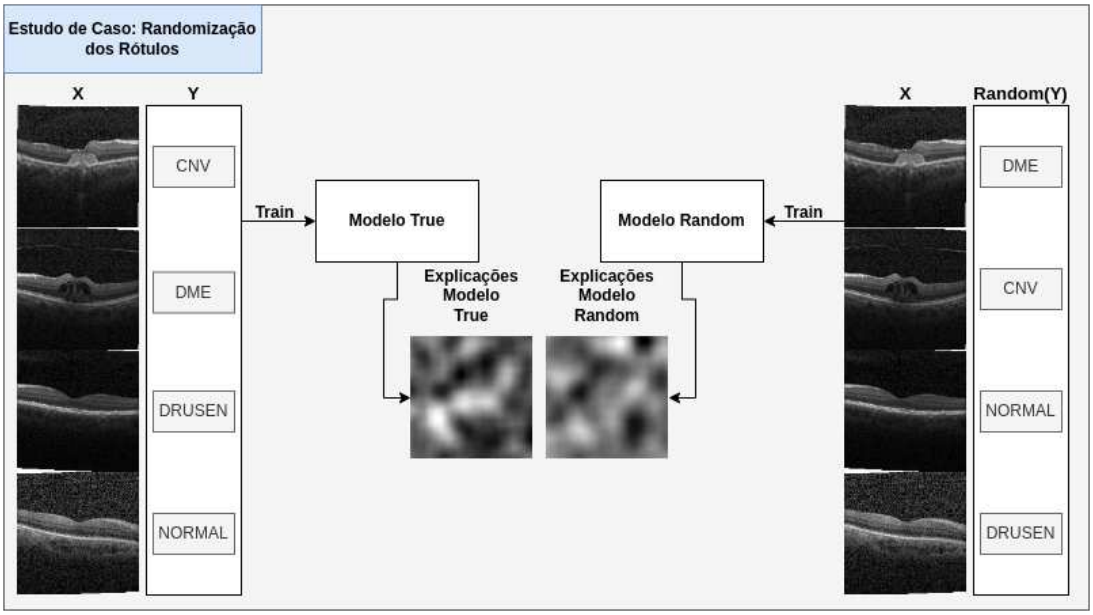


Figura 17 – Diagrama do experimento de randomização dos rótulos. Dois modelos são treinados e as explicações são comparadas.

4.4.3 Jogo de Apontar

A avaliação Jogo de Apontar (ou *Pointing Game*), proposta por Zhang et al. (2016), segue a abordagem de utilizar conjuntos de dados especiais para avaliar técnicas de interpretabilidade. Na avaliação original, um acerto é contabilizado se o ponto máximo de um mapa de atribuição estiver contido na anotação de localização do objeto classificado, caso contrário, um erro é contabilizado. Por exemplo, dada uma imagem de um cão corretamente classificada e uma explicação para essa classificação, se o ponto de máxima importância na explicação estiver inserido na anotação de onde está o cão, considera-se a explicação correta. A intuição da avaliação é a de que as técnicas não devem explicar as classificações utilizando o *backgroud* e sim o objeto principal da imagem.

A avaliação original é utilizada em técnicas que têm como saída mapas com atribuição ao nível de *pixels*, sendo imprecisa para técnicas que comunicam a atribuição no formato de regiões segmentadas, como nas técnicas LIME, XRAI e RISE, exemplificadas na Figura 5. Para essas técnicas não existe um ponto de máxima atribuição e sim uma região de máxima atribuição. Como adaptação, foi contabilizado a quantidade de *pixels* de máxima atribuição inseridos na localização sobre a quantidade total de *pixels* de máxima atribuição.

4.5 Detalhes de Implementação

O código produzido e instruções necessárias para reprodução estão organizados em um repositório público, disponível em matheusgmaia/CNN-Interpretability [1]. A seguir estão descritas as configurações da máquina utilizada para condução dos experimentos e principais bibliotecas de *software* utilizadas.

- Configurações da Máquina
  - Sistema Operacional — Pop!OS 20.04 LTS
  - Processador — AMD® Ryzen 5 3600x 6-core processor × 12
  - Memória — 16Gb DDR4
  - Placa de Vídeo — NVIDIA Corporation GeForce GTX 1660, 6GB, GDDR6, 192bit
- Principais Bibliotecas — Treinamento e Inferência
  - Tensorflow [2] — A biblioteca utilizada para o treinamento das redes neurais foi a Tensorflow, desenvolvida pelo Google e lançada em código aberto em 2015. A API de alto nível do TensorFlow tf.keras foi utilizada para criar e treinar os modelos de aprendizado profundo.
  - OpenCV [3] — Biblioteca com código aberto utilizada para manipulação de imagens digitais.
- Principais Bibliotecas — Interpretabilidade
  - Saliency [4] — Biblioteca com implementação de diversas técnicas de atribuição de CNNs desenvolvida pela equipe *Google PAIR People + AI Research* (RESEARCH, 2021).
  - RISEtf [5] — Implementação da técnica RISE para modelos treinados com a biblioteca Tensorflow (PALATNIK, 2022; PETSIUK; DAS; SAENKO, 2018).
  - LIME [6] — Biblioteca para utilização da técnica LIME, disponibilizada pelo autor do artigo original (RIBEIRO, 2022; RIBEIRO; SINGH; GUESTRIN, 2016).

[1] Disponível em: https://github.com/matheusgmaia/CNN-Interpretability. Acesso em: 17 abr. 2022
[2] Disponível em: https://www.tensorflow.org/. Acesso em: 17 abr. 2022
[3] Disponível em: https://opencv.org/. Acesso em: 17 abr. 2022
[4] Disponível em: https://github.com/PAIR-code/saliency. Acesso em: 17 abr. 2022
[5] Disponível em: https://github.com/palatos/RISE_tf. Acesso em: 17 abr. 2022
[6] Disponível em: https://github.com/marcotcr/lime. Acesso em: 17 abr. 2022

4.6 Considerações Finais

O estudo de caso conduzido utiliza dois conjuntos de dados médicos para realizar as etapas de treinamento, interpretabilidade e avaliação da interpretabilidade. As arquiteturas utilizadas para o treinamento se diferem em capacidade, para diversificar os resultados. As técnicas avaliadas se diferem na abordagem para distribuição da responsabilidade da predição entre as características da imagem.

Os resultados e discussões estão apresentados na Seção 5, onde para cada avaliação, os atores que influenciam no resultado e balizam a discussão são: base de dados, arquitetura e técnica de interpretabilidade.

5 Resultados e Discussões

No presente capítulo estão apresentados e discutidos os resultados das três avaliações explicadas nas Subseções 4.4.1, 4.4.2 e 4.4.3 do capítulo anterior. Duas das avaliações (Guia para Pertubações e Randomização dos Rótulos) produziram um resultado para cada combinação de conjunto de imagens (Raio-X ou OCT) e arquitetura convolucional (VGG16 ou Customizada). Para a terceira avaliação, Jogo de Apontar, apenas o conjunto de imagens de Raio-X foi utilizado, por ser uma avaliação que necessita de um conjunto com localizações anotadas.

5.1 Guia para Pertubações

Na Figura 18 estão apresentados os resultados da avaliação Guia para Pertubações. Cada gráfico apresenta o resultado obtido considerando a combinação de um conjunto particular de dados e arquitetura, a qual se encontra especificada no título do gráfico. O eixo X informa a quantidade de etapas de pertubação e o eixo Y a média da confiança do modelo nas imagens avaliadas. Em cada gráfico com resultado, cada curva colorida representa uma técnica diferente, conforme a legenda apresentada.

As técnicas LIME e GradCam foram o destaque negativo nos resultados obtidos. A técnica LIME é dada como segunda pior técnica nos quatro resultados, enquanto a técnica GradCam ocupa uma das três piores posições em três resultados. Como destaque positivo tem-se a técnica BlurIG, que ocupa uma das três melhores posições em três resultados. Existe uma significativa diferença entre o perfil dos resultados no conjunto de dados OCT e no conjunto de dados Raio-X, sendo as quedas do conjunto OCT mais abruptas.

Conforme as pertubações se acumulam no eixo X a confiança média nas predições tende a cair no eixo Y. As técnicas podem ser comparadas utilizando-se a área abaixo da curva, que está calculada para cada técnica nas legendas. A intuição é a de que quanto mais abrupto for o decaimento da confiança das predições, menor é a área abaixo da curva, melhor será a técnica, já que as pertubações guiadas pelas atribuições da técnica impactaram negativamente as predições do modelo via perturbações.

Os resultados foram pouco consistentes e uma hipótese para a inconsistência é o impacto do efeito de “entradas fora da distribuição”: conforme as imagens vão sendo perturbadas é normal que a rede não saiba lidar com entradas adulteradas, muito diferentes daquelas treinadas, produzindo classificações desamparadas em experiência. Quanto maior o número de classes de um problema, mais oportunidades para a classe resultante da “confusão” não ser a classe para a qual estamos acompanhando as confianças. A di-

ferença entre quantidade de classes, portanto, é uma hipótese plausível para a diferença entre o perfil de resultado entre os conjuntos Raio-X e OCT.

A concisão e a precisão em apontar o que mais impacta na predição é o que importa para a avaliação, favorecendo técnicas que atribuam com a maior importância as regiões de fato impactantes na predição. As principais dificuldades encontradas para realizar a avaliação desta seção foram: complexidade de implementação, susceptibilidade ao efeito de entradas fora da distribuição e elevado custo de processamento, já que o modelo precisa ser consultado para cada nova etapa de perturbação.

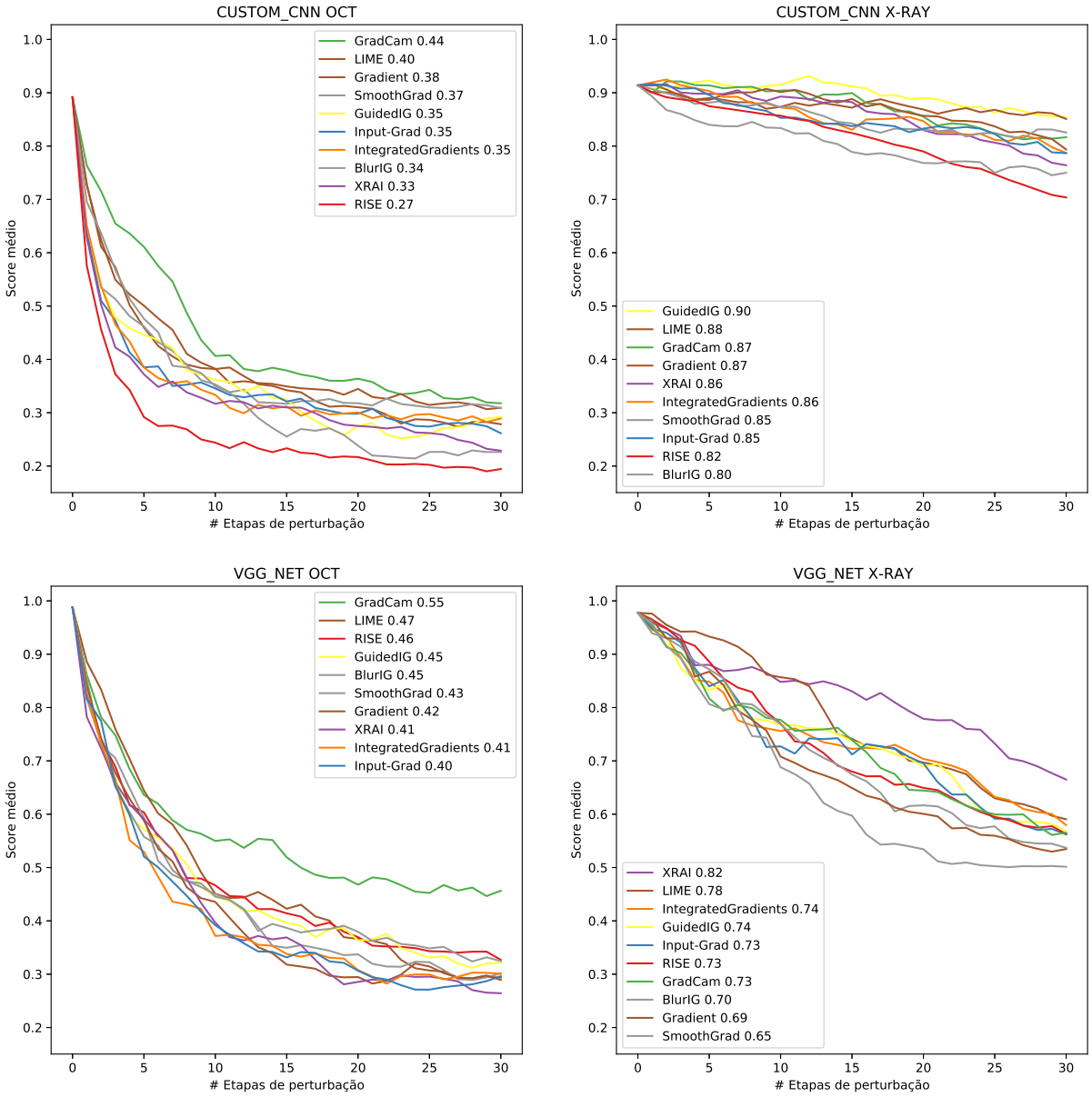


Figura 18 – Resultados do experimento Guia para Pertubações — Cada curva representa uma técnica e as técnicas são comparáveis por meio da área abaixo da curva (AUC, apresentada na legenda).

5.2 Randomização dos Rótulos

Nas Figuras 19, 20, 21 e 22 estão apresentados os resultados da avaliação da técnica de Randomização dos Rótulos. Em cada resultado tem-se no eixo X um intervalo de confiança para o valor absoluto da correlação entre duas imagens para as diferentes técnicas, listadas ao longo do eixo Y. As técnicas foram avaliadas em dois conjuntos de dados e com duas arquiteturas neurais diferentes, especificadas no título de cada resultado. Para cada técnica tem-se um intervalo de confiança de 95% que representa o intervalo de valores plausíveis para as correlações entre explicações, produzidos com o método de reamostragem *Bootstrap BCa* (DICICCIO; ROMANO, 1988). Cada técnica tem uma cor que indica como a atribuição é retratada pela técnica, detalhada na legenda da figura.

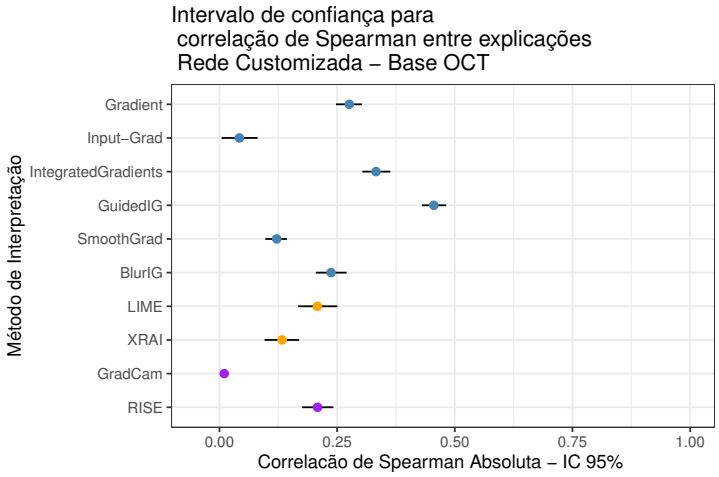


Figura 19 – Resultados da avaliação Randomização dos Rótulos para o conjunto de dados OCT e rede treinada com arquitetura customizada. As técnicas em azul retratam as atribuições ao nível de *pixels*. As técnicas em laranja utilizam regiões segmentadas. As técnicas em roxo utilizam regiões suavizadas.

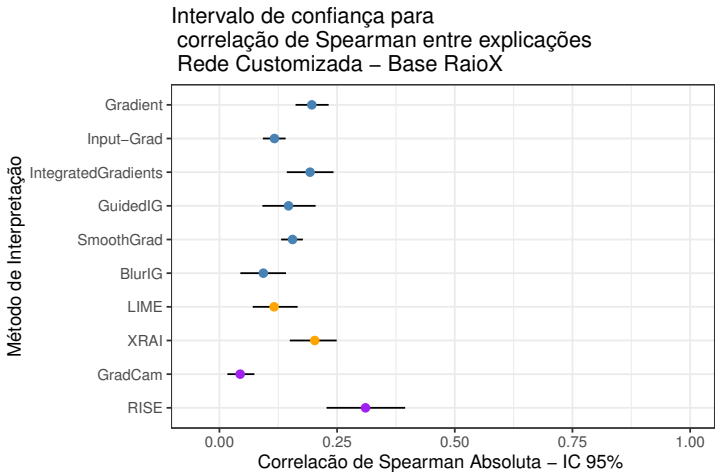


Figura 20 – Resultados da avaliação Randomização dos Rótulos para o conjunto de dados Raio-X e rede treinada com arquitetura customizada. As técnicas em azul retratam as atribuições ao nível de *pixels*. As técnicas em laranja utilizam regiões segmentadas. As técnicas em roxo utilizam regiões suavizadas.

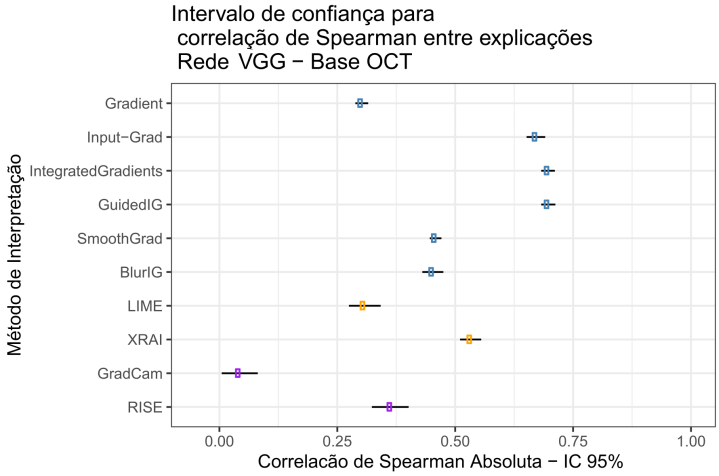


Figura 21 – Resultados da avaliação Randomização dos Rótulos para o conjunto de dados OCT e rede treinada com arquitetura VGG. As técnicas em azul retratam as atribuições ao nível de *pixels*. As técnicas em laranja utilizam regiões segmentadas. As técnicas em roxo utilizam regiões suavizadas.

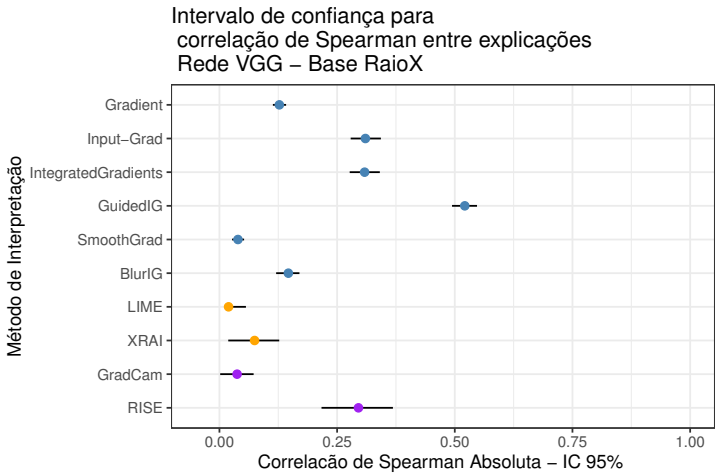
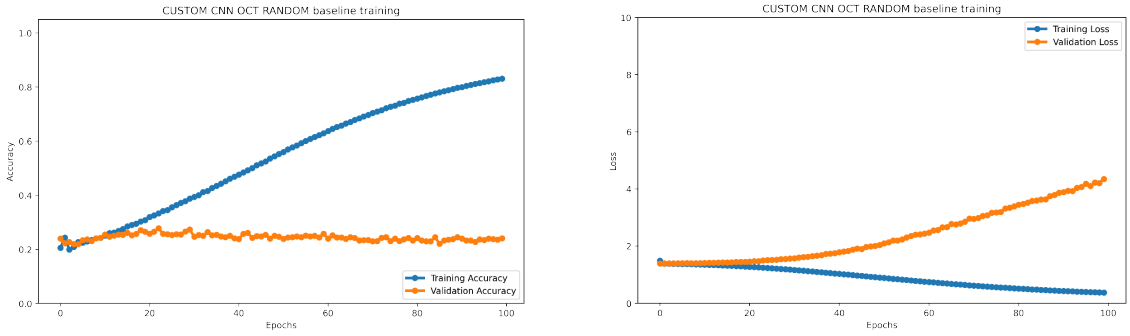


Figura 22 – Resultados da avaliação Randomização dos Rótulos para o conjunto de dados Raio-X e rede treinada com arquitetura VGG. As técnicas em azul retratam as atribuições ao nível de *pixels*. As técnicas em laranja utilizam regiões segmentadas. As técnicas em roxo utilizam regiões suavizadas.

A técnica GradCam foi o destaque positivo desta avaliação, em consonância com os resultados obtidos por Adebayo et al. (2018) para o mesmo tipo de avaliação em outros conjuntos de dados. A técnica GradCam utiliza os filtros convolucionais mais próximos da saída da rede e essa proximidade com os rótulos pode ser crucial para a sensibilidade da técnica relativa à mudança no modelo de rede neural sendo explicado.

A grande dificuldade para a execução da avaliação descrita nesta seção é a necessidade de um retreinamento com rótulos aleatórios que, pela incoerência nos rótulos, necessita de numerosas épocas de treinamento para que seja observada uma tendência de convergência (ou seja, ausência de variações de grande magnitude na função de perda usada para treinamento do modelo). A Figura 23 apresenta o gráfico para um dos treinamentos realizados com rótulos aleatórios. No conjunto de treinamento, curva de cor azul, os rótulos foram aleatorizados, e no conjunto de validação, curva de cor laranja, os rótulos continuam corretos. As redes treinadas com rótulos aleatórios alcançam alta acurácia em treino, mas têm acurácia similar a um classificador aleatório para o conjunto com rótulos originais. Analisando-se as explicações, mesmo com o despropósito da rede randomizada, algumas técnicas produziram explicações semelhantes para os diferentes treinamentos, como nas técnicas *GuidedIG*, *IntegratedGradients* e *Input-Grad*, o que representa é um comportamento não desejável para uma boa técnica de explicabilidade.



(a) Acurácia (b) *Loss*

Figura 23 – Histórico de (a) acurácia e (b) perda (*loss*) de treinamento para a rede randomizada customizada com a base OCT.

5.3 Jogo de Apontar

Para o experimento Jogo de Apontar é necessário ter-se disponível a anotação que separa a região maior de interesse do *background*. A técnica original, proposta por Zhang et al. (2016), utiliza as anotações presentes nos conjuntos de dados próprios para desafios de segmentação, como COCO (LIN et al., 2014) e Pascal (EVERINGHAM et al., 2012). Entretanto, para os conjuntos escolhidos não existem tais anotações. Para obter as anotações de localização foi utilizado um segmentador de pulmão em imagens de Raio-X desenvolvido por Kim et al. (2020). O segmentador proposto utiliza as saídas de cinco diferentes segmentadores, além de processamentos de imagens para produzir uma saída final, em um processo denominado *Ensemble*. Além dos processamentos originais foi realizado um preenchimento da área interna entre as duas segmentações retornadas pelo segmentador para destacar melhor a área de interesse em contraponto com o que é definitivamente *background* no contexto de classificação de pneumonia. Na Figura 24 estão apresentados três exemplos com etapa intermediária e resultado da segmentação realizada. Na Figura 24 as entradas estão na esquerda, no centro estão apresentadas as segmentações resultantes da união dos cinco segmentadores além do processamento, como no artigo original, e na direita, os exemplos foram preenchidos para destacar apenas a área de possível interesse do plano de fundo.

Nas Figuras 25 e 26 estão apresentados os dois resultados produzidos. As barras coloridas apontam a proporção dos *pixels* de máxima atribuição inseridos na região segmentada. As cores de cada barra do gráfico indicam o nível com que as atribuições são representadas, conforme discriminado nas legendas das figuras. Existe uma perceptível diferença entre o patamar das proporções nos resultados. Os resultados associados ao modelo customizado são inferiores aos resultados do modelo VGG. Pode-se dizer que as técnicas que explicam o modelo customizado utilizaram mais do *background*, o que pode ser uma falha nas técnicas, ou uma falha no próprio modelo, que pode conter um viés no

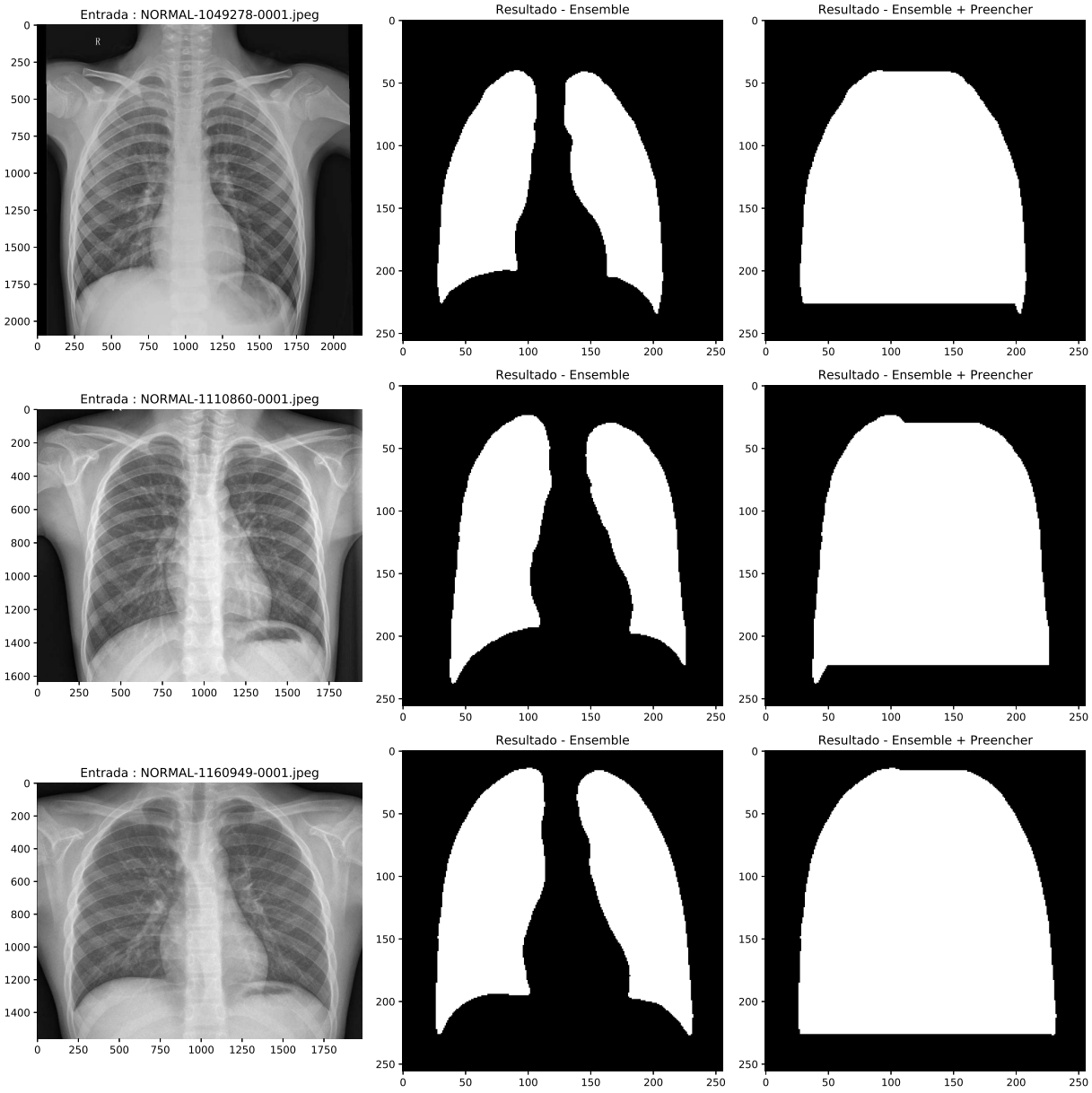


Figura 24 – Exemplos da segmentação utilizada para a avaliação Jogo de Apontar. Na primeira coluna estão as entradas para o segmentador. Na segunda coluna estão apresentados os resultados das segmentações utilizando a união da saída de cinco de segmentadores além de processamentos, como proposto por Zhang et al. (2016). Na terceira coluna as segmentações foram processadas para obter um maior destaque da área de plano de fundo, que não deve ser atribuída com importância para a classificação.

seu treinamento. A técnica RISE teve o melhor resultado, pois fez uso de menor área de *background* para explicar as predições.

Em situações ideais é esperado que uma explicação não atribua importância ao *backgorund*, entretanto, caso o modelo utilize informações do plano de fundo para realizar a classificação, a explicação deve ser fiel ao modelo e atribuir responsabilidade para essas regiões. É difícil apontar a causa da utilização do *background* nas explicações, podendo ser uma falha na técnica de atribuição ou a presença de um viés no conjunto de dados.

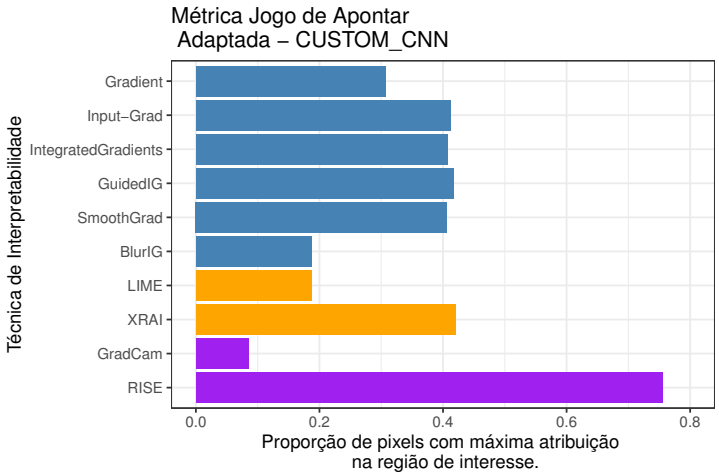


Figura 25 – Resultados da avaliação Jogo de Apontar com adaptações para o conjunto de dados raio-X e rede treinada com arquitetura customizada. As técnicas registradas em azul retratam as atribuições ao nível de *pixels*. As técnicas registradas em laranja utilizam regiões segmentadas. As técnicas registradas em roxo utilizam regiões suavizadas.

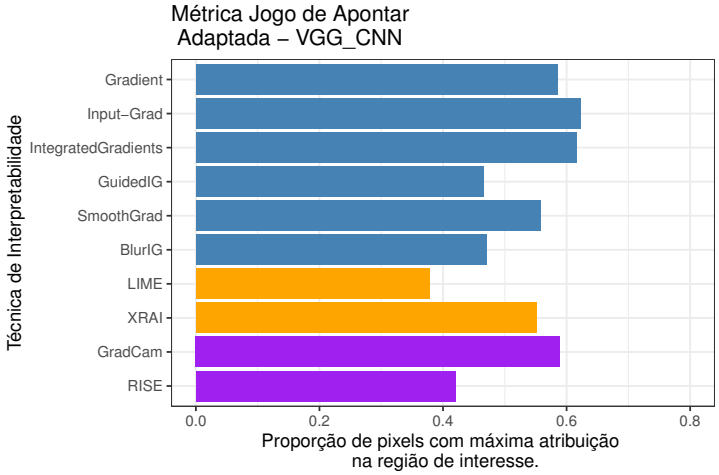


Figura 26 – Resultados da avaliação Jogo de Apontar com adaptações para o conjunto de dados Raio-X e rede treinada com a arquitetura VGG-16. As técnicas em azul retratam as atribuições ao nível de *pixels*. As técnicas em laranja utilizam regiões segmentadas. As técnicas em roxo utilizam regiões suavizadas.

Realizar a avaliação em questão é simples caso as anotações de região de interesse estejam disponíveis. A utilização de um segmentador para enriquecer o conjunto de dados com as anotações necessárias para a avaliação é uma inovação promissora, podendo ser replicada em outros conjuntos médicos que possuem segmentadores disponíveis. Conjuntos de imagens provenientes de exames de mamografia, utilizados tanto para a tarefa de classificação de câncer de mama, quanto para a tarefa segmentação de nódulos cancerígenos, são conjuntos promissores para avaliações semelhantes.

6 Considerações Finais

Este capítulo resume as contribuições deste estudo e introduz algumas oportunidades para futuras pesquisas relacionadas à interpretabilidade em CNNs. Na Seção 6.1 estão apresentadas as conclusões da pesquisa e na Seção 6.2 estão indicadas propostas de pesquisas futuras.

6.1 Conclusões

Os avanços no ramo de pesquisa de aprendizado profundo romperam limites sobre o que é possível ser realizado por um computador. Entretanto, a utilização de aprendizado profundo em áreas de alto risco, como a área médica, é inibida por desconfianças e regulações. A interpretabilidade se faz necessária não apenas para aumentar a confiança e o dinamismo entre o usuário e o sistema computadorizado, mas também atender conformidade às leis que asseguram direito de explicação, uma das barreiras para a ampla utilização de aprendizado profundo. A presente pesquisa contribui com o entendimento da área de interpretabilidade de Redes Neurais Convolucionais ao organizar e integrar conhecimentos presentes na literatura na forma de revisão bibliográfica e por meio de um estudo de caso.

Os conhecimentos sobre interpretabilidade de CNNs foram debatidos tendo como ponto de partida para a discussão uma fundamentação teórica sobre aprendizado de máquina e interpretabilidade. Diversos conceitos foram apresentados e organizados em taxonomias mais digeríveis, como a organização dos objetivos e características do uso da interpretabilidade proposta no Capítulo 2. Os objetivos relacionados ao uso da interpretabilidade foram agrupados em três: Ética e Regulação, Suporte à Decisão e Confiança e Entendimento. Os objetivos foram relacionados com as características de Auditabilidade, Explicabilidade e Transparência, esclarecendo esses termos comumente relacionados à interpretabilidade.

No Capítulo 3 as diferentes abordagens sobre interpretabilidade de redes convolucionais foram apresentadas e o vocabulário construído ao longo da dissertação foi utilizado para tipificar abordagens e técnicas proeminentes. As principais abordagens, Atribuição e Visualização, foram discutidas e o funcionamento das diferentes sub abordagens foi descrito. No decorrer da pesquisa foram apresentados debates científicos presentes na literatura, como a relação entre o campo da interpretabilidade e o campo dos ataques adversariais e a questão da avaliação objetiva das técnicas, discutida com mais ênfase na Seção 3.3.

As discussões apresentadas nos capítulos iniciais foram postas em prática em um estudo de caso sobre interpretabilidade de redes neurais contemplando a avaliação de técnicas do estado da arte. O estudo foi realizado no contexto de imagens médicas, não usual em pesquisas de comparação entre explicações correlatas, que utilizam, tipicamente, conjuntos de dados sintéticos ou gearados a partir de banco de imagens massivos, como *Mnist* (DENG, 2012), *FashionMnist* (XIAO; RASUL; VOLLGRAF, 2017) ou *Imagenet* (DENG et al., 2009). Foram realizadas de ponta-a-ponta as etapas de treinamento, interpretabilidade e avaliação das técnicas. Os estudos realizados podem ser facilmente reproduzidos em outros conjuntos de dados. Um pesquisador interessado na solução de determinado problema via treinamento de CNNs poderá usufruir dos códigos desenvolvidos para obter, além de um modelo treinado para o problema, também um modelo explicador devidamente avaliado.

As explicações estão diretamente atreladas à qualidade do modelo. Para uma justa comparação entre as técnicas, a utilização de conjuntos de dados mais robustos e experimentados, com menos vieses, seria o ideal. A inovação de utilizar conjuntos de imagens médicas para as avaliações trouxe desafios por serem conjuntos mais difíceis de se tratar. Foram encontrados vícios como forte desbalanceamento entre classes e presença de artefatos causados por aumento de dados prévio. Entretanto, o objetivo do estudo foi justamente identificar esses desafios e induzir aprendizados a partir das dificuldades enfrentadas. Seria difícil tecer uma conclusão global sobre as técnicas uma vez que as avaliações têm diferentes enfoques. A avaliação Guia para Pertubações, valoriza a concisão das técnicas, a avaliação Randomização dos Rótulos preza pela coerência entre a explicação e o contexto original do modelo e a avaliação Jogo de Apontar exalta a coerência espacial da explicação. Dentre as técnicas avaliadas, as que se destacaram foram as GradCAM e RISE.

Foram documentadas as dificuldades em realizar o estudo de caso. A avaliação Guia para Pertubações é uma avaliação custosa e complexa, além de estar sujeita a um fenômeno conhecido como entradas fora da distribuição. A avaliação Randomização dos Rótulos exige um retreino custoso, podendo servir mais como teste de sanidade do que como comparativo entre explicações. A avaliação Jogo de Apontar exige anotações presentes apenas em conjuntos específicos, no entanto, foi realizada com o auxílio de um segmentador robusto, o que é uma inovação promissora.

Os aprendizados extraídos da literatura e estudo de caso estão resumidos e apresentados nos Quadros 2 e 3. O Quadro 2 apresenta um resumo das contribuições dos Capítulos Fundamentação e Revisão Bibliográfica. O Quadro 3 apresenta os aprendizados extraídos a partir dos desafios enfrentados na condução do estudo de caso.

Quadro 2 – Quadro com resumo dos aprendizados identificados a partir da literatura e apresentados nos Capítulos Fundamentação e Revisão Bibliográfica.

Conceito-chave	Aprendizado
Objetivos das técnicas de interpretabilidade	Os objetivos relacionados com interpretabilidade podem ser agrupados em: Ética e Regulação, Suporte à Decisão e Confiança e Entendimento. Para determinar qual é o sentido pretendido na utilização do termo interpretabilidade, deve-se questionar o objetivo da interpretabilidade em cada uso.
Características de modelos interpretáveis	Auditabilidade, Explicabilidade e Transparência são características que um modelo pode possuir e estão relacionadas com os objetivos anteriormente apresentados.
Abordagens para interpretabilidade de CNNs	Os métodos de interpretabilidade de CNN mais comuns seguem duas abordagens principais, quais sejam: Visualização de Características e Atribuição de Características. Existem aplicações que tentam unir ambas as abordagens em uma única interface, assim como existem abordagens que não tentam expor características aprendidas, mas sim interpretar o modelo através de outros sinais.
Abordagens para avaliação de técnicas de interpretabilidade de CNNs	Um ponto de evolução perceptível nas pesquisas dessa área é a questão da validação das técnicas. Uma proposta de categorização para as avaliações quantitativas seria Sondagem das Saídas do Modelo, Comparação Entre Explicações e Utilização de Conjuntos de Dados Anotados.

Quadro 3 – Quadro com resumo dos aprendizados identificados a partir do estudo de caso e apresentados no Capítulo Resultados e Discussões.

Conceito-chave	Aprendizado
Guia para Pertubações	A concisão e a precisão em apontar o que mais impacta na predição é o que importa para a avaliação, favorecendo técnicas de interpretabilidade que atribuam com a maior importância as regiões de fato impactantes na predição. As dificuldades para realização da avaliação são: complexidade de implementação, susceptibilidade ao efeito de entradas fora da distribuição e elevado custo de processamento.
Randomização dos Rótulos	Avaliação que afere a discrepância entre as explicações produzidas por um modelo coerente e um modelo treinado com rótulos aleatórios. Técnicas coerentes, que realmente expõem as características discriminantes no contexto original são favorecidas. A maior dificuldade para a realização da avaliação é o custo em treinar os modelos com rótulos aleatórios até que alcancem convergência.
Jogo de Apontar	A avaliação Jogo de Apontar avalia a coerência espacial da explicação, depreciando técnicas que utilizam o plano de fundo como explicação. A avaliação necessita, idealmente, de dados anotados manualmente. Uma adaptação é utilizar um segmentador robusto para obter as anotações necessárias. Foram necessárias adaptações na métrica final para melhor avaliar técnicas que atribuem responsabilidade seguindo diferentes abordagens. Para alguma técnicas existe um ponto de máxima atribuição. Para outras existe uma região de máxima atribuição.

6.2 Propostas para Pesquisas Futuras

Uma continuação natural do trabalho seria expandir a quantidade de conjuntos de dados e arquiteturas avaliadas. Conjuntos de dados médicos, como aqueles apresentados por (WANG et al., 2017), podem enriquecer o debate por serem conjuntos mais volumosos do que os conjuntos utilizados. Com mais arquiteturas seria possível se aprofundar na relação entre características da rede e a atuação das técnicas de atribuição. Possíveis características a serem incluídas na avaliação são resolução das imagens de entrada, a profundidade, a largura das redes, a inclusão de *Batch Normalization*, entre outros. Tais

experimentos poderiam indicar quais características teriam maior influência na geração de modelos facilmente interpretáveis pelas técnicas investigadas nesta dissertação.

Uma etapa prática adicional seria o empacotamento das técnicas e avaliações em uma biblioteca ou ferramenta que permitisse maior reprodutibilidade dos experimentos. Tal ferramenta teria como produto para o usuário um explicador devidamente avaliado no contexto utilizado. A ferramenta teria como escopo a depuração de modelos convolucionais visando a aumentar a confiança nos modelos, fornecendo explicadores apropriados. A ferramenta poderia ser expandida com outras funcionalidades relacionadas com robustez de modelos, como avaliação da resistência a ataques adversariais.

Outra vertente identificada na literatura é a de técnicas que unem diferentes abordagens de forma interativa, proporcionando maior cooperação homem-máquina ao permitir uma investigação dos modelos treinados, indo além das explicações estáticas, como mapas de calor. A união das abordagens de visualização e atribuição também é uma vertente citada no trabalho alinhada à interatividade, como retratado com maior profundidade por Olah et al. (2018).

Referências

AAS, K.; JULLUM, M.; LØLAND, A. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, Elsevier, p. 103502, 2021. Citado na página 34.

ADEBAYO, J. et al. Sanity checks for saliency maps. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2018. p. 9505–9515. Citado 7 vezes nas páginas 29, 38, 41, 46, 50, 54 e 62.

AHERN, I. et al. Normlime: A new feature importance metric for explaining deep neural networks. *arXiv preprint arXiv:1909.04200*, 2019. Citado na página 34.

ALBER, M. et al. innvestigate neural networks! *Journal of Machine Learning Research*, v. 20, n. 93, p. 1–8, 2019. Disponível em: <http://jmlr.org/papers/v20/18-540.html>. Citado na página 22.

APLEY, D. W.; ZHU, J. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 82, n. 4, p. 1059–1086, 2020. Citado na página 20.

ARIK, S. O.; PFISTER, T. Protoattend: Attention-based prototypical learning. *Journal of Machine Learning Research*, v. 21, p. 1–35, 2020. Citado 3 vezes nas páginas 31, 32 e 37.

ARUN, N. et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, Radiological Society of North America, v. 3, n. 6, p. e200267, 2021. Citado na página 41.

BAU, D. et al. Network dissection: Quantifying interpretability of deep visual representations. *CoRR*, abs/1704.05796, 2017. Disponível em: <http://arxiv.org/abs-/1704.05796>. Citado na página 36.

BIBAL, A. et al. Legal requirements on explainability in machine learning. *Artificial Intelligence and Law*, Springer, v. 29, n. 2, p. 149–169, 2021. Citado na página 24.

BOZINOVSKI, S. Reminder of the first paper on transfer learning in neural networks, 1976. *Informatica*, v. 44, n. 3, 2020. Citado na página 52.

BRAAMS, J. *Babel, a multilingual package for use with LATEX's standard document classes*. [S.l.], 2008. Disponível em: <http://mirrors.ctan.org/info/babel/babel.pdf>. Acesso em: 17.2.2013. Citado na página 14.

BRAMHALL, S. et al. Qlime-a quadratic local interpretable model-agnostic explanation approach. *SMU Data Science Review*, v. 3, n. 1, p. 4, 2020. Citado na página 34.

CARTER, S. et al. Activation atlas. *Distill*, 2019. Https://distill.pub/2019/activation-atlas. Citado 3 vezes nas páginas 27, 32 e 36.

CHATTOPADHAY, A. et al. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: IEEE. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. [S.l.], 2018. p. 839–847. Citado 2 vezes nas páginas 32 e 33.

DAS, A.; RAD, P. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*, 2020. Citado 2 vezes nas páginas 24 e 29.

DENG, J. et al. Imagenet: A large-scale hierarchical image database. In: IEEE. *2009 IEEE conference on computer vision and pattern recognition*. [S.l.], 2009. p. 248–255. Citado 3 vezes nas páginas 14, 47 e 67.

DENG, L. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, IEEE, v. 29, n. 6, p. 141–142, 2012. Citado 2 vezes nas páginas 45 e 67.

DICICCIO, T. J.; ROMANO, J. P. A review of bootstrap confidence intervals. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 50, n. 3, p. 338–354, 1988. Citado na página 60.

DOMBROWSKI, A.-K. et al. Explanations can be manipulated and geometry is to blame. *arXiv preprint arXiv:1906.07983*, 2019. Citado na página 41.

DOSHI-VELEZ, F.; KIM, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017. Citado 2 vezes nas páginas 21 e 23.

DYK, D. A. V.; MENG, X.-L. The art of data augmentation. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 10, n. 1, p. 1–50, 2001. Citado na página 19.

ERHAN, D. et al. Visualizing higher-layer features of a deep network. *University of Montreal*, v. 1341, n. 3, p. 1, 2009. Citado na página 30.

ERHAN, D.; COURVILLE, A.; BENGIO, Y. Understanding representations learned in deep architectures. *Department dInformatique et Recherche Operationnelle, University of Montreal, QC, Canada, Tech. Rep*, v. 1355, p. 1, 2010. Citado na página 32.

ESTEVA, A. et al. A guide to deep learning in healthcare. *Nature medicine*, Nature Publishing Group, v. 25, n. 1, p. 24–29, 2019. Citado 2 vezes nas páginas 14 e 18.

EVERINGHAM, M. et al. *The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results*. 2012. Http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. Citado 4 vezes nas páginas 43, 44, 47 e 63.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, JSTOR, p. 1189–1232, 2001. Citado na página 20.

GARCÍA, M. V.; AZNARTE, J. L. Shapley additive explanations for no2 forecasting. *Ecological Informatics*, Elsevier, v. 56, p. 101039, 2020. Citado na página 34.

GHORBANI, A.; ABID, A.; ZOU, J. Interpretation of neural networks is fragile. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2019. v. 33, p. 3681–3688. Citado 5 vezes nas páginas 29, 38, 41, 42 e 46.

GHORBANI, A.; WEXLER, J.; KIM, B. Automating interpretability: Discovering and testing visual concepts learned by neural networks. *arXiv preprint arXiv:1902.03129*, 2019. Citado 5 vezes nas páginas 31, 32, 35, 38 e 46.

GOH, G. S. W. et al. Understanding integrated gradients with smoothtaylor for deep neural network attribution. In: *2020 25th International Conference on Pattern Recognition (ICPR)*. [S.l.: s.n.], 2021. p. 4949–4956. Citado 5 vezes nas páginas 40, 41, 42, 43 e 52.

GOLDSTEIN, A. et al. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 24, n. 1, p. 44–65, 2015. Citado na página 20.

GOODMAN, B.; FLAXMAN, S. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, v. 38, n. 3, p. 50–57, 2017. Citado na página 22.

GOYAL, Y.; SHALIT, U.; KIM, B. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*, 2019. Citado 2 vezes nas páginas 32 e 36.

GREENE, T. et al. Adjusting to the gdpr: The impact on data scientists and behavioral researchers. *Big data*, Mary Ann Liebert, Inc., publishers 140 Huguenot Street, 3rd Floor New . . . , v. 7, n. 3, p. 140–162, 2019. Citado na página 15.

HAMPEL, F. R. The influence curve and its role in robust estimation. *Journal of the american statistical association*, Taylor & Francis, v. 69, n. 346, p. 383–393, 1974. Citado na página 31.

HE, K. et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 770–778. Citado na página 21.

HUA, K.-L. et al. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *OncoTargets and therapy*, Dove Press, v. 8, 2015. Citado na página 15.

ILYAS, A. et al. Adversarial examples are not bugs, they are features. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2019. p. 125–136. Citado na página 26.

JO, T.; NHO, K.; SAYKIN, A. J. Deep learning in alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. *Frontiers in aging neuroscience*, Frontiers, v. 11, p. 220, 2019. Citado na página 15.

KAPISHNIKOV, A. et al. Xrai: Better attributions through regions. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. [S.l.: s.n.], 2019. p. 4948–4957. Citado na página 29.

KERMANY, D. S. et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, Elsevier, v. 172, n. 5, p. 1122–1131, 2018. Citado 2 vezes nas páginas 48 e 49.

KERMANY, D. S.; ZHANG, K.; GOLDBAUM, M. H. Labeled optical coherence tomography (oct) and chest x-ray images for classification. In: . [S.l.: s.n.], 2018. Citado na página 48.

KIM, B. et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: *International conference on machine learning*. [S.l.: s.n.], 2018. p. 2668–2677. Citado 2 vezes nas páginas 32 e 35.

KIM, M. et al. Web applicable computer-aided diagnosis of glaucoma using deep learning. *arXiv preprint arXiv:1812.02405*, 2018. Citado na página 22.

KIM, Y.-G. et al. Deep learning-based four-region lung segmentation in chest radiography for covid-19 diagnosis. *arXiv preprint arXiv:2009.12610*, 2020. Citado na página 63.

KINDERMANS, P.-J. et al. The (un) reliability of saliency methods. Springer, p. 267–280, 2019. Citado 3 vezes nas páginas 38, 41 e 46.

KOH, P. W.; LIANG, P. Understanding black-box predictions via influence functions. *arXiv preprint arXiv:1703.04730*, 2017. Citado 4 vezes nas páginas 31, 32, 34 e 35.

KOLESNIKOV, A.; LAMPERT, C. H. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: SPRINGER. *European conference on computer vision*. [S.l.], 2016. p. 695–711. Citado na página 44.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2012. p. 1097–1105. Citado na página 21.

KUO, C.-C. J. et al. Interpretable convolutional neural networks via feedforward design. *Journal of Visual Communication and Image Representation*, Elsevier, v. 60, p. 346–359, 2019. Citado 2 vezes nas páginas 32 e 37.

LABS, D. *Deal Labs*. [S.l.], 2018. Disponível em: <"https://www.deal.com.br/tecnologia-e-inovacao/laboratorio-iot-ai-machine-learning/">. Acesso em: 04.09.2018. Citado na página 19.

LEOPOLD, H. A. et al. Deep learning for ophthalmology using optical coherence tomography. In: *State of the Art in Neural Networks and their Applications*. [S.l.]: Elsevier, 2021. p. 239–269. Citado na página 15.

LI, X.-H. et al. Quantitative evaluations on saliency methods: An experimental study. *arXiv preprint arXiv:2012.15616*, 2020. Citado 2 vezes nas páginas 38 e 42.

LIN, T.-Y. et al. Microsoft coco: Common objects in context. In: SPRINGER. *European conference on computer vision*. [S.l.], 2014. p. 740–755. Citado 3 vezes nas páginas 43, 47 e 63.

LIPTON, Z. C. The mythos of model interpretability. *Queue*, ACM New York, NY, USA, v. 16, n. 3, p. 31–57, 2018. Citado 2 vezes nas páginas 21 e 23.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2017. p. 4765–4774. Citado 2 vezes nas páginas 32 e 34.

MAHENDRAN, A.; VEDALDI, A. Understanding deep image representations by inverting them. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2015. p. 5188–5196. Citado na página 30.

MISHRA, S.; STURM, B. L.; DIXON, S. Local interpretable model-agnostic explanations for music content analysis. In: *ISMIR*. [S.l.: s.n.], 2017. p. 537–543. Citado na página 34.

MOLNAR, C. Interpretable machine learning. *Lulu. com*, 2019. Citado 2 vezes nas páginas 20 e 24.

NGUYEN, A. et al. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *arXiv preprint arXiv:1605.09304*, 2016. Citado na página 30.

NGUYEN, A.; YOSINSKI, J.; CLUNE, J. Understanding neural networks via feature visualization: A survey. *CoRR*, abs/1904.08939, 2019. Disponível em: <http://arxiv.org/abs/1904.08939>. Citado na página 31.

NGUYEN, A. M.; YOSINSKI, J.; CLUNE, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *CoRR*, abs/1412.1897, 2014. Disponível em: <http://arxiv.org/abs/1412.1897>. Citado 2 vezes nas páginas 27 e 30.

NGUYEN, A. M.; YOSINSKI, J.; CLUNE, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *CoRR*, abs/1412.1897, 2014. Disponível em: <http://arxiv.org/abs/1412.1897>. Citado na página 30.

OLAH, C.; MORDVINTSEV, A.; SCHUBERT, L. Feature visualization. *Distill*, 2017. Https://distill.pub/2017/feature-visualization. Citado 3 vezes nas páginas 29, 30 e 31.

OLAH, C. et al. The building blocks of interpretability. *Distill*, 2018. Https://distill.pub/2018/building-blocks. Citado na página 69.

PALATNIK, I. *RISE tf*. [S.l.], 2022. Disponível em: <"https://github.com/palatos-/RISE_tf">. Acesso em: 02.02.2022. Citado 2 vezes nas páginas 52 e 56.

PAN, S. J.; YANG, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, IEEE, v. 22, n. 10, p. 1345–1359, 2009. Citado na página 19.

PETSIUK, V.; DAS, A.; SAENKO, K. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018. Citado 8 vezes nas páginas 26, 28, 29, 32, 34, 39, 40 e 56.

PROKHORENKOVA, L. et al. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, v. 31, p. 6638–6648, 2018. Citado na página 34.

RESEARCH, G. P. P. . A. *Saliency (PAIR-code)*. [S.l.], 2021. Disponível em: <"https://-/pair-code.github.io/saliency">. Acesso em: 11.4.2022. Citado 3 vezes nas páginas 15, 52 e 56.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should i trust you?"explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2016. p. 1135–1144. Citado 6 vezes nas páginas 22, 28, 29, 32, 34 e 56.

RIBEIRO, M. T. et al. Beyond accuracy: Behavioral testing of nlp models with checklist. In: *Association for Computational Linguistics (ACL)*. [S.l.: s.n.], 2020. Citado na página 22.

RIBEIRO, M. T. C. *LIME Lib*. [S.l.], 2022. Disponível em: <"https://github-.com/marcotcr/lime">. Acesso em: 02.02.2022. Citado 2 vezes nas páginas 52 e 56.

RIEGER, L.; HANSEN, L. K. A simple defense against adversarial attacks on heatmap explanations. *arXiv preprint arXiv:2007.06381*, 2020. Citado na página 41.

ROTH, A. E. *The Shapley value: essays in honor of Lloyd S. Shapley*. [S.l.]: Cambridge University Press, 1988. Citado na página 34.

RUDIN, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, Nature Publishing Group, v. 1, n. 5, p. 206–215, 2019. Citado 2 vezes nas páginas 21 e 26.

RUSSAKOVSKY, O. et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, v. 115, n. 3, p. 211–252, 2015. Citado 2 vezes nas páginas 21 e 50.

SELVARAJU, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2017. p. 618–626. Citado 9 vezes nas páginas 22, 26, 29, 32, 33, 38, 43, 44 e 46.

SENGUPTA, S. et al. Ophthalmic diagnosis using deep learning with fundus images–a critical review. *Artificial Intelligence in Medicine*, Elsevier, v. 102, p. 101758, 2020. Citado na página 15.

SHI, S.; ZHANG, X.; FAN, W. A modified perturbed sampling method for local interpretable model-agnostic explanation. *arXiv preprint arXiv:2002.07434*, 2020. Citado na página 34.

SHRIKUMAR, A.; GREENSIDE, P.; KUNDAJE, A. Learning important features through propagating activation differences. *arXiv preprint arXiv:1704.02685*, 2017. Citado 7 vezes nas páginas 28, 29, 32, 33, 42, 45 e 46.

SIMONYAN, K.; VEDALDI, A.; ZISSERMAN, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. Iclr, 2014. Citado 5 vezes nas páginas 29, 32, 33, 42 e 50.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. Citado 2 vezes nas páginas 49 e 50.

SINGH, A.; SENGUPTA, S.; LAKSHMINARAYANAN, V. Explainable deep learning models in medical image analysis. *Journal of Imaging*, Multidisciplinary Digital Publishing Institute, v. 6, n. 6, p. 52, 2020. Citado 2 vezes nas páginas 14 e 15.

SMILKOV, D. et al. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. Citado 3 vezes nas páginas 32, 33 e 50.

SPRINGENBERG, J. T. et al. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. Citado 2 vezes nas páginas 32 e 33.

SUNDARARAJAN, M.; TALY, A.; YAN, Q. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*, 2017. Citado 6 vezes nas páginas 28, 29, 32, 33, 42 e 51.

SZEGEDY, C. et al. Going deeper with convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2015. p. 1–9. Citado na página 14.

TJOA, E.; GUAN, C. A survey on explainable artificial intelligence (XAI): towards medical XAI. *CoRR*, abs/1907.07374, 2019. Disponível em: <http://arxiv.org/abs/1907-.07374>. Citado na página 24.

WANG, X. et al. Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: *IEEE CVPR*. [S.l.: s.n.], 2017. v. 7. Citado na página 68.

WU, J. et al. Expert identification of visual primitives used by cnns during mammogram classification. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. *Medical Imaging 2018: Computer-Aided Diagnosis*. [S.l.], 2018. v. 10575, p. 105752T. Citado na página 22.

WU, Z.; SHEN, C.; HENGEL, A. V. D. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, Elsevier, v. 90, p. 119–133, 2019. Citado na página 26.

XIAO, H.; RASUL, K.; VOLLGRAF, R. *Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms*. 2017. Citado na página 67.

YANG, M.; KIM, B. Benchmarking attribution methods with relative feature importance. *arXiv preprint arXiv:1907.09701*, 2019. Citado 2 vezes nas páginas 44 e 45.

YEH, C.-K. et al. On completeness-aware concept-based explanations in deep neural networks. *arXiv preprint arXiv:1910.07969*, 2019. Citado 2 vezes nas páginas 32 e 36.

YOUNG, T. et al. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, IEEE, v. 13, n. 3, p. 55–75, 2018. Citado na página 14.

ZEILER, M. D.; FERGUS, R. Visualizing and understanding convolutional networks. In: SPRINGER. *European conference on computer vision*. [S.l.], 2014. p. 818–833. Citado na página 28.

ZHANG, A. et al. *Dive into Deep Learning*. [S.l.: s.n.], 2020. `https://d2l.ai`. Citado na página 52.

ZHANG, J. et al. Top-down neural attention by excitation backprop. In: SPRINGER. *European Conference on Computer Vision*. [S.l.], 2016. p. 543–559. Citado 5 vezes nas páginas 36, 43, 55, 63 e 64.

ZHANG, Q. et al. Interpreting cnn knowledge via an explanatory graph. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2018. v. 32, n. 1. Citado 2 vezes nas páginas 32 e 36.

ZHANG, Q.; WU, Y. N.; ZHU, S.-C. Interpretable convolutional neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2018. p. 8827–8836. Citado 2 vezes nas páginas 32 e 37.

ZHANG, Q. et al. Interpreting cnns via decision trees. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2019. Citado 2 vezes nas páginas 32 e 36.

ZHOU, B. et al. Learning deep features for discriminative localization. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 2921–2929. Citado 2 vezes nas páginas 32 e 33.

A Arquitetura das Redes Neurais

A seguir estão apresentadas as arquiteturas neurais utilizadas no estudo de caso, nos Capítulos 4 e 5. Nas Figuras 27, 28, 29 e 30 estão apresentadas as arquiteturas utilizadas para o estudo de caso. Para o conjunto de dados OCT a resolução utilizada foi 224 × 224. Para o conjunto de dados Raio-X a resolução utilizada foi 150 × 150.

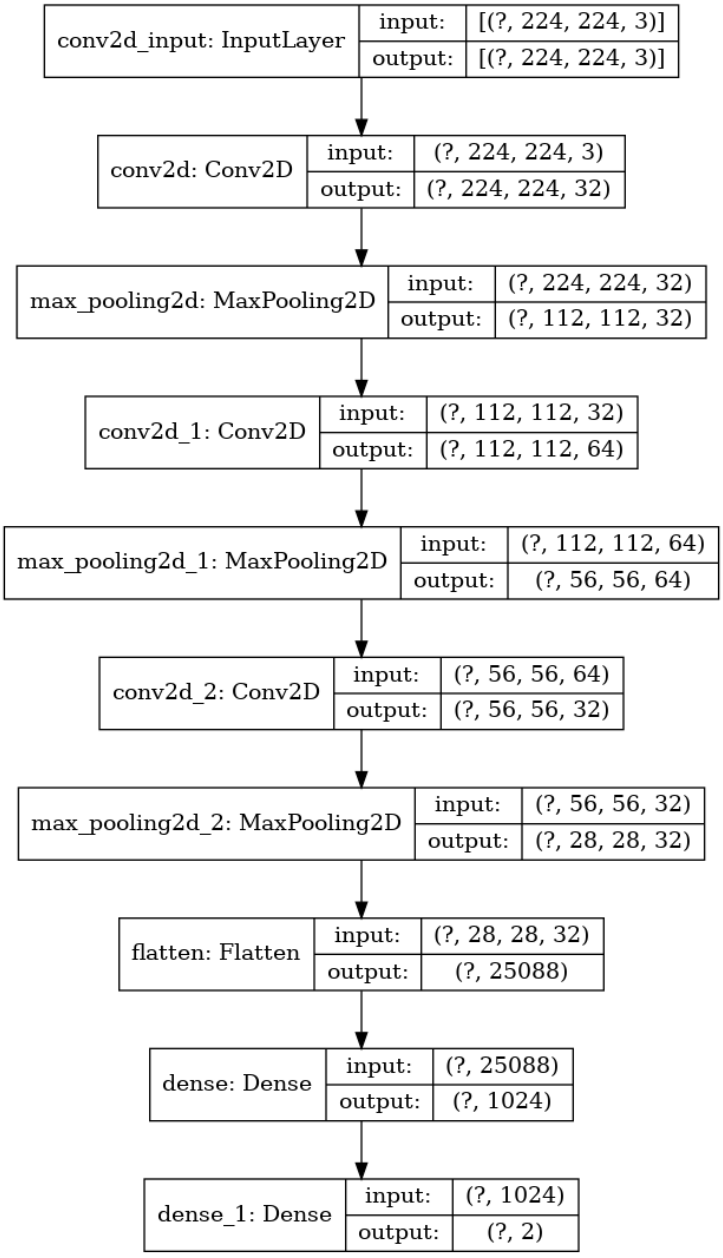


Figura 27 – Arquitetura Customizada com duas saídas e 224 × 224 de resolução, apropriada para o conjunto de dados de Raio-X.

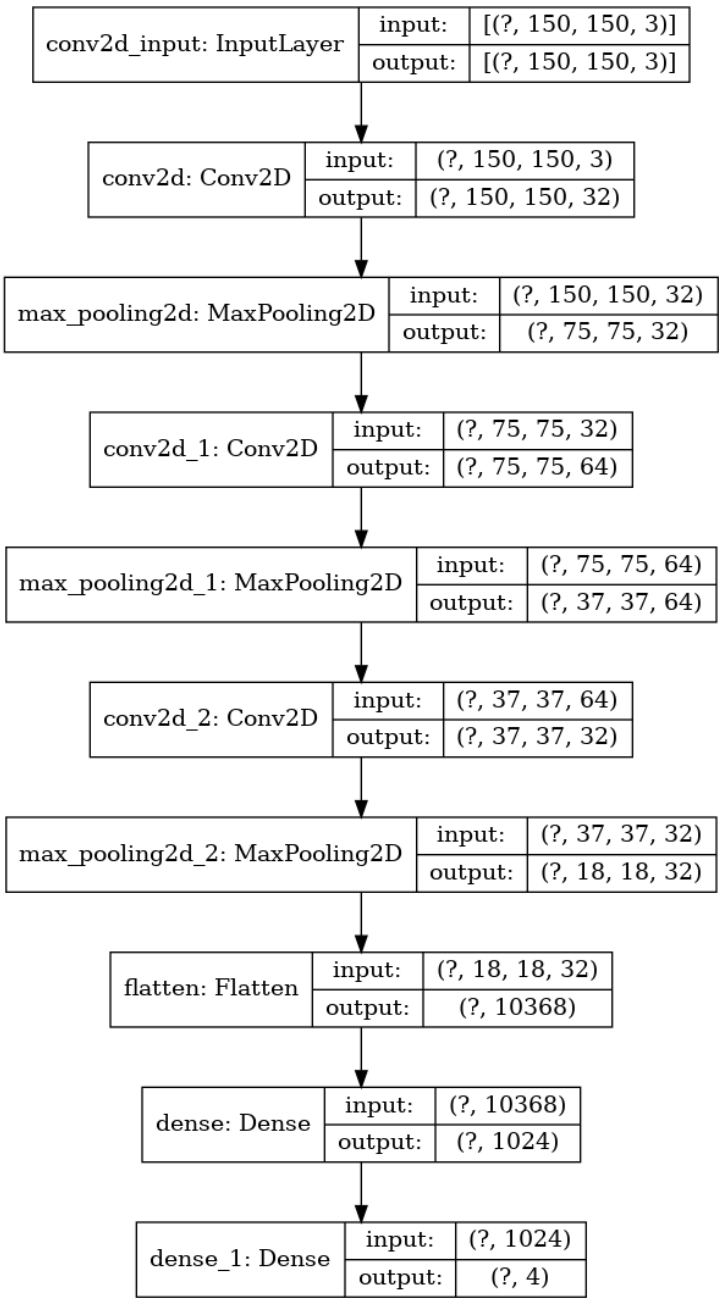


Figura 28 – Arquitetura Customizada com quatro saídas e 150 × 150 de resolução, apropriada para o conjunto de dados de OCT.

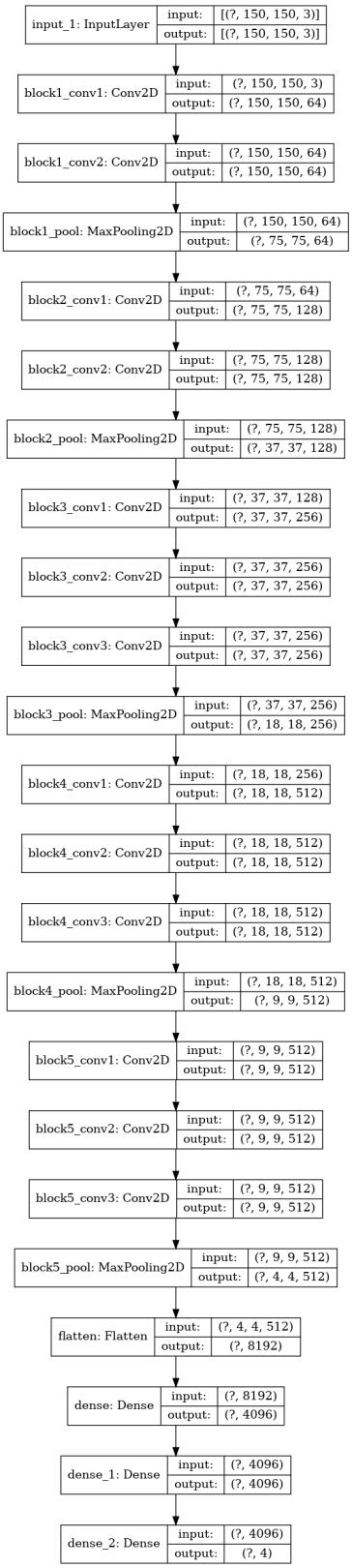


Figura 29 – Arquitetura VGG16 com quatro saídas e 150 × 150 de resolução, apropriada para o conjunto de dados de OCT.

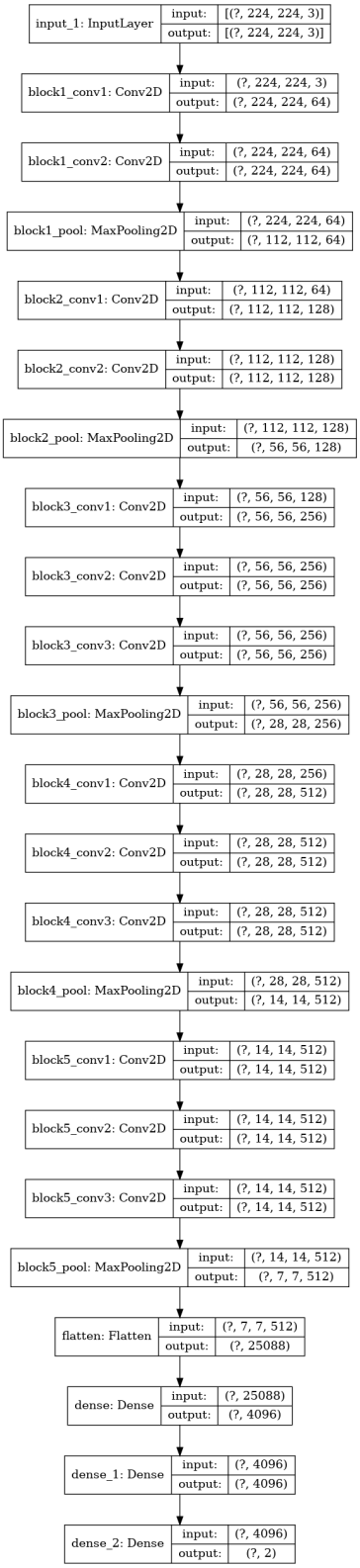


Figura 30 – Arquitetura VGG16 com duas saídas e 224 × 224 de resolução, apropriada para o conjunto de dados de Raio-X.

B Históricos de Treinamentos

A seguir estão apresentados os históricos de treinamento para os modelos utilizados. Para cada treinamento estão apresentadas as curvas de acurácia e perda (*Loss*) para os conjuntos de treinamento e validação.

B.1 Arquitetura Customizada - Base OCT

Nas Figuras 31, 32, 33, 34, 35 e 36 estão apresentadas as curvas de treinamento para a rede customizada e conjunto OCT.

Para o treinamento com os rótulos originais (Figuras 31 e 32), os valores de acurácia e perda sinalizam um treinamento saudável, com altas acurácias e baixas perdas. Observa-se uma rápida convergência e pouca variação nas métricas após a décima época.

Por ser uma arquitetura com baixa capacidade de ajuste, o treinamento com os rótulos aleatórios foi dividido em duas etapas, primeiro foi treinado com o conjunto completo de dados e depois com um conjunto reduzido. O conjunto de validação possui os rótulos originais. Nas Figuras 33, 34, 35 e 36 é perceptível como a acurácia relacionada com o conjunto original permaneceu no mesmo patamar enquanto a perda aumentou vertiginosamente, indicando um modelo aleatório, despropositado com o conjunto original de rótulos. Observou-se uma convergência nas métricas relacionadas com o conjunto de treino.

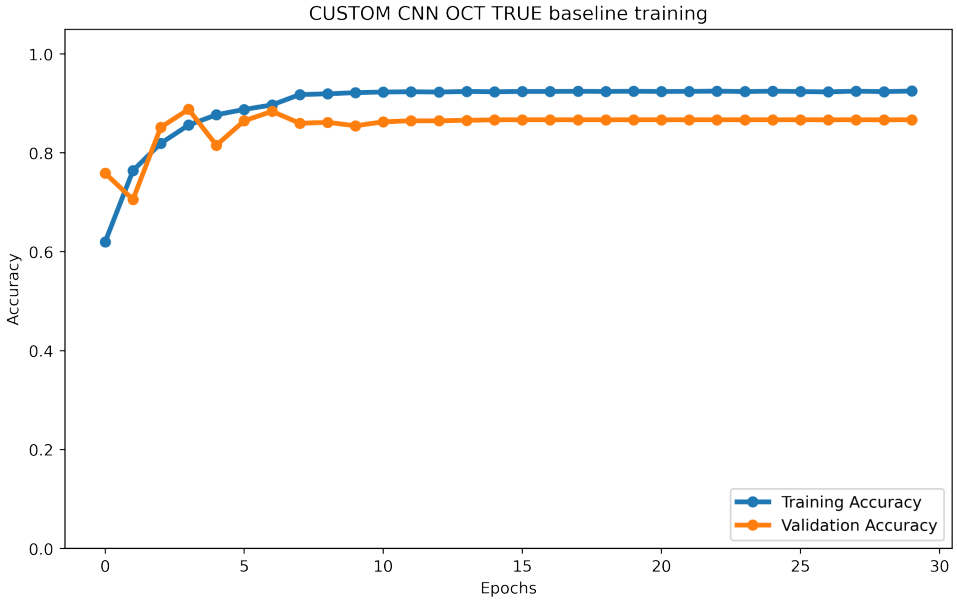


Figura 31 – Acurácia nos conjuntos de treino e validação para o treinamento da rede customizada, com o conjunto de dados OCT.

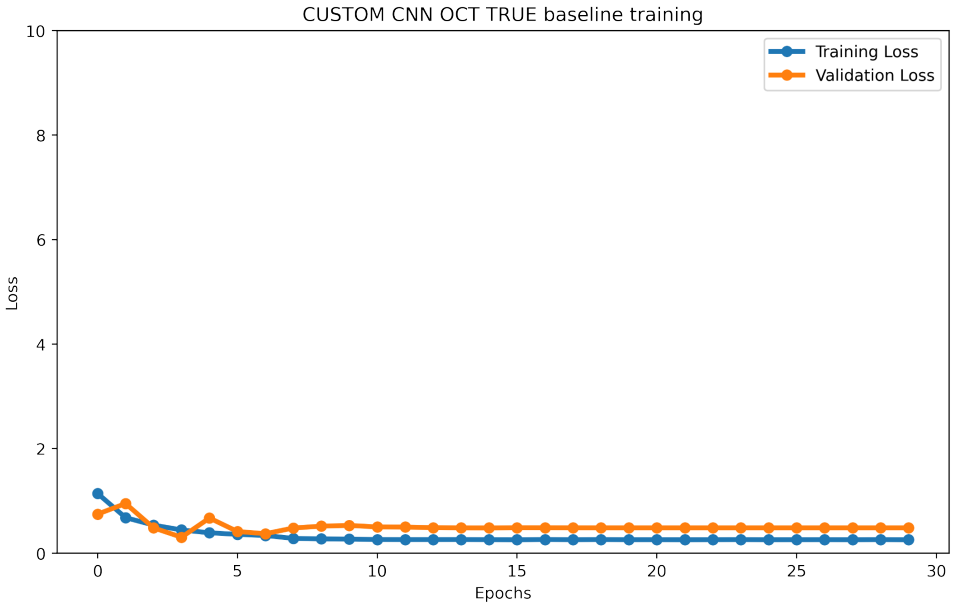


Figura 32 – Perda (*Loss*) nos conjuntos de treino e validação para o treinamento da rede customizada, com o conjunto de dados OCT.

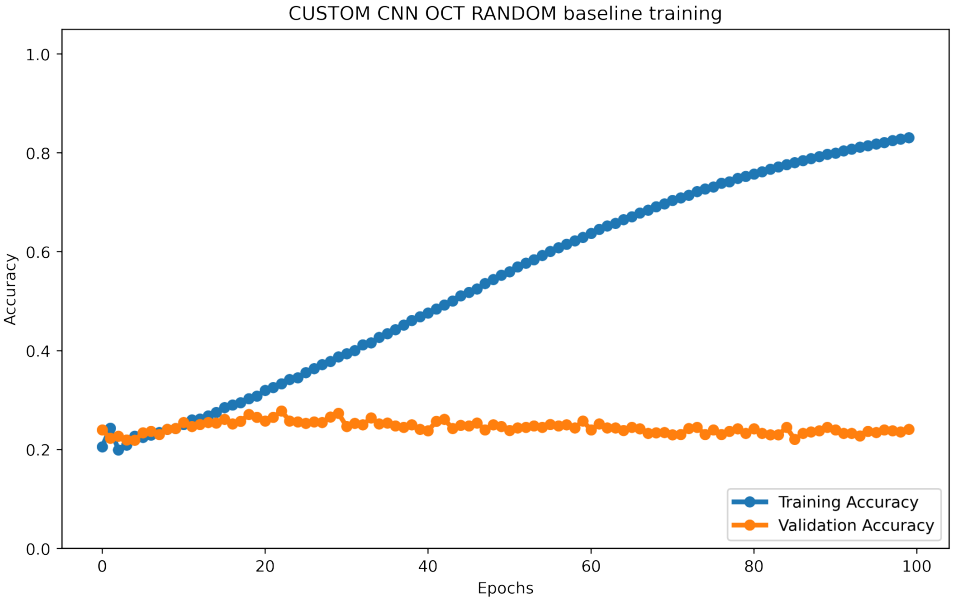


Figura 33 – Acurácia nos conjuntos de treino e validação para o treinamento da rede customizada, com o conjunto de dados OCT. Os rótulos do conjunto de treino são os aleatórios. Os rótulos do conjunto de validação são originais.

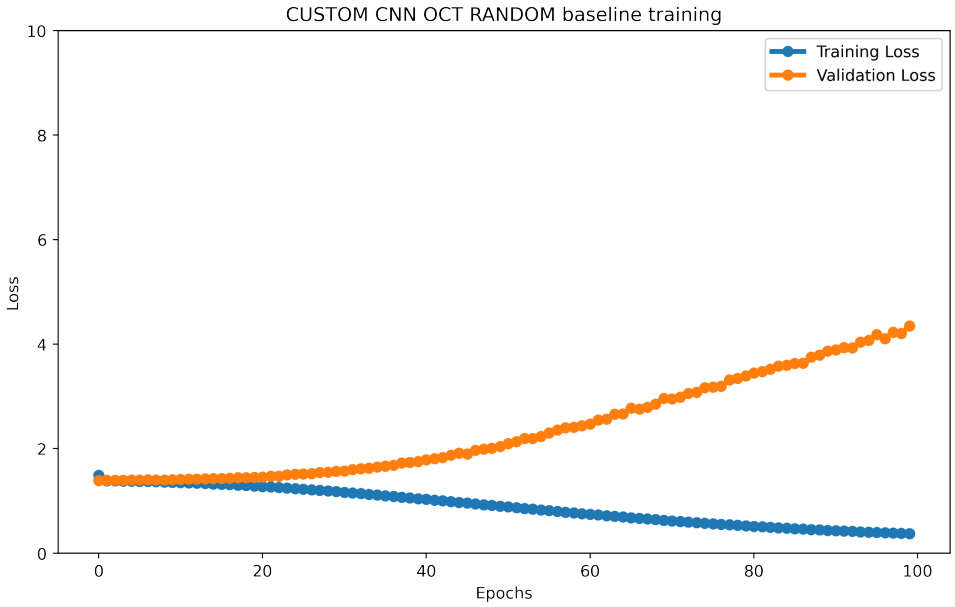


Figura 34 – Perda (*Loss*) nos conjuntos de treino e validação para o treinamento da rede customizada, com o conjunto de dados OCT. Os rótulos do conjunto de treino são os aleatórios. Os rótulos do conjunto de validação são originais.

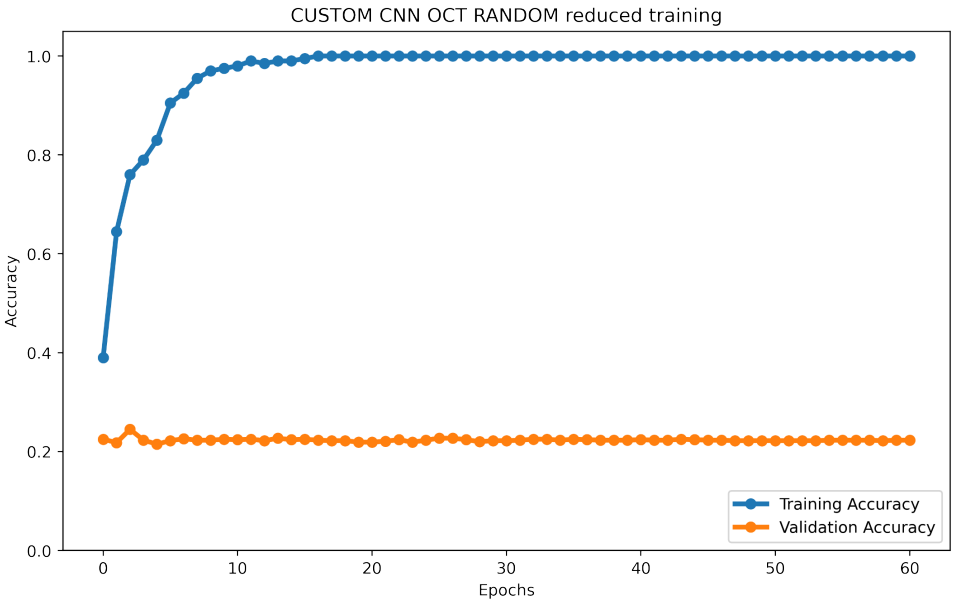


Figura 35 – Acurácia nos conjuntos de treino e validação para o treinamento da rede customizada, com o conjunto de dados OCT. Ajuste final com um conjunto de dados reduzido para atingir convergência. Os rótulos do conjunto de treino são os aleatórios. Os rótulos do conjunto de validação são originais.

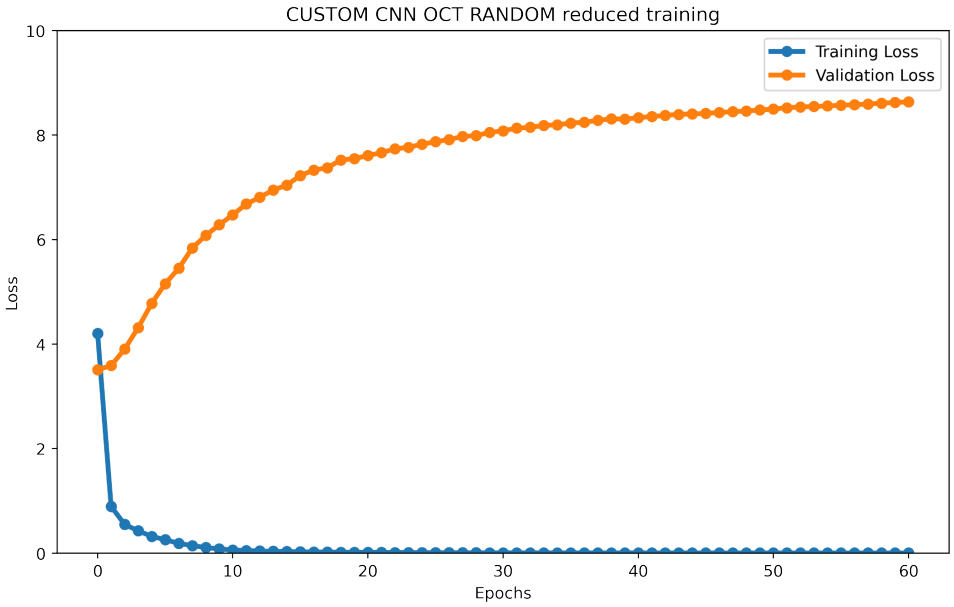


Figura 36 – Perda (*Loss*) nos conjuntos de treino e validação para o treinamento da rede customizada., com o conjunto de dados OCT. Ajuste final com um conjunto de dados reduzido para atingir convergência. Os rótulos do conjunto de treino são os aleatórios. Os rótulos do conjunto de validação são originais.

B.2 Arquitetura Customizada - Base Raio-X

Nas Figuras 37, 38, 39, 40, 41 e 42 estão apresentadas as curvas de treinamento para a rede customizada e conjunto Raio-X.

Para o treinamento com os rótulos originais (Figuras 37, 38), os valores de acurácia e perda sinalizam um treinamento saudável, com altas acurácias e baixas perdas. Observa-se uma rápida convergência e uma melhora pequena, mas constante, nas métricas ao longo do treinamento.

Por ser uma arquitetura com baixa capacidade de ajuste, o treinamento com os rótulos aleatórios foi dividido em duas etapas, primeiro foi treinado com o conjunto completo de dados e depois com um conjunto reduzido. O conjunto de validação possui os rótulos originais. Nas Figuras 39, 40, 41 e 42 é perceptível como a acurácia relacionada com o conjunto original permaneceu no mesmo patamar enquanto a perda aumentou vertiginosamente, indicando um modelo aleatório, despropositado com o conjunto original de rótulos. Observou-se uma convergência nas métricas relacionadas com o conjunto de treino.

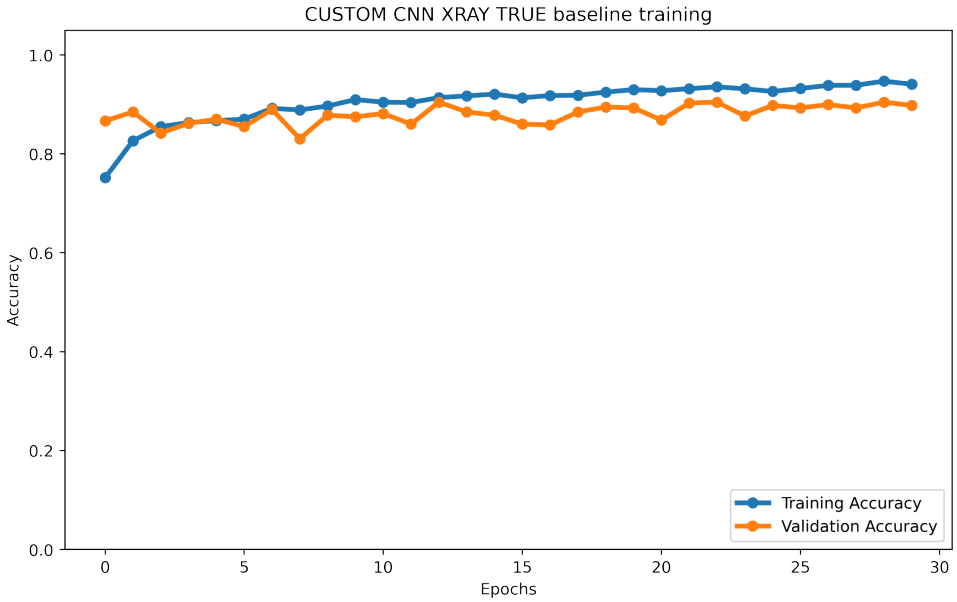


Figura 37 – Acurácia nos conjuntos de treino e validação para o treinamento da rede customizada, com o conjunto de dados XRAY.

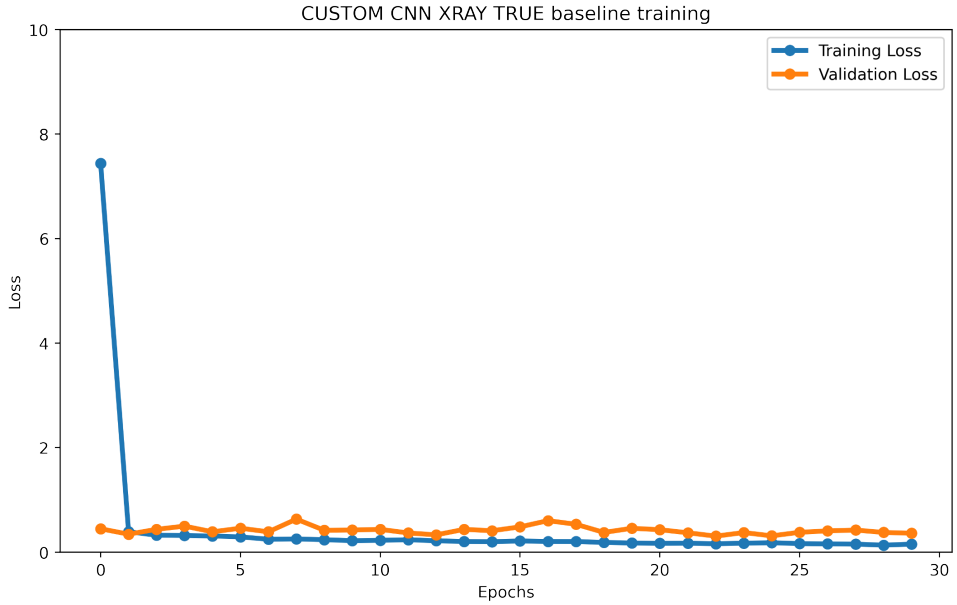


Figura 38 – Perda (*Loss*) nos conjuntos de treino e validação para o treinamento da rede customizada, com o conjunto de dados XRAY.

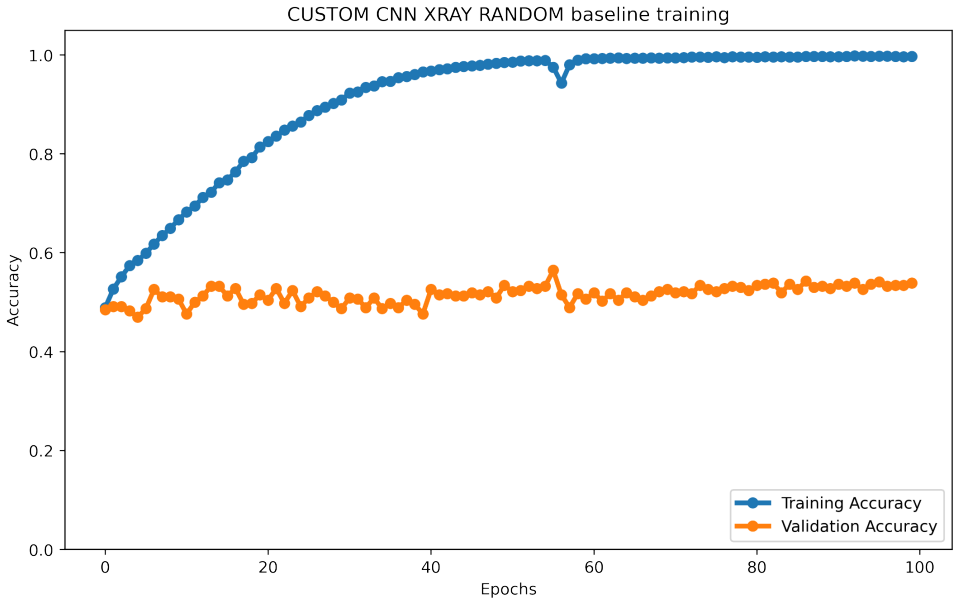


Figura 39 – Acurácia nos conjuntos de treino e validação para o treinamento da rede customizada, com o conjunto de dados Raio-X. Os rótulos do conjunto de treino são os aleatórios. Os rótulos do conjunto de validação são originais.

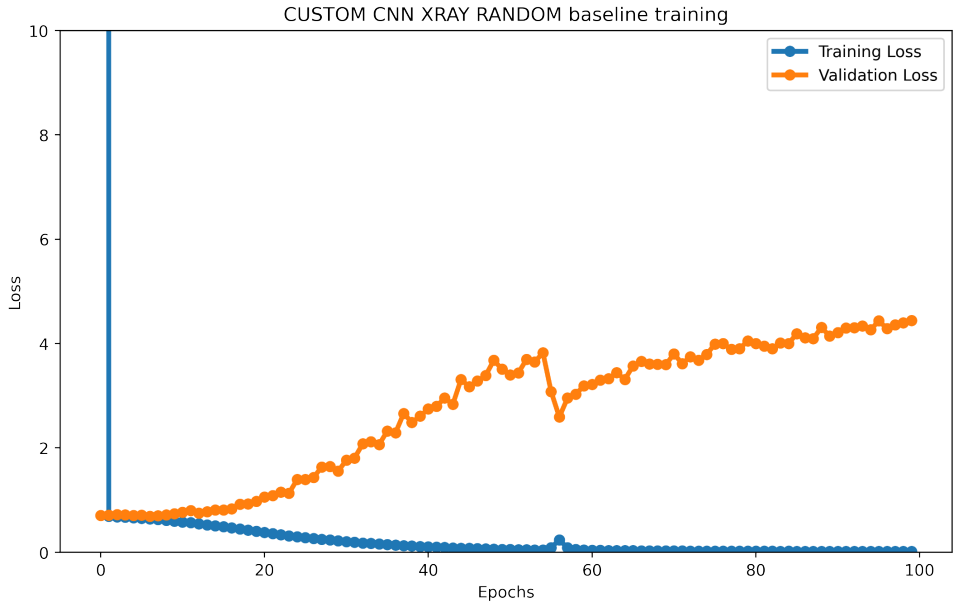


Figura 40 – Perda (*Loss*) nos conjuntos de treino e validação para o treinamento da rede customizada, com o conjunto de dados Raio-X. Os rótulos do conjunto de treino são os aleatórios. Os rótulos do conjunto de validação são originais.

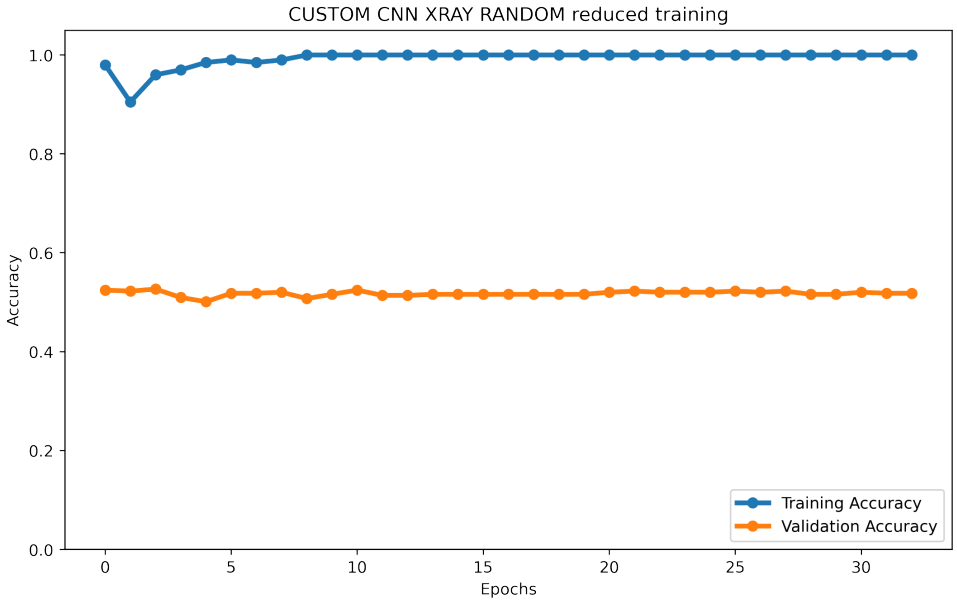


Figura 41 – Acurácia nos conjuntos de treino e validação para o treinamento da rede customizada, com o conjunto de dados Raio-X. Ajuste final com um conjunto de dados reduzido para atingir convergência. Os rótulos do conjunto de treino são os aleatórios. Os rótulos do conjunto de validação são originais.

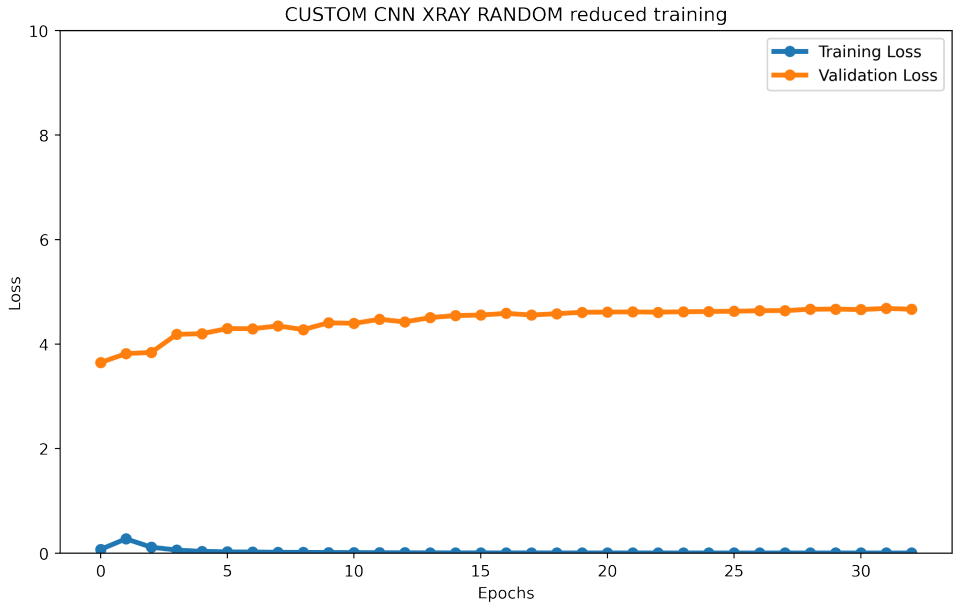


Figura 42 – Perda (*Loss*) nos conjuntos de treino e validação para o treinamento da rede customizada, com o conjunto de dados OCT. Ajuste final com um conjunto de dados reduzido para atingir convergência. Os rótulos do conjunto de treino são os aleatórios. Os rótulos do conjunto de validação são originais.

B.3 Arquitetura VGG - Base OCT

Nas Figuras 43, 44, 45, 46, 47 e 48 estão apresentadas as curvas de treinamento para a rede VGG e conjunto de dados OCT.

Para a arquitetura VGG o treinamento com os rótulos originais foi dividido em duas etapas: transferência de conhecimento e ajuste fino. Por ser uma arquitetura popular na literatura, a arquitetura VGG permite a utilização de técnicas facilitadoras que possibilitam uma rápida convergência. Para a fase de transferência de conhecimento (Figuras 43 e 44), os valores de acurácia e perda sinalizam um treinamento saudável, com altas acurácias e baixas perdas. Observa-se que os pesos herdados produzem boas métricas já nas primeiras épocas de treinamento. Para a fase de ajuste fino (Figuras 45 e 46), já na primeira época, as métricas indicam um modelo muito próximo do ideal.

O treinamento de um modelo aleatório visando comparar explicações foi realizado em uma só etapa. Por ter uma maior capacidade, a arquitetura VGG consegue atingir uma convergência mesmo com rótulos aleatórios. As métricas de treino são aferidas com o conjunto aleatório e as métricas de validação são aferidas em um conjunto com rótulos originais. Nas Figuras 47 e 48 é perceptível como a acurácia relacionada com o conjunto original permaneceu no mesmo patamar enquanto a perda aumentou vertiginosamente, indicando um modelo aleatório, despropositado com o conjunto original de rótulos. Os modelos apresentaram convergência nas métricas relacionadas ao conjunto aleatório.

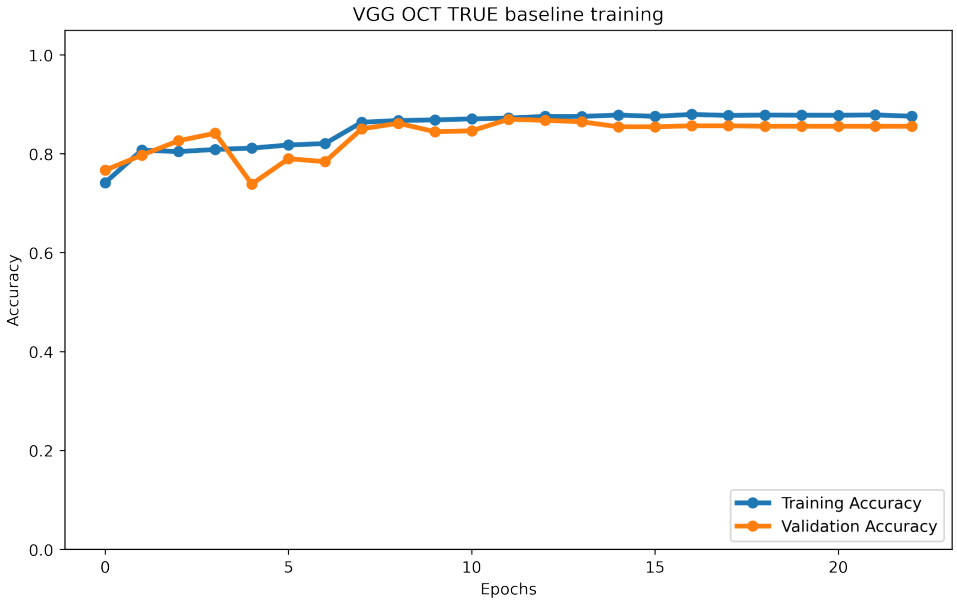


Figura 43 – Acurácia nos conjuntos de treino e validação para o treinamento da rede VGG, com o conjunto de dados Raio-X.

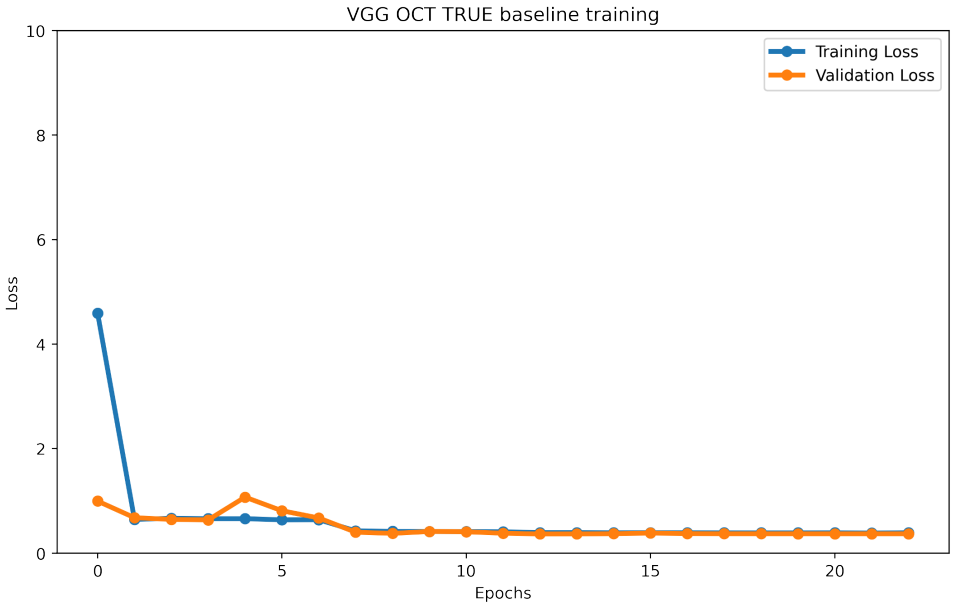


Figura 44 – Perda (*Loss*) nos conjuntos de treino e validação para o treinamento da rede VGG, com o conjunto de dados OCT.

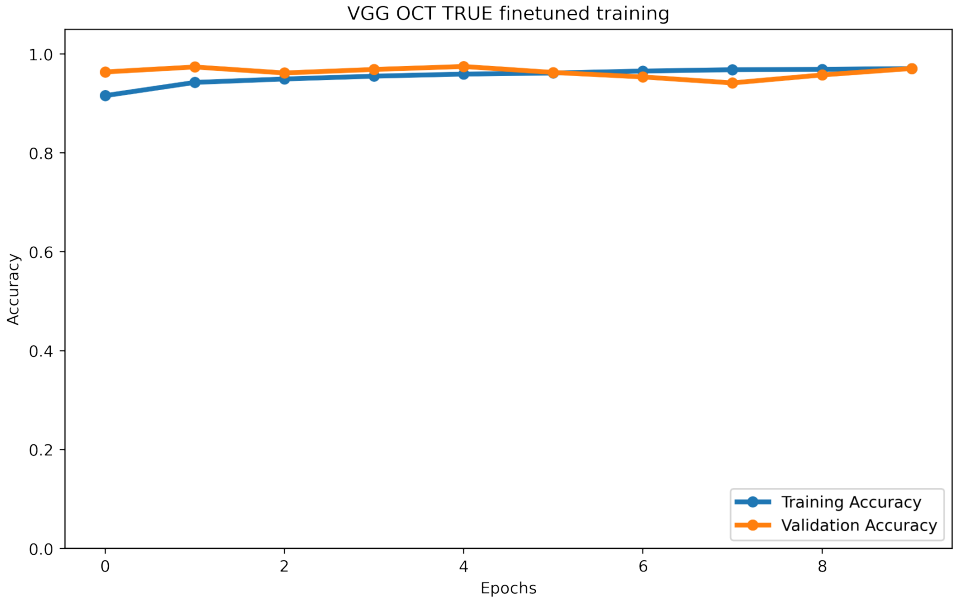


Figura 45 – Acurácia nos conjuntos de treino e validação para o treinamento da rede VGG, com o conjunto de dados OCT com rótulos originais. Etapa de ajuste fino.

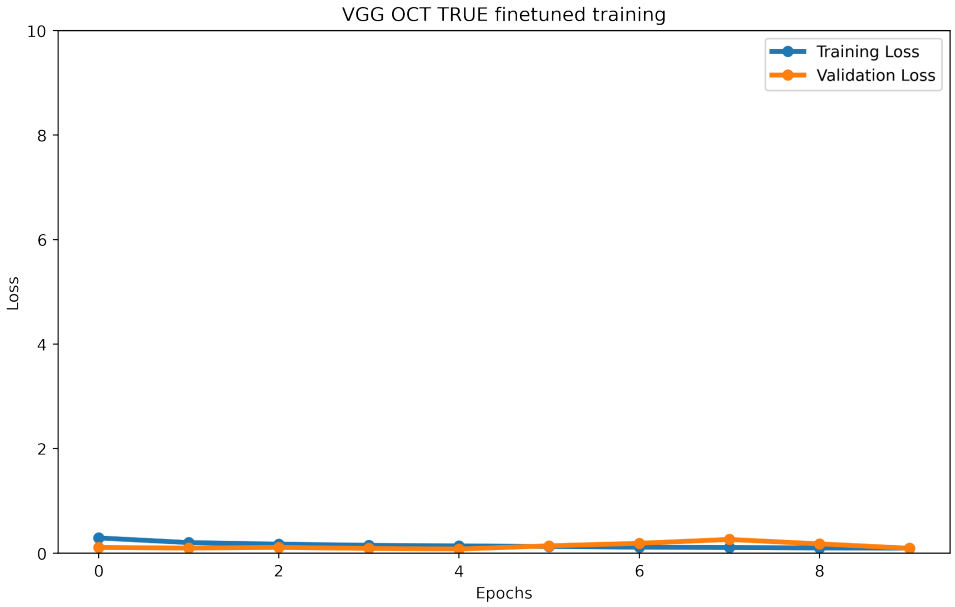


Figura 46 – Perda (*Loss*) nos conjuntos de treino e validação para o treinamento da rede VGG, com o conjunto de dados OCT com rótulos original. Etapa de ajuste fino.

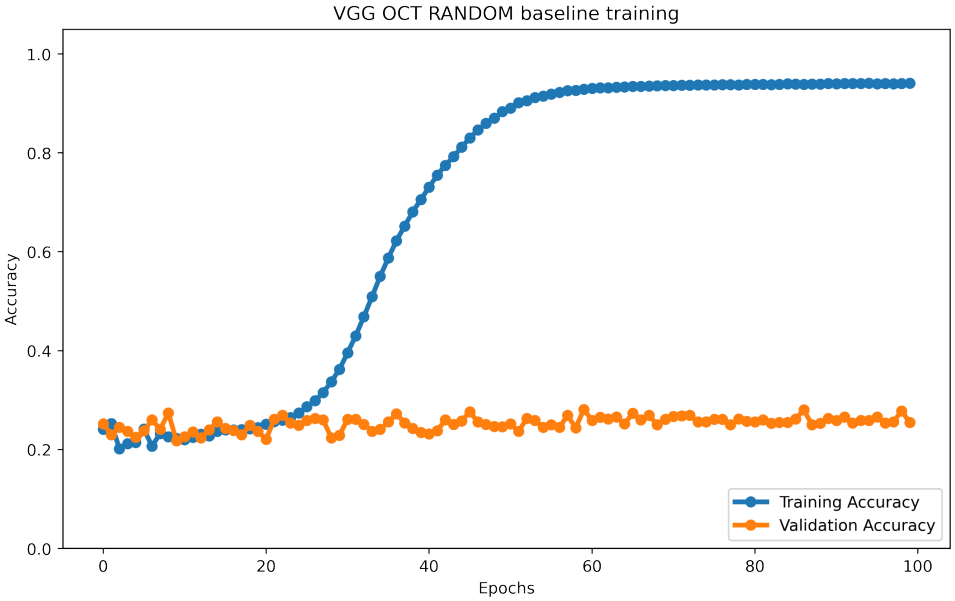


Figura 47 – Acurácia nos conjuntos de treino e validação para o treinamento da rede VGG, com o conjunto de dados OCT com rótulos aleatorizados, para o experimento Randomização dos Rótulos. Os rótulos do conjunto de treino são os aleatórios. Os rótulos do conjunto de validação são originais.

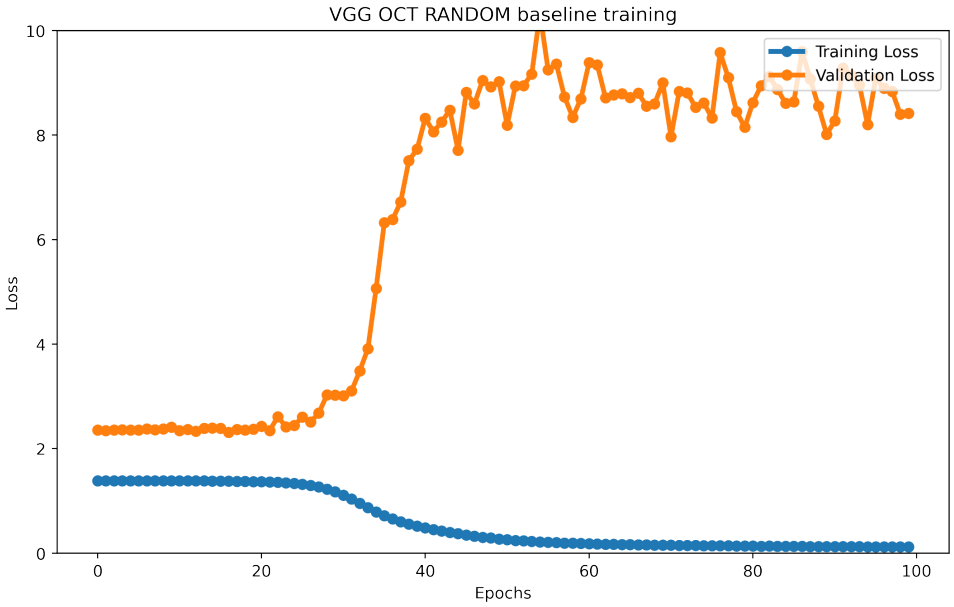


Figura 48 – Perda (*Loss*) nos conjuntos de treino e validação para o treinamento da rede VGG, com o conjunto de dados OCT com rótulos aleatorizados, para o experimento Randomização dos Rótulos. Os rótulos do conjunto de treino são os aleatórios. Os rótulos do conjunto de validação são originais.

B.4 Arquitetura VGG - Base Raio-X

Nas Figuras 49, 50, 51, 52, 53 e 54 estão apresentadas as curvas de treinamento para a rede VGG e conjunto de dados Raio-X.

Para a arquitetura VGG o treinamento com os rótulos originais foi dividido em duas etapas: transferência de conhecimento e ajuste fino. Por ser uma arquitetura popular na literatura, a arquitetura VGG permite a utilização de técnicas facilitaras que permitem uma convergência rápida. Para a fase de transferência de conhecimento (Figuras 49 e 50), os valores de acurácia e perda sinalizam um treinamento saudável, com altas acurácias e baixas perdas. Observa-se que os pesos herdados produzem boas métricas nas primeiras dez épocas de treinamento. Para a fase de ajuste fino (Figuras 51 e 52), as métricas indicam um modelo muito próximo do ideal.

O treinamento de um modelo aleatório visando comparar explicações foi realizado em uma só etapa. Por ter uma maior capacidade, a arquitetura VGG consegue atingir uma convergência mesmo com rótulos aleatórios. As métricas de treino são aferidas com o conjunto aleatório e as métricas de validação são aferidas em um conjunto com rótulos originais. Nas Figuras 47 e 48 é perceptível como a acurácia relacionada com o conjunto original permaneceu no mesmo patamar enquanto a perda aumentou vertiginosamente, indicando um modelo aleatório, despropositado com o conjunto original de rótulos. Os modelos apresentaram convergência nas métricas relacionadas ao conjunto aleatório.

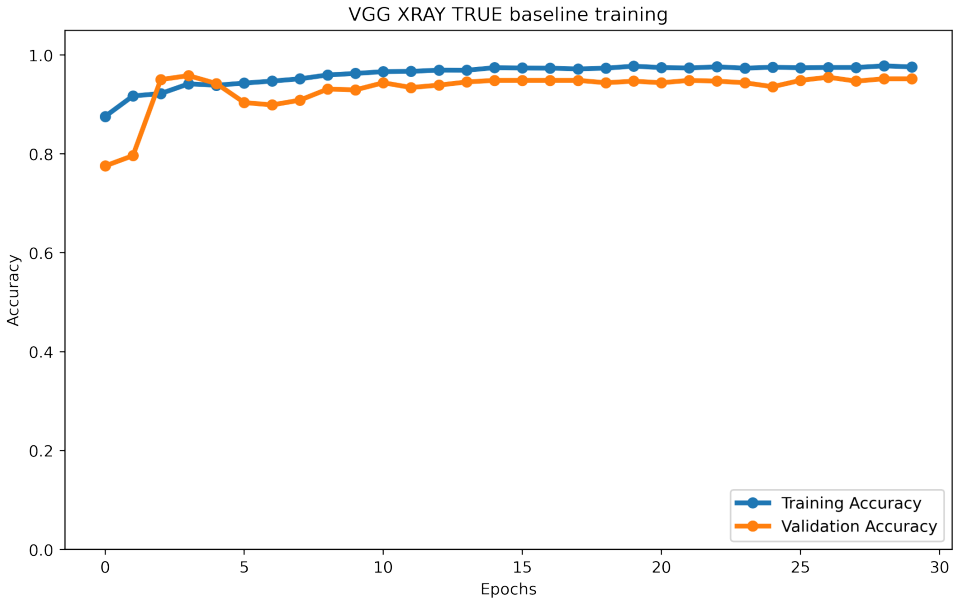


Figura 49 – Acurácia nos conjuntos de treino e validação para o treinamento da rede VGG, com o conjunto de dados Raio-X.

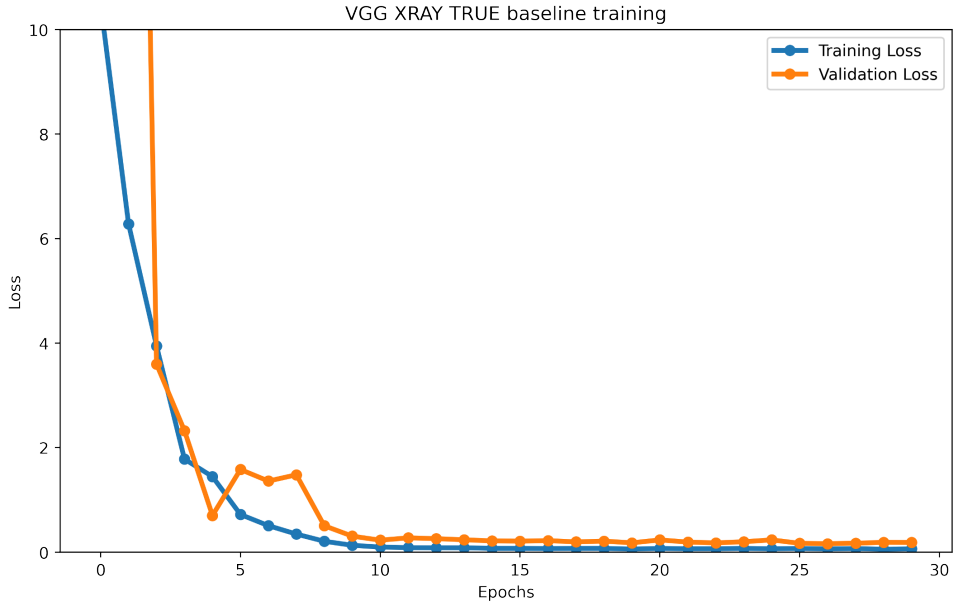


Figura 50 – Perda (*Loss*) nos conjuntos de treino e validação para o treinamento da rede VGG, com o conjunto de dados Raio-X.

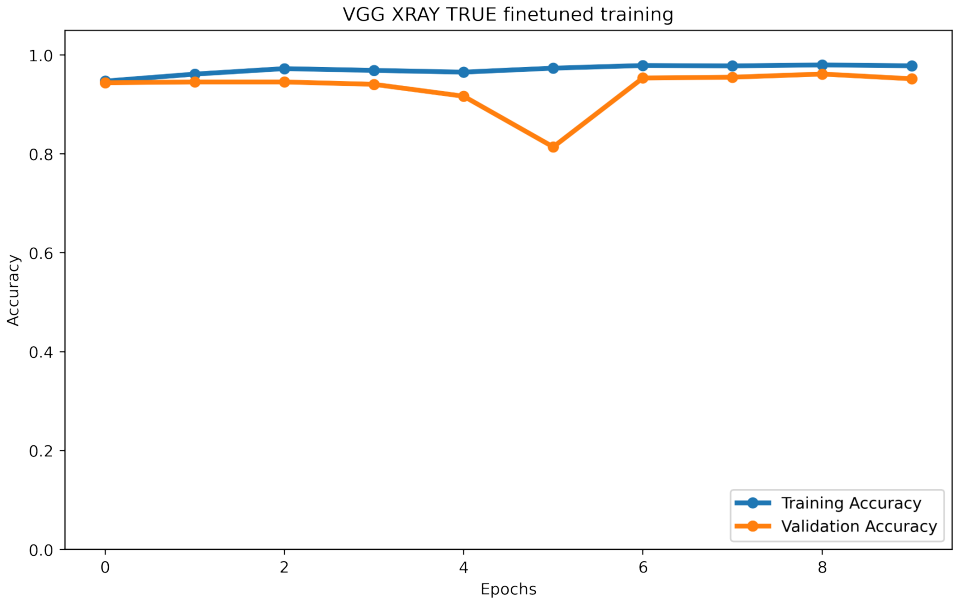


Figura 51 – Perda (*Loss*) nos conjuntos de treino e validação para o treinamento da rede VGG, com o conjunto de dados Raio-X com rótulos rótulos original. Etapa de ajuste fino.

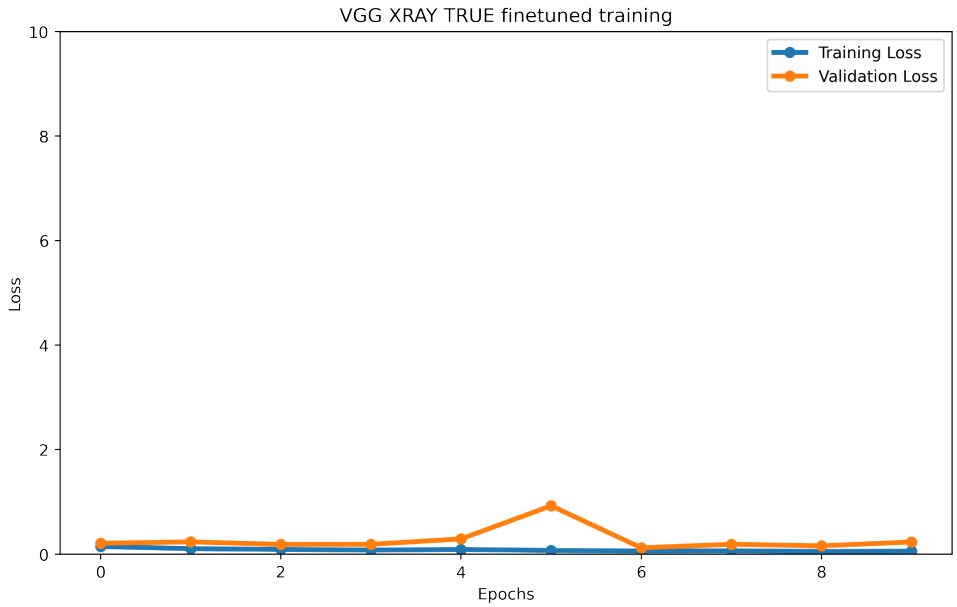


Figura 52 – Perda (*Loss*) nos conjuntos de treino e validação para o treinamento da rede VGG, com o conjunto de dados Raio-X com rótulos rótulos original. Etapa de ajuste fino.

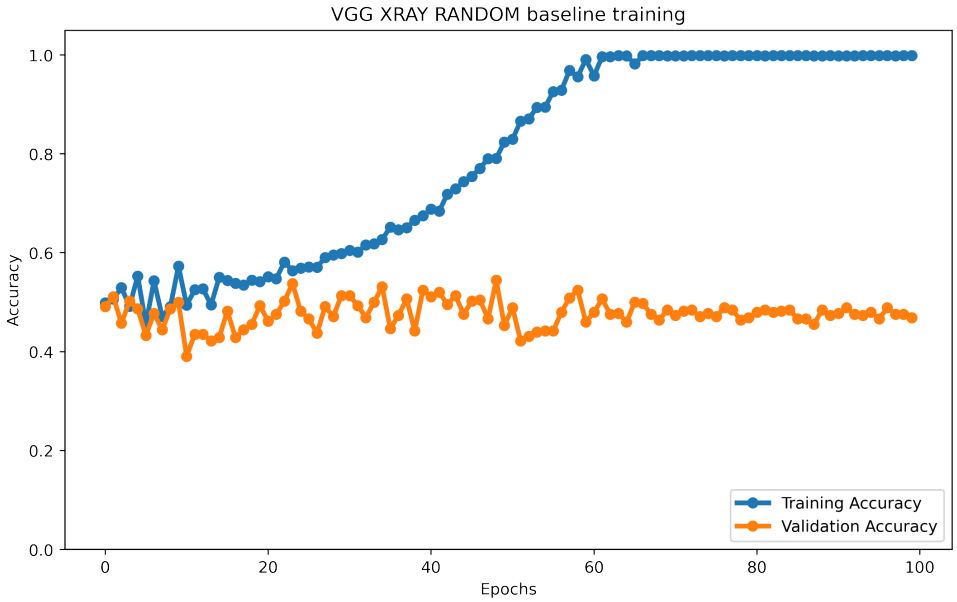


Figura 53 – Acurácia nos conjuntos de treino e validação para o treinamento da rede VGG, com o conjunto de dados Raio-X com rótulos aleatorizados, para o experimento Randomização dos Rótulos. Os rótulos do conjunto de treino são os aleatórios. Os rótulos do conjunto de validação são originais.

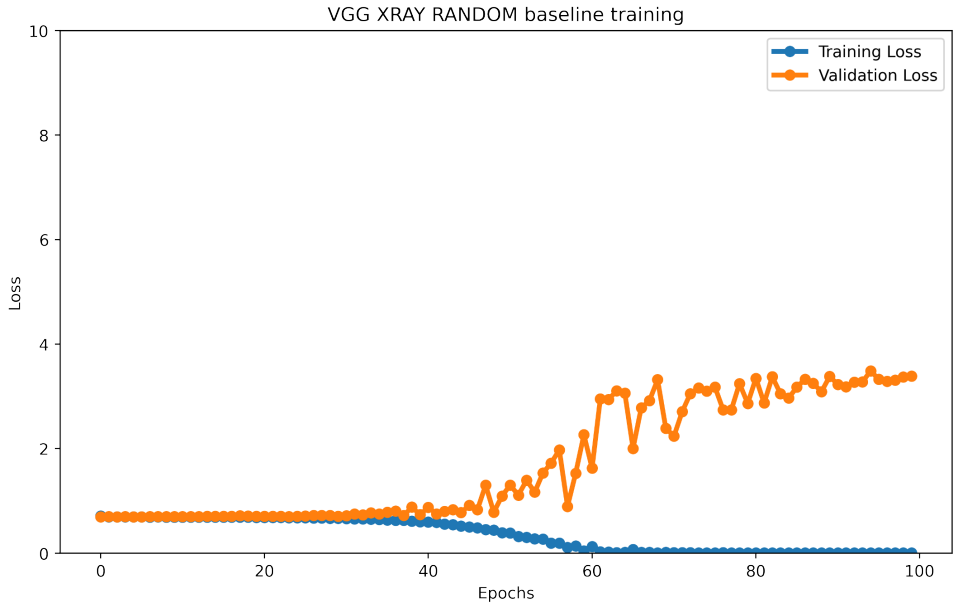


Figura 54 – Perda (*Loss*) nos conjuntos de treino e validação para o treinamento da rede VGG, com o conjunto de dados Raio-X com rótulos aleatorizados, para o experimento Randomização dos Rótulos. Os rótulos do conjunto de treino são os aleatórios. Os rótulos do conjunto de validação são originais.

C Randomização dos Rótulos - Distribuição

Nas Figuras 55, 56, 57 e 58 estão apresentadas as distribuições para a avaliação Randomização dos Rótulos. Em cada resultado tem-se no eixo X o valor da correlação entre duas imagens para as diferentes técnicas, listadas ao longo do eixo Y. As técnicas foram avaliadas em dois conjuntos de dados e com duas arquiteturas neurais diferentes, especificados no título de cada resultado. Na Seção Randomização dos Rótulos as correlações foram apresentadas de maneira agregada em intervalos de confiança.

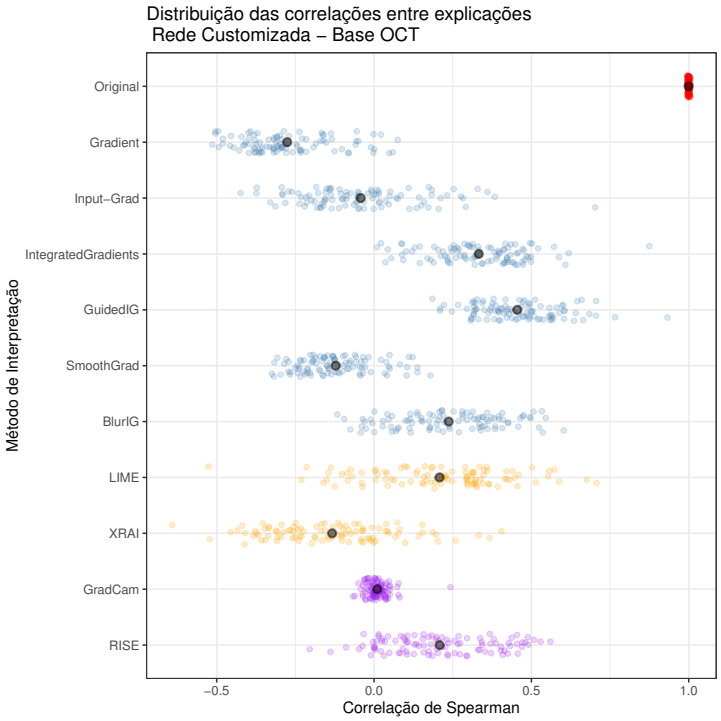


Figura 55 – Resultados da avaliação randomização dos rótulos com adaptações para o conjunto de dados Raio-X e rede treinada com arquitetura customizada.

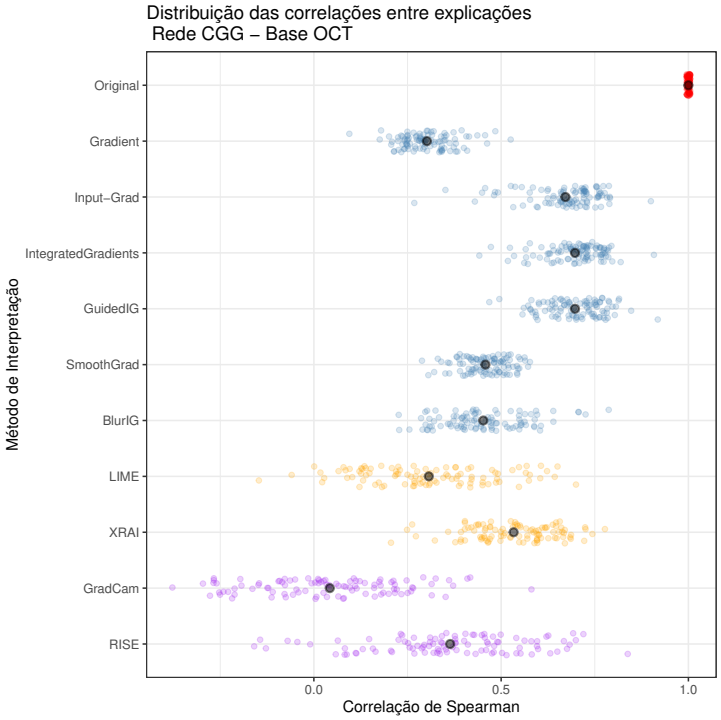


Figura 56 – Resultados da avaliação randomização dos rótulos com adaptações para o conjunto de dados Raio-X e rede treinada com arquitetura customizada.

Figura 57 – Resultados da avaliação randomização dos rótulos com adaptações para o conjunto de dados Raio-X e rede treinada com arquitetura customizada.

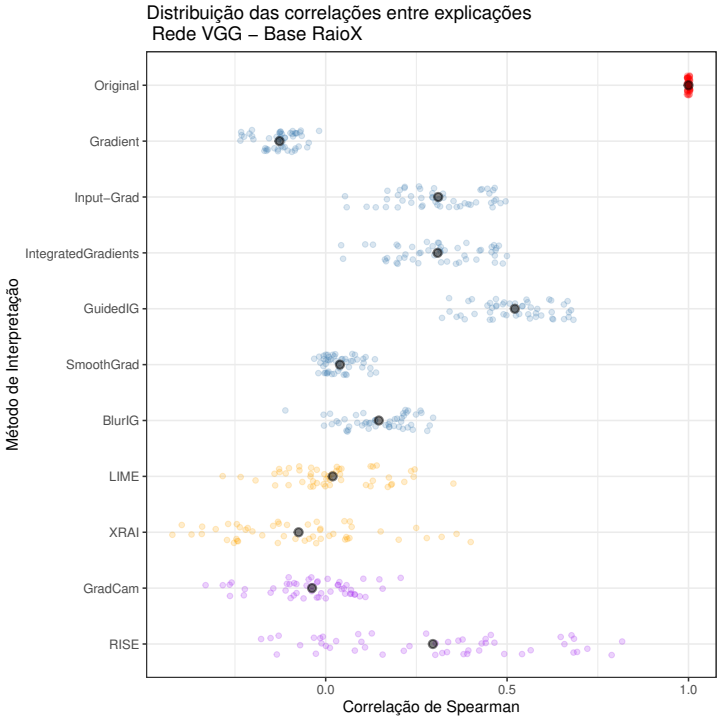


Figura 58 – Resultados da avaliação randomização dos rótulos com adaptações para o conjunto de dados Raio-X e rede treinada com arquitetura customizada.

D Técnicas de Atribuição - Exemplos

Nas Figuras a seguir estão apresentados exemplos de explicações utilizados nas avaliações para os modelos e técnicas apresentados no Capítulo Materiais e Métodos. Mais informações sobre as técnicas estão descritas em Técnicas.

Nas Figuras 59 e 60 estão exemplificadas explicações para o modelo customizado, treinado com o conjunto OCT com rótulos originais e aleatórios. Percebe-se como as explicações são variadas e podem ser influenciadas por artefatos provenientes de um aumento de dados prévio, como no caso dos exemplos da técnica LIME.

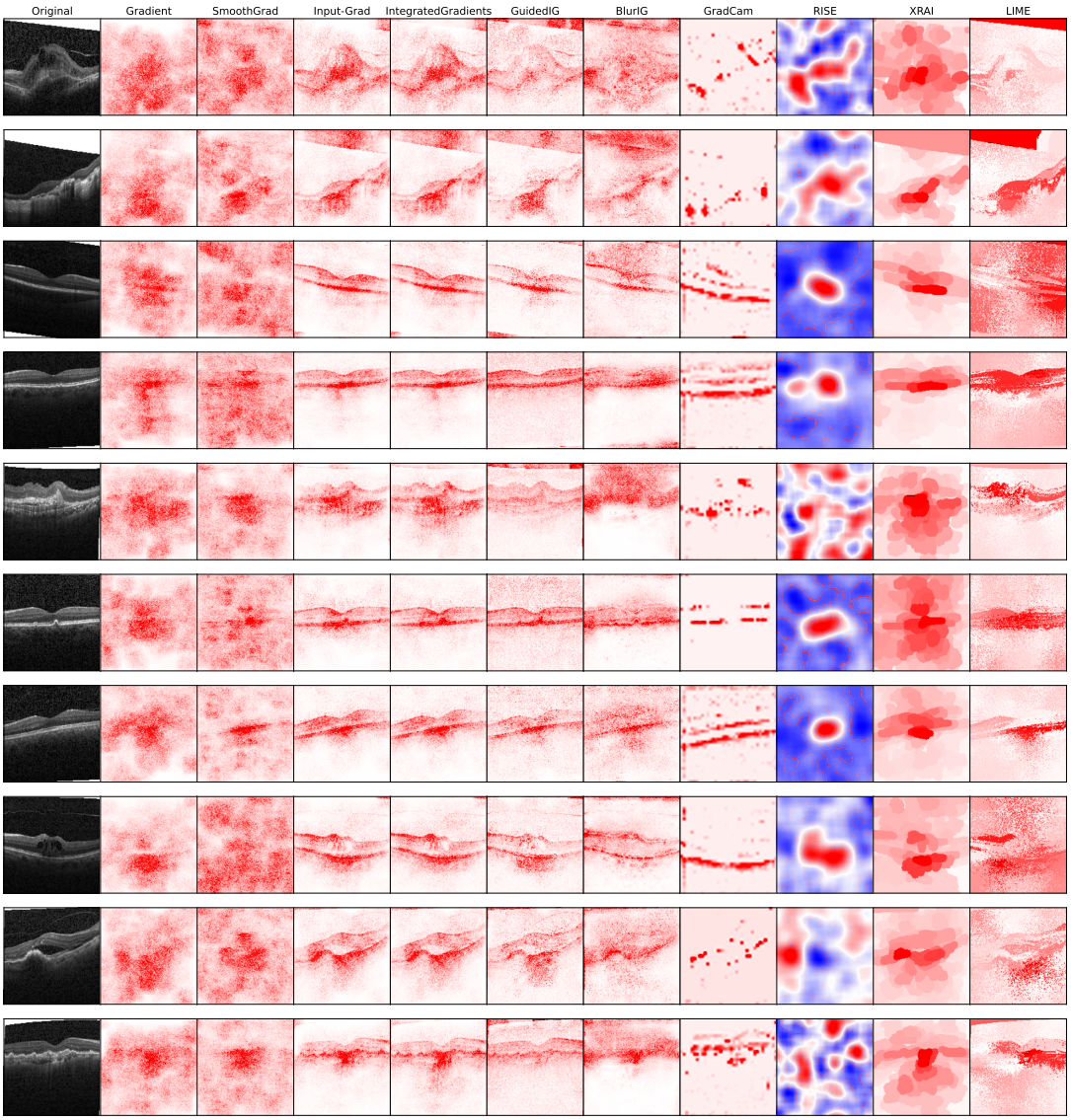


Figura 59 – Exemplos de explicações para o modelo treinado com o conjunto de dados OCT, com rótulo originais, e rede customizada. Na primeira coluna estão apresentadas as imagens originais seguidas, respectivamente, pelas explicações das técnicas: *Gradient, SmoothGrad, Input-Grad, Integrated Gradients, Guided IG, BlurIG, GradCam*, RISE, XRAI e LIME.

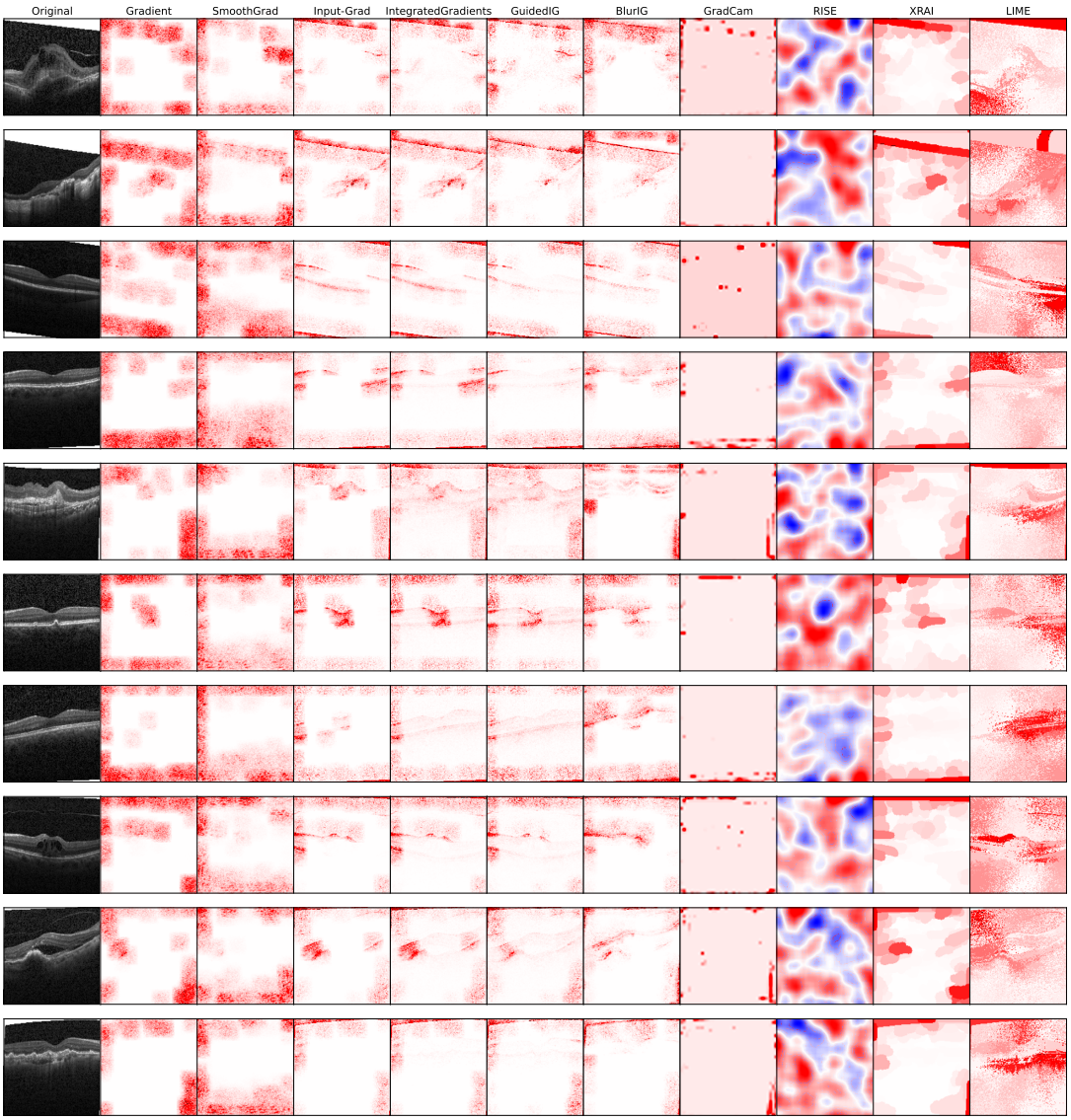


Figura 60 – Exemplos de explicações para o modelo treinado com o conjunto de dados OCT, com rótulo aleatórios, e rede customizada. Na primeira coluna estão apresentadas as imagens originais seguidas, respectivamente, pelas explicações das técnicas: *Gradient, SmoothGrad, Input-Grad, Integrated Gradients, Guided IG, BlurIG, GradCam*, RISE, XRAI e LIME.

Nas Figuras 61 e 62 estão exemplificadas explicações para o modelo customizado, treinado com o conjunto Raio-X com rótulos originais e aleatórios. Percebe-se como algumas das explicações se assemelham a um mero detector de bordas, como nas técnicas Guided-IG. Segundo as explicações do modelo aleatório, o modelo aparenta depender do *backgorund* das imagens.

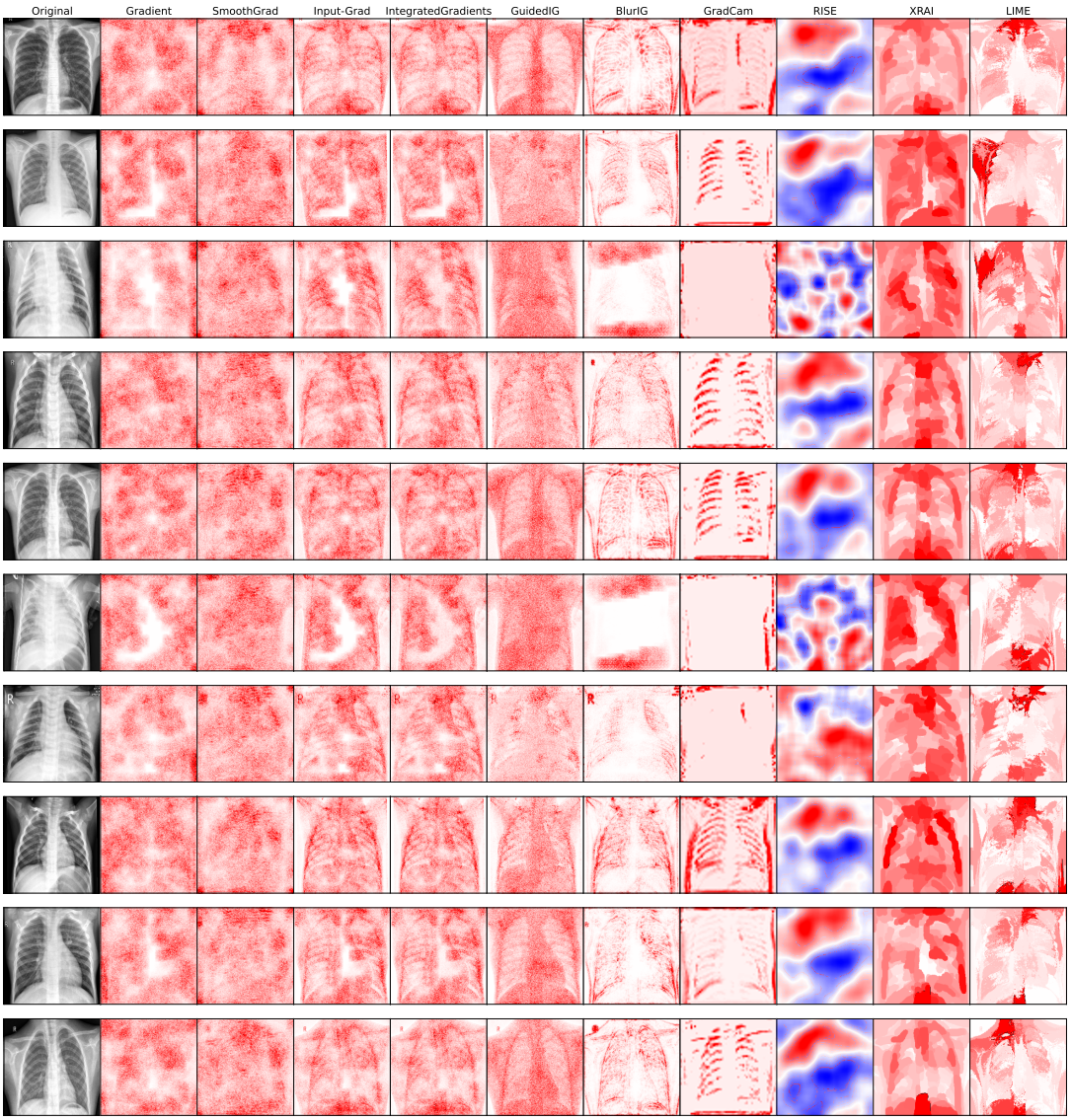


Figura 61 – Exemplos de explicações para o modelo treinado com o conjunto de dados Raio-X, com rótulo originais, e rede customizada. Na primeira coluna estão apresentadas as imagens originais seguidas, respectivamente, pelas explicações das técnicas: *Gradient, SmoothGrad, Input-Grad, Integrated Gradients, Guided IG, BlurIG, GradCam*, RISE, XRAI e LIME.

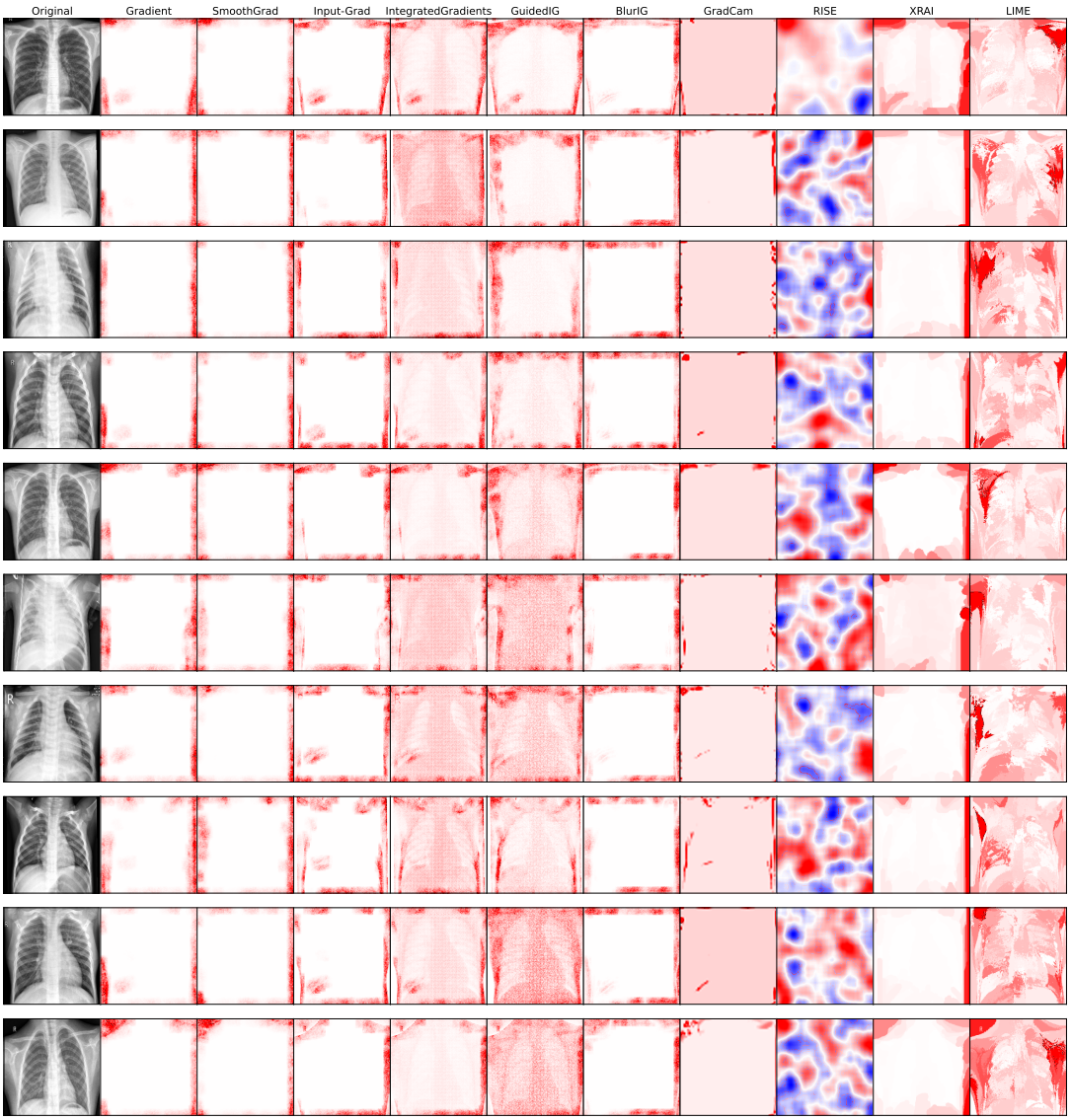


Figura 62 – Exemplos de explicações para o modelo treinado com o conjunto de dados Raio-X, com rótulo aleatórios, e rede customizada. Na primeira coluna estão apresentadas as imagens originais seguidas, respectivamente, pelas explicações das técnicas: *Gradient, SmoothGrad, Input-Grad, Integrated Gradients, Guided IG, BlurIG, GradCam*, RISE, XRAI e LIME.

Nas Figuras 63 e 64 estão exemplificadas explicações para o modelo VGG, treinado com o conjunto OCT com rótulos originais e aleatórios. Algumas das técnicas parecem atribuir responsabilidade ao *backgound* em alguns exemplos. Percebe-se como as explicações para o modelo treinado com rótulos aleatórios parecem "difusas".

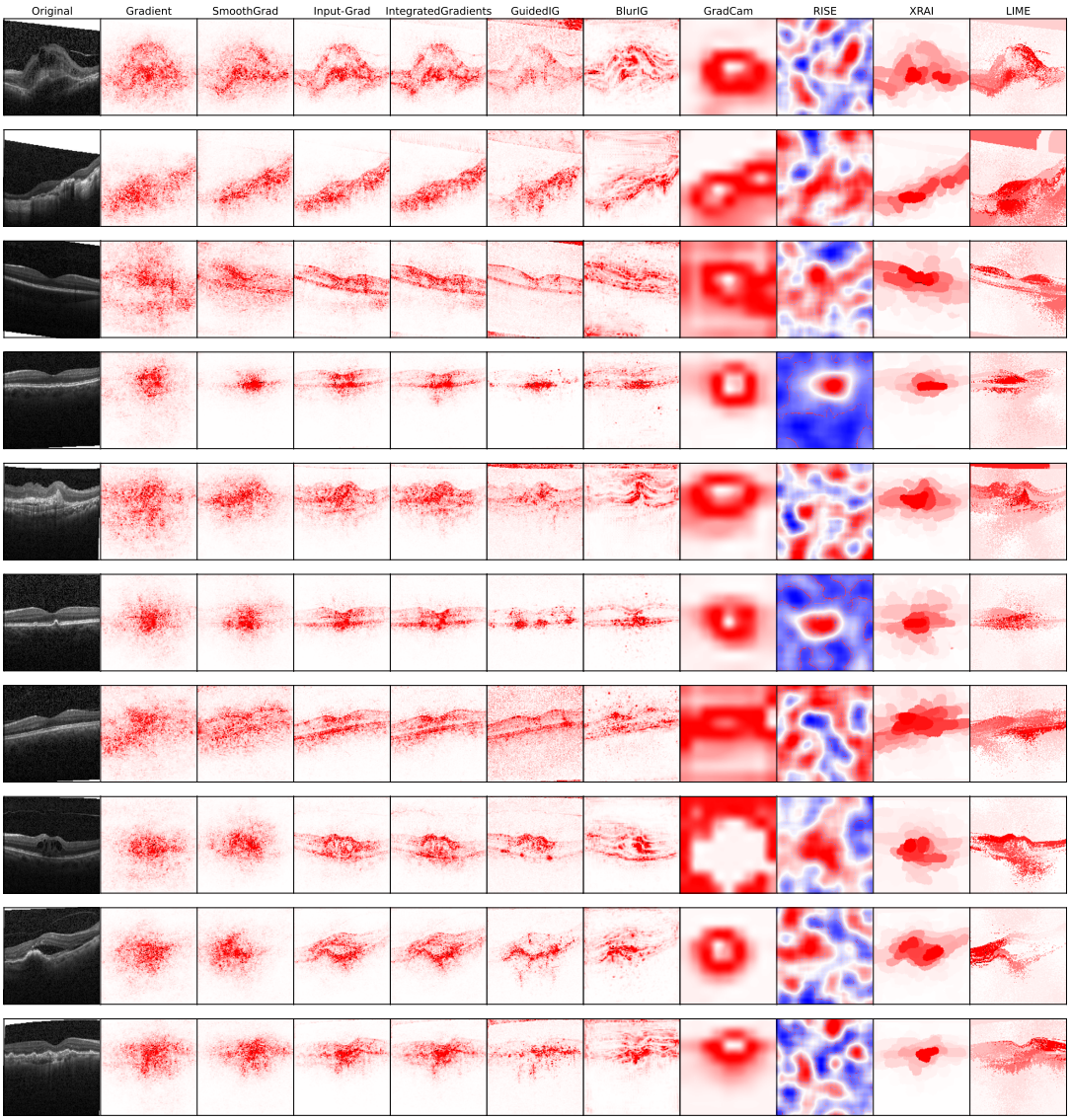


Figura 63 – Exemplos de explicações para o modelo treinado com o conjunto de dados OCT, com rótulo originais, e rede VGG. Na primeira coluna estão apresentadas as imagens originais seguidas, respectivamente, pelas explicações das técnicas: *Gradient, SmoothGrad, Input-Grad, Integrated Gradients, Guided IG, BlurIG, GradCam*, RISE, XRAI e LIME.

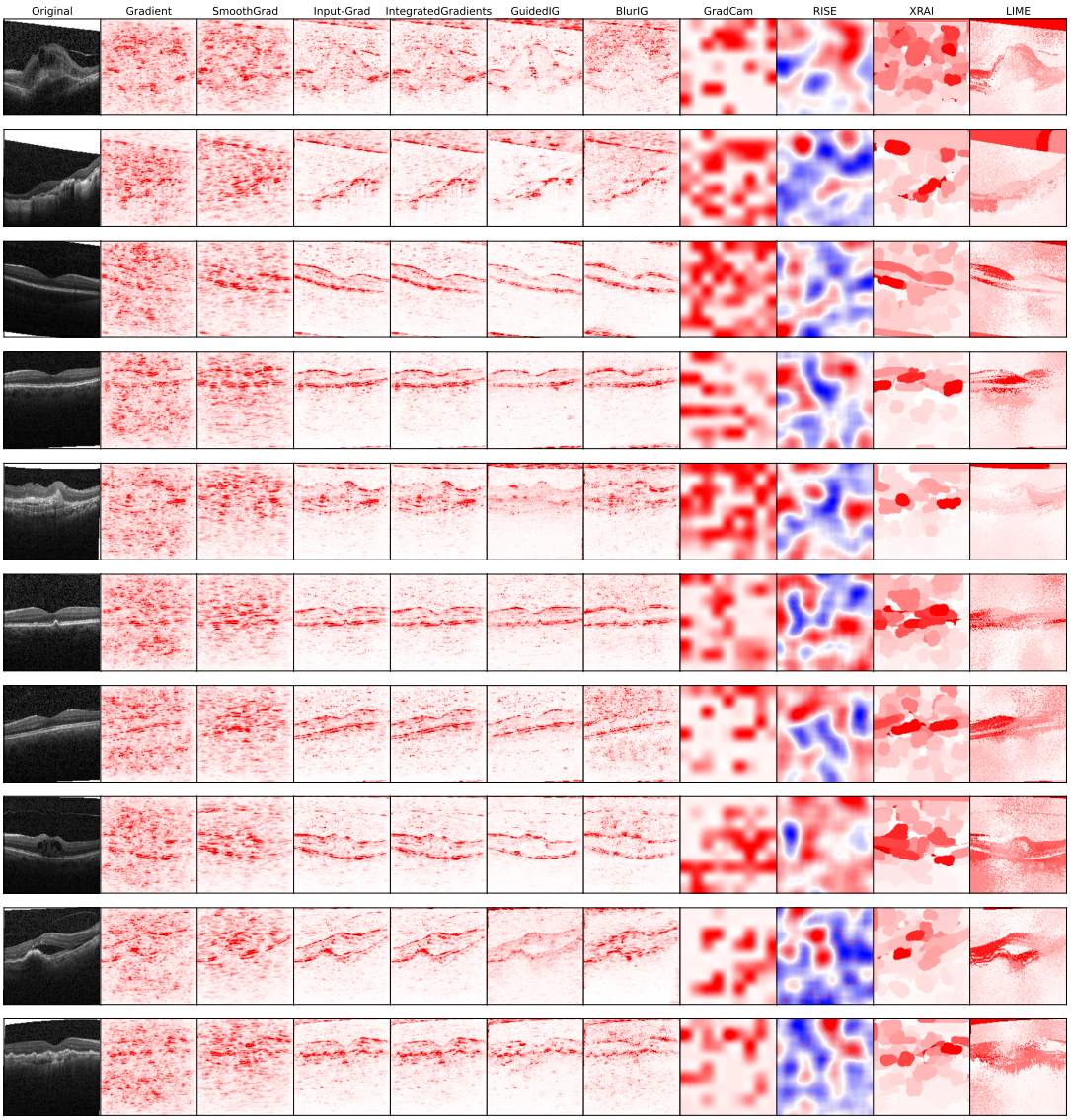


Figura 64 – Exemplos de explicações para o modelo treinado com o conjunto de dados OCT, com rótulo aleatórios, e rede VGG. Na primeira coluna estão apresentadas as imagens originais seguidas, respectivamente, pelas explicações das técnicas: *Gradient, SmoothGrad, Input-Grad, Integrated Gradients, Guided IG, BlurIG, GradCam*, RISE, XRAI e LIME.

Nas Figuras 65 e 66 estão exemplificadas explicações para o modelo VGG, treinado com o conjunto Raio-X com rótulos originais e aleatórios. Percebe-se uma falta de concisão nas explicações, se assemelhando em algumas técnicas a um detector de bordas. Percebe-se como as explicações para o modelo treinado com rótulos aleatórios parecem "difusas".

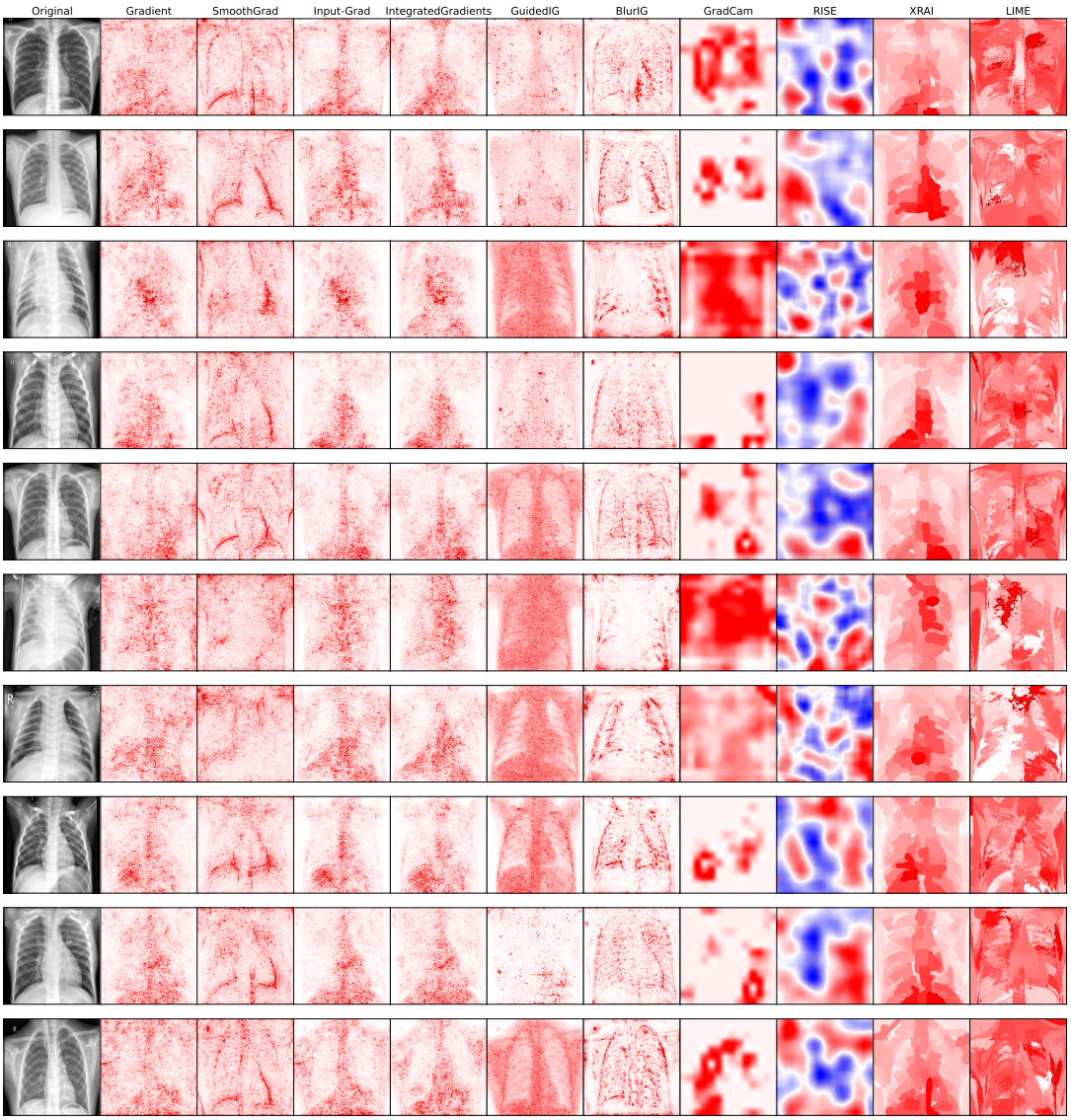


Figura 65 – Exemplos de explicações para o modelo treinado com o conjunto de dados Raio-X, com rótulo originais, e rede VGG. Na primeira coluna estão apresentadas as imagens originais seguidas, respectivamente, pelas explicações das técnicas: *Gradient, SmoothGrad, Input-Grad, Integrated Gradients, Guided IG, BlurIG, GradCam*, RISE, XRAI e LIME.

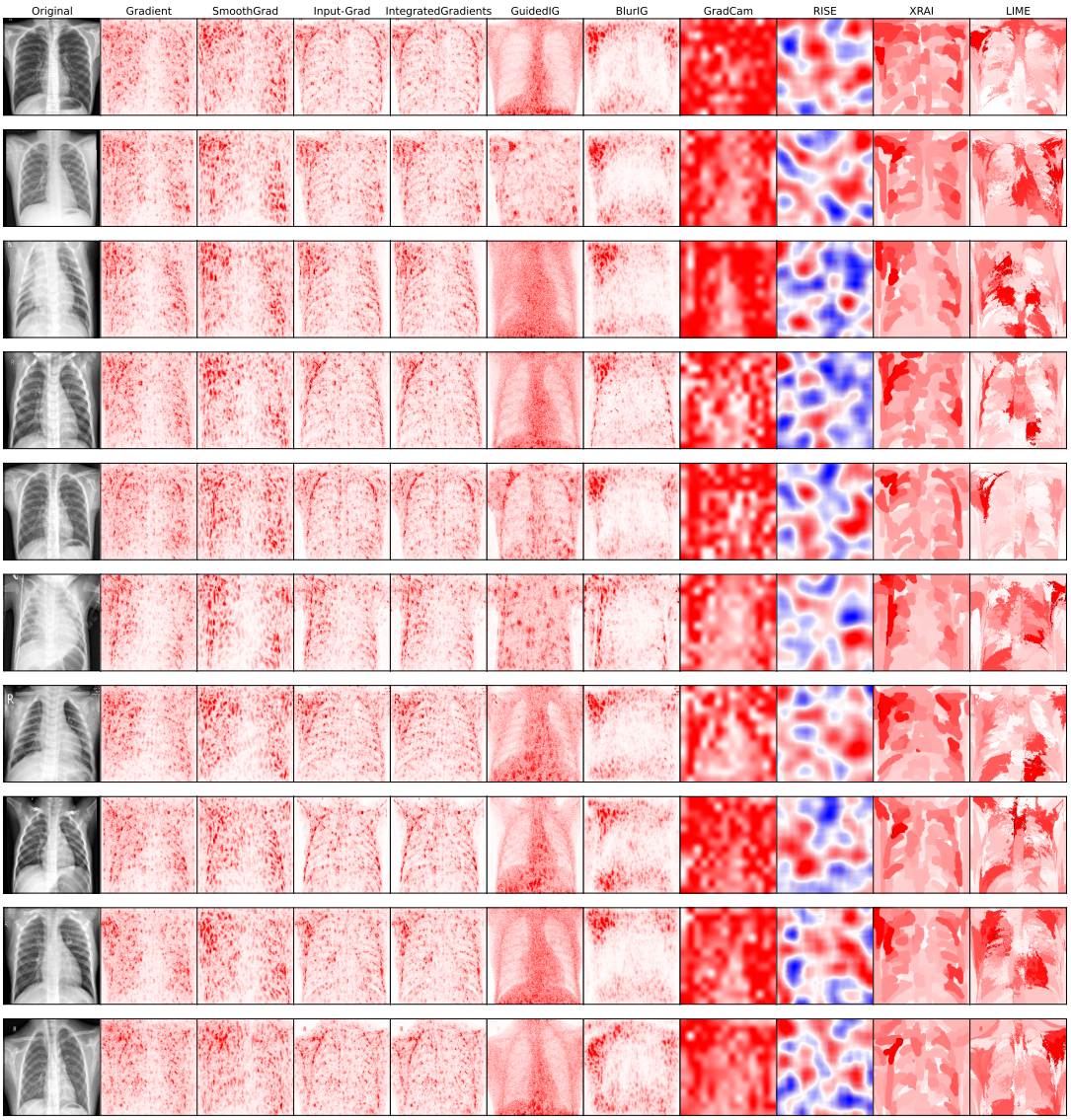


Figura 66 – Exemplos de explicações para o modelo treinado com o conjunto de dados Raio-X, com rótulo aleatórios, e rede VGG. Na primeira coluna estão apresentadas as imagens originais seguidas, respectivamente, pelas explicações das técnicas: *Gradient, SmoothGrad, Input-Grad, Integrated Gradients, Guided IG, BlurIG, GradCam*, RISE, XRAI e LIME.