UNIVERSIDADE FEDERAL DE SÃO JOÃO DEL-REI

Cleverson Marques Vieira

Inteligência artificial explicável aplicada à classificação de imagens de retinografia para apoio no diagnóstico do glaucoma

São João del-Rei

2024

UNIVERSIDADE FEDERAL DE SÃO JOÃO DEL-REI

Cleverson Marques Vieira

Inteligência artificial explicável aplicada à classificação de imagens de retinografia para apoio no diagnóstico do glaucoma

Dissertação apresentada como requisito para obtenção do título de Mestre em Ciência da Computação no Curso de Mestrado do Programa de Pós Graduação em Ciência da Computação da UFSJ.

Orientador: Dr. Diego Roberto Colombo Dias

Universidade Federal de São João del-Rei – UFSJ

Mestrado em Ciência da Computação

São João del-Rei

2024

Ficha catalográfica elaborada pela Divisão de Biblioteca (DIBIB)
e Núcleo de Tecnologia da Informação (NTINF) da UFSJ,
com os dados fornecidos pelo(a) autor(a)

Vieira, Cleverson Marques.
V658i Inteligência artificial explicável aplicada à classificação de imagens de retinografia para apoio no diagnóstico do glaucoma / Cleverson Marques Vieira ; orientador Diego Roberto Colombo Dias. -- São João del-Rei, 2024.
127 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciência da Computação) -- Universidade Federal de São João del-Rei, 2024.

1. Aprendizado de Máquina. 2. Inteligência Artificial . 3. Inteligência Artificial Explicável. 4. Redes Neurais Convolucionais. I. Dias, Diego Roberto Colombo, orient. II. Título.

Cleverson Marques Vieira

Inteligência artificial explicável aplicada à classificação de imagens de retinografia para apoio no diagnóstico do glaucoma

Dissertação apresentada como requisito para obtenção do título de Mestre em Ciência da Computação no Curso de Mestrado do Programa de Pós Graduação em Ciência da Computação da UFSJ.

Trabalho aprovado. São João del-Rei, 28 de fevereiro de 2024:

Dr. Diego Roberto Colombo Dias
Orientador

Dr. Edimilson Batista dos Santos
Convidado 1

Dr. Leonardo Chaves Dutra da Rocha
Convidado 2

Dr. Rodrigo Bonacin
Convidado 3

São João del-Rei
2024

Dedico este trabalho aos meus amados pais e à minha querida esposa, que sempre estiveram ao meu lado, apoiando-me incondicionalmente em cada etapa desta jornada acadêmica. A dedicação deste trabalho é uma singela forma de expressar minha gratidão por tudo o que vocês representam em minha vida.

Agradecimentos

Gostaria de expressar meu profundo agradecimento a todas as pessoas que contribuíram para a realização desta dissertação de mestrado. Em primeiro lugar, expresso minha gratidão a Deus pelo dom da vida e por todo o amparo durante essa caminhada. Aos meus pais, cujo amor, apoio incondicional e encorajamento foram fundamentais para que eu chegasse até aqui. Agradeço à minha esposa, que sempre esteve ao meu lado, compreendendo e incentivando cada passo dessa jornada acadêmica. Manifesto minha gratidão ao meu orientador por todas as sugestões, dicas, críticas e pelo constante apoio, os quais foram muito importantes para o desenvolvimento deste trabalho. Agradeço a todos os professores do Programa de Pós-Graduação em Ciência da Computação (PPGCC) da Universidade Federal de São João del-Rei (UFSJ), que compartilharam seus conhecimentos e proporcionaram um ambiente estimulante de aprendizado. Aos colegas do Instituto Federal do Sudeste de Minas Gerais - Campus Rio Pomba, lotados na Gerência de Tecnologia da Informação (GTI), que me auxiliaram no que foi possível nas atividades diárias para que eu conseguisse realizar todos os créditos necessários durante o curso. Expresso, sinceramente, meus agradecimentos a todos os mencionados e a todos aqueles que, de alguma forma, contribuíram para o sucesso deste trabalho.

"A mente que se abre a uma nova ideia jamais voltará ao seu tamanho original."

Albert Einstein

Resumo

Modelos de aprendizado de máquina têm sido utilizados extensivamente em diversas áreas do conhecimento e possuem inúmeras aplicações em quase todos os segmentos da atividade humana. Na área da saúde, o uso de técnicas de inteligência artificial tem revolucionado o diagnóstico de doenças com excelentes desempenhos na classificação de imagens. Embora esses modelos tenham alcançado resultados extraordinários, a falta de explicabilidade das decisões tomadas pelos modelos tem sido uma limitação significativa para a adoção generalizada na prática clínica. Em ambientes médicos, compreender as decisões dos modelos de inteligência artificial é crucial não apenas para a confiança dos profissionais de saúde, mas também para cumprir requisitos regulatórios, assegurar a segurança dos pacientes e viabilizar a responsabilização em caso de falhas, o que dificulta ainda mais a adoção generalizada dessas técnicas. O glaucoma é uma doença ocular neurodegenerativa que pode levar à cegueira de forma irreversível. A sua detecção precoce é crucial para prevenir a perda de visão. A detecção automatizada do glaucoma tem sido objeto de intensa pesquisa em visão computacional com diversos estudos propondo o uso de redes neurais convolucionais (*Convolutional Neural Networks* - CNNs) para analisar automaticamente imagens de fundo de retina e diagnosticar a doença. No entanto, essas propostas carecem de explicabilidade, o que é crucial para que os oftalmologistas compreendam as decisões tomadas pelos modelos e possam justificá-las aos seus pacientes. Este trabalho tem a finalidade de explorar e aplicar técnicas de inteligência artificial explicável (*eXplainable Artificial Intelligence* - XAI) em diferentes arquiteturas de CNNs para a classificação do glaucoma. Busca-se realizar uma análise comparativa sobre quais métodos de explicação fornecem os melhores recursos para a interpretação humana, servindo de apoio no diagnóstico clínico. Uma abordagem denominada SCIM (*SHAP-CAM Interpretable Mapping*) é proposta demonstrando resultados promissores. Os experimentos conduzidos, com a colaboração de um especialista com mais de sete anos de experiência na área de oftalmologia, revelam que a técnica de interpretabilidade fundamentada no mapeamento de ativação de classe (*Class Activation Mapping* - CAM), aplicada nas arquiteturas VGG16 e VGG19, se destaca como o recurso mais eficaz para promover a interpretabilidade e apoio ao diagnóstico. Para o futuro, propõe-se ampliar a pesquisa, explorando mais métodos de explicação, aprimorando a compreensão das decisões dos modelos de inteligência artificial e buscando reduzir a subjetividade na avaliação humana da interpretabilidade, por meio do desenvolvimento de métricas objetivas e critérios mais claros. Essas investigações têm o potencial de aumentar significativamente a confiança dos profissionais de saúde na adoção dessas inovadoras tecnologias.

Palavras-chave: Aprendizado de Máquina, Inteligência Artificial, Inteligência Artificial Explicável, XAI, Redes Neurais Convolucionais, CNN, Glaucoma.

Abstract

Machine learning models are extensively used in various fields of knowledge and have numerous applications in almost every domain of human activity. In the healthcare field, the use of artificial intelligence techniques has revolutionized disease diagnosis, with excellent performance in image classification. Although these models have achieved extraordinary results, the lack of explainability in model decisions has been a significant limitation for the widespread adoption of these techniques in clinical practice. In medical environments, understanding the decisions of artificial intelligence models is crucial not only for the confidence of healthcare professionals but also to meet regulatory requirements, ensure patient safety, and facilitate accountability in case of failures. These challenges further complicate the widespread adoption of such techniques. Glaucoma is a neurodegenerative eye disease that can lead to irreversible blindness. Early detection is crucial for preventing vision loss. Automated glaucoma detection has been the subject of intense research in computer vision, with several studies proposing the use of convolutional neural networks (CNNs) to automatically analyze retinal fundus images and diagnose the disease. However, these proposals lack explainability, which is crucial for ophthalmologists to understand model decisions and justify them to their patients. This work aims to explore and apply eXplainable Artificial Intelligence (XAI) techniques to different architectures of CNNs for glaucoma classification. The goal is to conduct a comparative analysis of which explanation methods provide the best resources for human interpretation, serving as support in clinical diagnosis. An approach called SCIM (SHAP-CAM Interpretable Mapping) is proposed, demonstrating promising results. The experiments conducted, with the collaboration of one expert with over seven years of experience in the field of ophthalmology, reveal that the interpretability technique based on Class Activation Mapping (CAM), applied to the VGG16 and VGG19 architectures, stands out as the most effective tool for promoting interpretability and diagnostic support. For the future, it is proposed to expand the research by exploring more explanation methods, refining the understanding of decisions made by artificial intelligence models, and seeking to reduce subjectivity in human assessment of interpretability through the development of objective metrics and clearer criteria. These investigations have the potential to significantly increase the confidence of healthcare professionals in adopting these innovative technologies.

Key-words: Machine Learning, Artificial Intelligence, Explainable Artificial Intelligence, XAI, Convolutional Neural Networks, CNN, Glaucoma.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AUC	*Area Under the Curve*
ANN	*Artificial Neural Network*
AG-CNN	*Attention Guided Convolutional Neural Network*
AM	*Attention Mining*
CAD	*Computer Aided Diagnosis*
CAM	*Class Activation Mapping*
CDR	*Cup-to-Disc Ratio*
CLAHE	*Contrast Limited Adaptative Histogram Equalization*
CNN	*Convolutional Neural Network*
DA	*Data Augmentation*
DC	*Distance Correlation*
DCNN	*Deep Convolutional Neural Networks*
DISSIM	*Dissimilarity*
DL	*Deep Learning*
Grad-CAM	*Gradient-weighted Class Activation Mapping*
IA	Inteligência Artificial
ISNT	Inferior, Superior, Nasal, Temporal
LAG	*Large-scale Attention based Glaucoma*
LCD	*Liquid Crystal Display*
LIME	*Local Interpretable Model-agnostic Explanations*
ML	*Machine Learning*
MMP	*Meaningful Perturbation*
OCT	*Optical Coherence Tomography*

PIO	Pressão Intraocular
R-CNN	*Region-based Convolutional Neural Network*
ReLu	*Rectified Linear Unit*
RNA	Rede Neural Artificial
RNFL	*Retinal Nerve Fiber Layer*
ROC	*Receiver Operating Characteristic*
SCIM	*SHAP-CAM Interpretable Mapping*
SHAP	*SHapley Additive exPlanations*
SLIC	*Simple Linear Iterative Clustering*
TL	*Transfer Learning*
VCD	*Vertical of Cup Diameter*
VDD	*Vertical of Disc Diameter*
XAI	*eXplainable Artificial Intelligence*

Sumário

1 Introdução

Modelos de aprendizado de máquina - *Machine Learning* (ML) - estão sendo utilizados extensivamente em diversas áreas do conhecimento e possuem inúmeras aplicações em quase todos os segmentos da atividade humana. Esses modelos têm se tornado cada vez mais sofisticados e complexos à medida que uma grande quantidade de dados são envolvidos na resolução de problemas. Quanto maior a complexidade computacional desses modelos, mais incompreensíveis se tornam para os humanos (AGARWAL; DAS, 2020).

O uso de técnicas de inteligência artificial (IA) na área da saúde tem revolucionado o diagnóstico de doenças com excelentes desempenhos na classificação de imagens. Mesmo com resultados extraordinários, a adoção generalizada dessas técnicas na prática clínica ainda ocorre em ritmo moderado (NAZIR, 2023). Uma das principais limitações está relacionada a dificuldade em obter informações ou justificativas sobre como as decisões estão sendo tomadas pelo modelo (KUMAR, 2019). Em ambientes médicos, compreender as decisões dos modelos de IA é crucial não apenas para a confiança dos profissionais de saúde, mas também para cumprir requisitos regulatórios, assegurar a segurança dos pacientes e viabilizar a responsabilização em caso de falhas, o que dificulta ainda mais a adoção generalizada dessas técnicas (NAZIR, 2023). Neste contexto, uma das doenças abordadas por essas técnicas é o glaucoma.

O glaucoma é uma doença ocular neurodegenerativa de caráter irreversível, considerada uma das principais causadoras de cegueira no mundo. Como pode ser assintomática, a detecção precoce e o tratamento são importantes para prevenir a perda da visão. Esta doença ocular silenciosa é caracterizada principalmente pela perda da fibra do nervo óptico e que se dá pelo aumento da pressão intraocular (PIO) e/ou pela perda de fluxo sanguíneo para a região do nervo óptico. No entanto, a medição da PIO não é específica nem sensível o suficiente para ser um indicador eficaz de glaucoma, pois o dano visual pode estar presente sem PIO aumentada (DIAZ-PINTO, 2019).

Segundo Sarhan, Nasseri, Zapp, Maier, Lohmann, Navab e Eslami (2020), o glaucoma afeta o nervo óptico de forma progressiva e é comumente detectado por meio de três abordagens: detecção de aumento da PIO, identificação do campo de visão normal e avaliação do dano do nervo óptico por cálculo da relação copa-disco (CDR). A avaliação da cabeça do nervo óptico é uma das técnicas de triagem clinicamente mais significativas para o glaucoma. Nesta avaliação, algumas medidas são propostas como condições clínicas para a triagem de glaucoma, por exemplo a razão vertical copa-disco (CDR vertical), diâmetro do disco, área da borda e regra ISNT (STEFAN, 2020).

A detecção automatizada do glaucoma é uma área de pesquisa muito ativa entre

os pesquisadores de visão computacional. Na literatura há diversas propostas de trabalhos que utilizam imagens oftalmológicas, de fundo de retina, como fonte de dados para algoritmos baseados em redes neurais convolucionais - *Convolutional Neural Networks* (CNNs), que são capazes de analisar de forma automática as características anatômicas do nervo óptico e realizar um possível diagnóstico do olho avaliado (NOROUZIFARD, 2018; DIAZ-PINTO, 2019; GóMEZ-VALVERDE, 2019; SERENER; SERTE, 2019; MARTINS, 2020). Apesar de muito assertivas e apresentarem um excelente desempenho com relação ao diagnóstico, essas propostas são limitadas por realizarem predições com pouca ou mesmo nenhuma explicabilidade. Essa é uma limitação importante, pois o oftalmologista precisa compreender o porquê do diagnóstico apresentado para seu paciente.

Os modelos de aprendizado profundo - *Deep Learning* (DL), principalmente as CNNs, têm obtido excelentes resultados para a classificação do glaucoma através da análise de imagens de fundo de retina. No entanto, uma compreensão mais aprofundada com relação a interpretabilidade por trás da previsão e decisão desses modelos não tem sido muito estudada.

Este trabalho tem a finalidade de explorar e aplicar técnicas de inteligência artificial explicável - *eXplainable Artificial Intelligence* (XAI) - em diferentes arquiteturas de CNNs para a classificação do glaucoma em imagens de retinografia e comparar quais métodos fornecem os melhores recursos para a análise e interpretação humana, servindo de apoio no diagnóstico do glaucoma. Uma abordagem inédita baseada na confluência de outras duas técnicas, SHAP (LUNDBERG; LEE, 2017) e CAM (ZHOU, 2016), denominada SCIM (*SHAP-CAM Interpretable Mapping*) é proposta, neste trabalho, demonstrando resultados promissores. Os experimentos conduzidos, com a colaboração de um especialista com mais de sete anos de experiência na área de oftalmologia, sugerem que a técnica de interpretabilidade fundamentada no mapeamento de ativação de classe - *Class Activation Mapping* (CAM) - aplicada nas arquiteturas VGG16 (SIMONYAN; ZISSERMAN, 2015) e VGG19 (SIMONYAN; ZISSERMAN, 2015), se destaca como o recurso mais eficaz para promover a interpretabilidade e apoio ao diagnóstico.

1.1 Objetivos

O objetivo principal deste trabalho é explorar e aplicar técnicas de XAI em diferentes arquiteturas de CNNs para a classificação de imagens de retinografia e comparar quais métodos de explicação fornecem os melhores recursos para a análise e interpretação humana, servindo de apoio no diagnóstico do glaucoma.

Para alcançar esse objetivo principal, os seguintes objetivos específicos são perseguidos:

- Implementar diferentes arquiteturas de CNNs, incluindo VGG16, VGG19, InceptionV3 (SZEGEDY, 2015), DenseNet (HUANG, 2018), Xception (CHOLLET, 2016) e ResNet50 (HE, 2015), para a classificação de imagens de retinografia;
- Treinar e avaliar os modelos em um conjunto de dados de imagens de retinografia, disponíveis publicamente, que estão rotuladas como saudáveis (sem glaucoma) ou com glaucoma;
- Implementar e aplicar técnicas de interpretabilidade visual, incluindo o Mapeamento de Ativação de Classe (CAM), Mapeamento de Ativação de Classe Ponderada por Gradiente (Grad-CAM) (SELVARAJU, 2017), Explicações Aditivas Shapley (SHAP), Explicações Interpretáveis Locais Independentes de Modelo (LIME) (RIBEIRO, 2016), Vanilla Gradients (SIMONYAN, 2014) e SmoothGrad (SMILKOV, 2017);
- Avaliar a eficácia das técnicas de XAI na identificação das áreas da imagem que mais influenciaram a decisão do modelo de classificação;
- Propor e implementar uma nova abordagem fundamentada na região de confluência de duas técnicas distintas (CAM e SHAP). Essa nova abordagem, denominada SCIM, visa gerar uma interpretação visual mais abrangente, precisa e convergente para as predições da CNN.

Ao alcançar esses objetivos específicos, este trabalho representa uma contribuição significativa para o aprofundamento e aplicação das técnicas de XAI no campo da IA. Adicionalmente, este trabalho visa o desenvolvimento de modelos de IA mais transparentes, interpretáveis e confiáveis, proporcionando suporte médico no diagnóstico da doença.

1.2 Justificativa

A principal motivação deste trabalho é a falta de transparência e interpretabilidade dos modelos de DL, principalmente as CNNs, aplicados na classificação de imagens de retinografia para o diagnóstico do glaucoma. Embora os modelos de IA possam alcançar taxas de desempenho notáveis (acurácia, precisão, sensibilidade, *f1-score* e AUC) na classificação automatizada de imagens de retinografia para o diagnóstico do glaucoma, uma compreensão mais aprofundada com relação a interpretabilidade por trás da previsão para o processo de decisão desses modelos não tem sido muito estudada. Essa falta de transparência pode levar à desconfiança dos médicos e pacientes em relação ao uso de modelos de IA no apoio ao diagnóstico da doença.

Além disso, a interpretabilidade dos modelos de IA é particularmente importante na área médica, pois os médicos precisam entender as justificativas para as decisões do

modelo de forma a tomar decisões mais embasadas e precisas. Sem essa compreensão, os médicos podem ter dificuldades em interpretar e aplicar as recomendações do modelo em seus diagnósticos.

Portanto, o problema de pesquisa abordado neste trabalho é como tornar os modelos de IA aplicados na classificação de imagens de retinografia mais confiáveis, transparentes e interpretáveis, a fim de permitir que os médicos entendam como as decisões são tomadas e para que possam tomar decisões mais embasadas e precisas no diagnóstico do glaucoma. A utilização de técnicas de XAI pode contribuir significativamente para o entendimento e resolução desse problema.

1.3 Organização do Texto

Este trabalho está dividido em seis capítulos, incluindo o capítulo de introdução (Capítulo 1) que compreende a contextualização, o problema de pesquisa e os objetivos.

O Capítulo 2 apresenta o referencial teórico, que explica e exemplifica as teorias e conceitos fundamentais que sustentam o desenvolvimento deste trabalho. No Capítulo 3 são apresentados os trabalhos relacionados, ou seja, uma revisão da literatura de publicações que desenvolvem ideias similares e/ou relacionadas de alguma forma com a temática deste trabalho. O Capítulo 4, metodologia e experimentos, descreve sobre a condução deste trabalho de pesquisa, incluindo os conjuntos de dados que foram utilizados, instrumentos e técnicas implementadas no decorrer deste estudo. No Capítulo 5, são apresentados os resultados derivados dos experimentos realizados, acompanhados por uma análise comparativa e uma avaliação crítica. Por fim, no Capítulo 6, as considerações finais e conclusão são apresentadas sumarizando as contribuições, limitações e sugestões para pesquisas futuras.

2 Referencial Teórico

Neste capítulo, são apresentados os aspectos teóricos relativos aos temas Glaucoma, IA, CNNs, XAI e Técnicas de XAI, básicos para o entendimento da proposta central deste trabalho de pesquisa.

2.1 Glaucoma

O glaucoma é uma doença oftalmológica de difícil diagnóstico e uma das principais doenças relacionadas à deficiência visual que apresenta neurodegeneração progressiva irreversível (MANASSAKORN, 2022), em outras palavas, a maior causadora de cegueira irreversível do mundo. A lesão do nervo óptico, que é a principal causa do glaucoma, é produzida pelo aumento da PIO (STEFAN, 2020).

De acordo com Stefan, Paraschiv, Ovreiu e Ovreiu (2020), para uma pessoa saudável, a PIO média deve estar entre 14 e 16 milímetros de mercúrio (mmHg) e uma PIO de 22 ou mais é considerada anormal e pode causar danos às células do nervo óptico. Todavia, apenas a medição da PIO não é suficiente para o diagnóstico.

Segundo Sarhan, Nasseri, Zapp, Maier, Lohmann, Navab e Eslami (2020), o glaucoma afeta o nervo óptico de forma progressiva e é comumente detectado por meio de três abordagens: detecção de aumento da PIO, identificação do campo de visão normal e avaliação do dano do nervo óptico pelo cálculo do CDR. A primeira abordagem é a mais utilizada em consultórios oftalmológicos pela facilidade de se realizar a avaliação. Todavia, é comum diagnósticos de falso negativo pois a PIO pode ficar inalterada em todos os estágios da doença. A segunda é dependente de exames complementares para avaliação do campo visual. Já a terceira abordagem costuma ser mais eficaz e precisa no diagnóstico, realizada por fundoscopia, ou seja, avaliações de imagens detalhadas do nervo óptico identificando possíveis alterações em sua anatomia, mais precisamente pelo aumento vertical da copa do disco óptico.

A avaliação da cabeça do nervo óptico é uma das técnicas de triagem clinicamente mais significativas para o glaucoma (STEFAN, 2020). Nesta avaliação, algumas medidas são propostas como condições clínicas para a triagem de glaucoma, como a razão vertical copa-disco (CDR vertical), diâmetro do disco, área da borda e regra ISNT.

De acordo com Zangalli, Gupta e Spaeth (2011), o tamanho e a forma do disco óptico bem como a escavação óptica é um aspecto importante a ser levado em consideração durante o diagnóstico de glaucoma. Dessa forma, o aumento vertical da escavação, ou seja, o aumento do CDR é uma característica da neuropatia óptica glaucomatosa. Analisando a

Figura 1 (c) e (d), é possível identificar um aumento na escavação se comparado à Figura 1 (a) e (b), o que representa um sinal claro de glaucoma (BARROS, 2020).

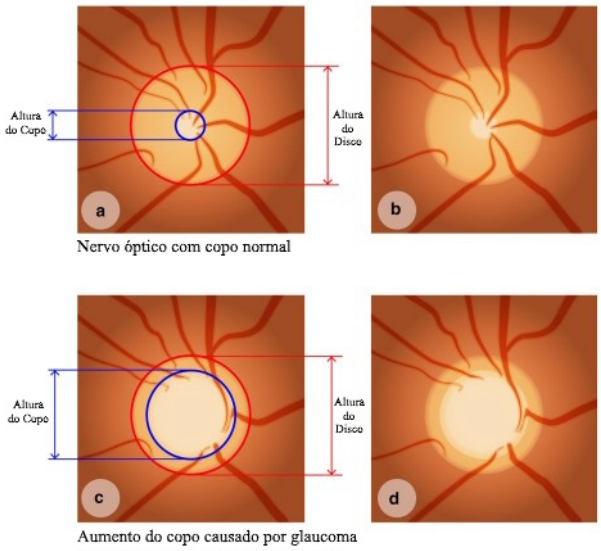


Figura 1 – Representação do nervo óptico com escavação normal (a)(b) e escavação aumentada (c)(d) causada pelo glaucoma - Fonte: Adaptado de (BARROS, 2020).

O CDR vertical corresponde à razão entre a vertical do diâmetro da copa (VCD) para o diâmetro vertical do disco (VDD). Para ilustrar essa métrica, observe a Figura 2, em que o VCD é indicado pela letra v e o VDD indicado pela letra V. Basicamente, o CDR mede o alargamento da copa em relação ao disco (afinamento do borda neurorretiniana). Quanto maior o CDR, maior o risco para o glaucoma (KUMAR, 2016; ZHAO, 2020).

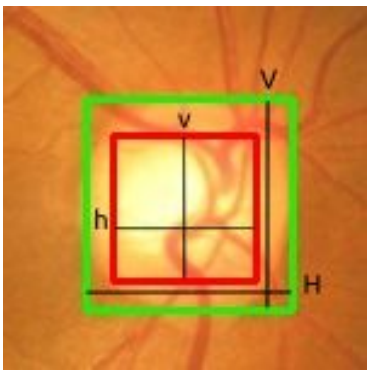


Figura 2 – Exemplo de CDR. Fonte: (OLIVEIRA, 2023).

A regra ISNT é utilizada para se diferenciar o nervo óptico normal do nervo óptico glaucomatoso. Em olhos normais a espessura da borda do disco de Inferior (I) é maior que a espessura da borda Superior (S) que é maior que a espessura da borda Nasal (N) que é maior que a espessura da borda Temporal (T) (KUMAR, 2016). A Figura 3 representa uma avaliação clínica da regra ISNT para um nervo óptico normal.

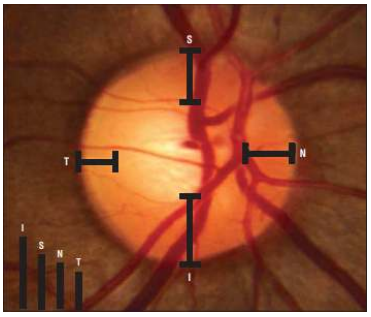


Figura 3 – Regra ISNT em um nervo óptico normal. Fonte: (HARIZMAN, 2006).

Para que se obtenham medidas precisas, é necessário submeter o paciente a um exame conhecido como Tomografia de Coerência Óptica (*Optical Coherence Tomography* - OCT) (VESSANI, 2008), que por sua vez tem um custo elevado.

2.2 Inteligência Artificial (IA)

A IA é uma das ciências mais recentes, teve início após a Segunda Guerra Mundial e, atualmente, abrange uma enorme variedade de subcampos, desde áreas de uso geral, como aprendizado e percepção, até tarefas específicas como jogos de xadrez, demonstração de teoremas matemáticos, criação de poesia e diagnóstico de doenças. A IA sistematiza e automatiza tarefas intelectuais e, portanto, é potencialmente relevante para qualquer esfera da atividade intelectual humana. Nesse sentido, ela é um campo universal (NORVIG; RUSSELL, 2013).

Ao longo do tempo a IA seguiu quatro linhas de pensamento:

- I - Sistemas que pensam como seres humanos: “O novo e interessante esforço para fazer os computadores pensarem... máquinas com mentes, no sentido total e literal” (HAUGELAND, 1985).
- II - Sistemas que atuam como seres humanos: “A arte de criar máquinas que executam funções que exigem inteligência quando executadas por pessoas” (KURZWEIL, 1990).
- III - Sistemas que pensam racionalmente: “O estudo das faculdades mentais pelo seu uso de modelos computacionais” (CHARNIAK; MCDERMOTT, 1985) .
- IV - Sistemas que atuam racionalmente: “A Inteligência Computacional é o estudo do projeto de agentes inteligentes” (POOLE, 1998).

2.2.1 Aplicações da Inteligência Artificial

De acordo com Monard e Baranauskas (2000), a IA é um ramo da ciência da computação cujo interesse é fazer com que os computadores pensem ou se comportem de forma inteligente. Por se tratar de um tópico muito amplo, IA também está relacionada com psicologia, biologia, lógica matemática, linguística, engenharia, filosofia, entre outras áreas científicas, conforme mostra a Figura 4.

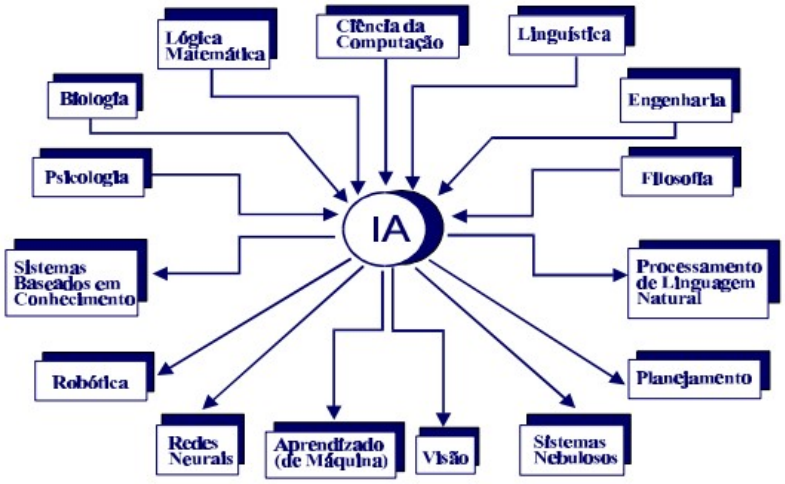


Figura 4 – Áreas relacionadas com a IA. Fonte: (MONARD; BARANAUSKAS, 2000).

Desde então, a área de IA vem se desenvolvendo em várias linhas de pesquisa - e.g. Sistemas Baseados em Conhecimento, Robótica, Redes Neurais, Aprendizado de Máquina, Visão, Lógica Nebulosa, Planejamento, Processamento e Interpretação de Linguagem Natural, Reconhecimento de Padrões - com o objetivo de fornecer ao computador as habilidades para efetuar funções antes desempenhadas apenas através da inteligência humana (MONARD; BARANAUSKAS, 2000).

2.2.2 Redes Neurais Artificiais (RNAs)

As RNAs são algoritmos computacionais que apresentam um modelo matemático inspirado na estrutura de organismos inteligentes, os quais possibilitam inserir simplificadamente o funcionamento do cérebro humano em computadores. Dessa forma, a exemplo do cérebro humano, a RNA é capaz de aprender e tomar decisões baseadas em seu próprio aprendizado (FLECK, 2016).

De acordo com Haykin (2001), a RNA se assemelha ao cérebro humano em dois aspectos básicos: (1) o conhecimento é adquirido pela rede a partir de seu ambiente, por intermédio do processo de aprendizagem e (2) as forças de conexão entre neurônios (pesos sinápticos) são utilizadas para armazenar o conhecimento adquirido (FLECK, 2016).

Uma RNA extrai seu poder computacional através de sua habilidade de aprender e de generalizar. A generalização se refere ao fato da RNA produzir saídas adequadas para entradas inexistentes durante a aprendizagem, tornando possível a resolução de problemas complexos, atualmente intratáveis (FLECK, 2016).

Diversos tipos de RNAs com distintas topologias estão disponíveis. No âmbito desta pesquisa, serão empregadas diversas arquiteturas de CNNs.

2.2.3 Redes Neurais Convolucionais (CNNs)

As CNNs são um tipo especializado de RNA para processamento de dados extremamente bem-sucedidas em aplicações práticas para classificação de imagens. O nome "rede neural convolucional" indica que a rede emprega uma operação matemática chamada convolução. A convolução é um tipo especializado de operação linear. CNNs são simplesmente RNAs que usam convolução no lugar da multiplicação geral de matrizes em pelo menos uma de suas camadas (GOODFELLOW, 2016).

Uma CNN pode captar uma imagem de entrada, atribuir importância (pesos e vieses que podem ser aprendidos) a vários aspectos/objetos da imagem e ser capaz de diferenciar um do outro. O pré-processamento exigido em uma CNN é muito menor em comparação com outros algoritmos de classificação (NIELSEN, 2018).

Uma CNN é capaz de capturar com sucesso as dependências espaciais e temporais em uma imagem através da aplicação de filtros relevantes. A arquitetura executa um melhor ajuste ao conjunto de dados da imagem devido à redução no número de parâmetros envolvidos e à capacidade de reutilização dos pesos. Em outras palavras, a rede pode ser treinada para entender melhor a sofisticação da imagem (NIELSEN, 2018).

Tipicamente, uma CNN possui um conjunto de camadas denominadas camadas convolucionais, camadas de *pooling* e a camada *fully-connected* (NIELSEN, 2018).

2.2.3.1 Camada Convolucional

As camadas convolucionais são o bloco fundamental de uma CNN. Elas são responsáveis por extrair recursos (*features*) relevantes das imagens de entrada. Cada camada convolucional é composta por um conjunto de filtros convolucionais, que são pequenas matrizes de pesos. Esses filtros deslizam sobre a imagem de entrada, calculando produtos de convolução em cada posição, o que resulta em um mapa de características. Cada filtro é projetado para detectar uma característica específica, como bordas, texturas ou padrões mais complexos. Ao combinar múltiplos filtros, a camada convolucional pode capturar diferentes tipos de características simultaneamente. O resultado é um conjunto de mapas de características que representam a imagem de entrada (GOODFELLOW, 2016).

Matematicamente, isso é feito primeiro com um detector de recurso ou *kernel* de tamanho k × k, deslizando-o no espaço de entrada e executando a convolução entre o *kernel* e o *patch* de entrada em cada ponto. O tamanho do *kernel* geralmente é menor que o espaço de entrada. A profundidade do *kernel*, no entanto, deve ser a mesma que a profundidade de entrada. Vários núcleos são usados em cada camada convolutiva para preservar melhor as dimensões espaciais. Cada *kernel* procura um recurso específico no espaço de entrada e produz um mapa de recursos. Como a convolução é uma operação linear (multiplicação sábia de elementos dos valores do *kernel* e do *patch* de entrada seguida pela soma), executar várias convoluções em várias camadas acaba em uma grande operação linear e, portanto, limita a capacidade de aprendizado da rede. Para resolver esse problema, a saída de cada camada convolutiva é passada através de uma função não linear. A unidade linear retificada (ReLU), definida como f(x) = max (0, x), é a função não linear mais popular usada para aumentar as propriedades não lineares da rede. (BAJWA, 2019).

2.2.3.2 Camada de *Pooling*

As camadas de *pooling* (agrupamento) são usadas para reduzir a amostra dos mapas de recursos sem perder informações significativas. Isso é feito pegando uma pequena janela de tamanho p × p de cada fatia do mapa de recursos e fornecendo, por exemplo, o valor médio dessa janela. A operação de *pooling* é aplicada a cada mapa de características, substituindo cada região por um único valor representativo. A forma mais comum de *pooling* é o *MaxPooling*, que seleciona o valor máximo em cada região. Isso ajuda a tornar a rede mais robusta a pequenas variações de posição e tamanho nas características detectadas, além de reduzir a quantidade de parâmetros da rede (BAJWA, 2019).

2.2.3.3 Camada *Fully-Connected*

A camada *fully-connected* (totalmente conectada) é uma camada densa tradicional de uma RNA, na qual todos os neurônios estão conectados a todos os neurônios da camada anterior. Ela segue as camadas convolucionais e de *pooling* em uma CNN. Essa camada é responsável pela classificação final dos recursos extraídos pela rede. Os mapas de características das camadas anteriores são achatados em um vetor unidimensional e alimentados para a camada *fully-connected*. Essa camada usa esses vetores como entrada e aplica operações lineares e não lineares para mapear esses recursos em classes específicas, fornecendo uma previsão final (GOODFELLOW, 2016).

A Figura 5 representa a estrutura de uma CNN, destacando a etapa de entrada, as camadas convolucionais (C), as camadas de *pooling* (P), a camada *fully-connected* (FC) e a saída da rede. Essa representação ilustra a habilidade intrínseca dessa arquitetura na extração e interpretação hierárquica de características presentes em dados visuais.

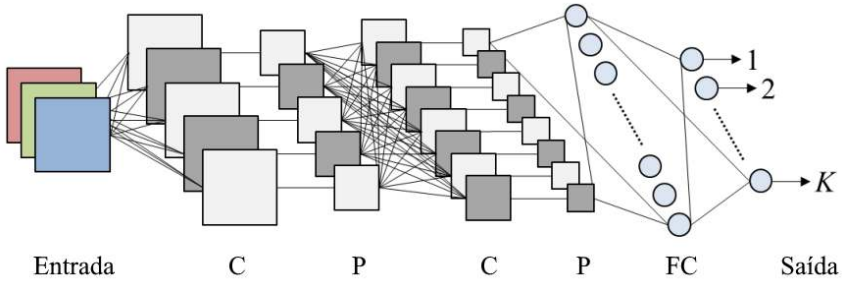


Figura 5 – Exemplo de uma arquitetura de CNN, onde C, P e FC representam as camadas convolucionais, camadas de *pooling* e a camada *fully-connected*, respectivamente. Fonte: Adaptado de (HIDAKA; KURITA, 2017).

2.3 Inteligência Artificial Explicável (XAI)

A XAI consiste em um agente que possui a capacidade de revelar as causas por trás das tomadas de decisões dele ou de um agente externo. Assim, ele consiste em um problema de interação humano-agente, que pode ser definida como a interseção entre os campos de IA, ciências sociais e interação humano-computador (MILLER, 2017). Neste sentido, a Figura 6 representa o escopo da XAI.

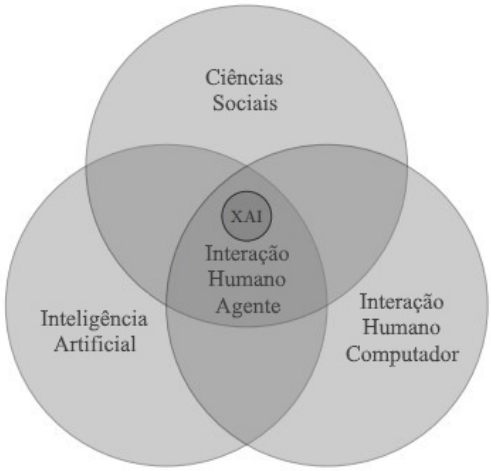


Figura 6 – Escopo da Inteligência Artificial Explicável (XAI). Fonte: Adaptado de (MILLER, 2017).

De acordo com Nazir, Dickson e Akram (2023), as técnicas de XAI visam fornecer informações adicionais sobre a decisão de um modelo de ML, melhorando assim a confiança nas suas decisões. Um modelo explicável é aquele que fornece explicações para suas previsões no nível humano para uma tarefa específica. Um modelo interpretável é aquele para o qual algumas conclusões podem ser tiradas sobre as partes internas/previsões do modelo (LALONDE, 2020).

Segundo Agarwal e Das (2020), o modelo de ML interpretável pode ser definido como uma extensão em que um humano pode compreender as decisões tomadas por

modelos de ML em seu processo de tomada de decisão. A importância do ML interpretável se deve às seguintes características:

- Viés: as previsões de modelos de ML não são tendenciosas e não discriminam contra turmas/grupos.
- Privacidade: A confidencialidade dos dados é assegurada através da interpretabilidade do modelo.
- Robustez: Garante que as previsões do modelo permaneceram consistentes e não têm alterações quando poucas alterações são feitas nos dados.
- Causalidade: garante que apenas relacionamentos causais sejam escolhidos.
- Confiabilidade: Os indivíduos podem facilmente confiar no sistema interpretável comparado ao sistema de "caixa-preta"(GILL, 2018).

Os **modelos agnósticos** são abordagens que visam fornecer explicações para o funcionamento de qualquer modelo de ML, independentemente de sua arquitetura ou algoritmo subjacente. São projetados para serem aplicáveis a uma ampla variedade de técnicas de ML, desde redes neurais profundas até algoritmos de árvore de decisão. A principal vantagem trazida por eles é a flexibilidade por não dependerem do funcionamento interno do sistema alvo (RIBEIRO, 2016).

Segundo Ribeiro, Singh e Guestrin (2016), o uso de modelos substitutos, que consistem em sistemas do tipo caixa-branca, de interpretabilidade conhecida e intuitiva, treinados para gerar aproximações do modelo preditivo podem ser **locais**, focados na análise de exemplos específicos, ou **globais**, visando explicar o modelo preditivo como um todo.

Para a condução deste trabalho, devido a flexibilidade dos modelos agnósticos com relação a independência da arquitetura do modelo de ML e por gerarem interpretações intuitivas, a exploração e aplicação das técnicas de XAI foram baseadas na análise de exemplos locais pelos métodos CAM, Grad-CAM, LIME, SHAP, Vanilla Gradients e SmoothGrad.

2.4 Interpretabilidade x Explicabilidade

Segundo Molnar (2022), é difícil, matematicamente, definir a interpretabilidade. Uma definição (não matemática) seria o grau em que um ser humano pode entender a causa de uma decisão ou o grau em que um ser humano pode prever consistentemente o resultado do modelo. Quanto maior a interpretabilidade de um modelo de ML, mais fácil é alguém compreender por que certas decisões ou previsões foram tomadas. Neste contexto, um modelo de ML é melhor interpretável que outro modelo se suas decisões

forem mais fáceis de serem compreendidas pelo ser humano do que as decisões do outro modelo.

De acordo com Lipton (2016), explicabilidade é uma noção contextual e não absoluta. Neste sentido, busca-se identificar as propriedades que são desejáveis para os sistemas interpretáveis, destacando a transparência, confiança e interpretabilidade *post-hoc*. Esse último relacionado à capacidade do sistema de oferecer informações úteis sobre seus resultados para os diversos perfis de usuários. Interpretabilidade é o grau em que um humano pode prever consistentemente o resultado do modelo.

2.5 Classificação dos Métodos de Interpretabilidade

Os métodos para interpretabilidade de ML podem ser classificados de acordo com vários critérios (MOLNAR, 2022).

2.5.1 Caixa-branca x Caixa-preta

Modelos contento combinações lineares, como a regressão linear e a regressão logistica, possuem estruturas mais simplificadas. Essa estrutura mais simplificada facilita o entendimento sobre o peso de cada parâmetro na predição do modelo. Esses algoritmos (regressão linear, regressão logística, modelos baseados em árvore, *k-nearest neighbors*, aprendizagem baseado em regras, modelos aditivos gerais e modelos bayesianos) são denominados caixa-branca, pois oferecem algum tipo de interpretabilidade em seus mecanismos (ARRIETA, 2019).

Em problemas mais complexos, onde um grande número de parâmetros são envolvidos, outros métodos se fazem necessário. Os algoritmos de caixa-preta são considerados opacos, uma vez que não é possível entender seus mecanismos de decisão. Dentre os modelos de caixa-preta destacam-se os métodos *ensemble*, redes neurais multicamadas, redes neurais profundas (CNNs e redes neurais recorrentes) (ARRIETA, 2019).

2.5.2 Interpretabilidade Intrínseca x *Post-hoc*

Esse critério distingue se a interpretabilidade é alcançada restringindo a complexidade do modelo de ML (intrínseco) ou aplicando métodos que analisam o modelo após o treinamento (*post-hoc*). Interpretabilidade intrínseca refere-se a modelos de ML que são considerados interpretáveis devido à sua estrutura simples, como árvores de decisão curtas ou modelos lineares esparsos. Interpretabilidade *post-hoc* refere-se à aplicação de métodos de interpretação após o treinamento do modelo. O recurso de permutação, também conhecido como importância de recursos (*feature importance*) ou importância de variáveis, é uma técnica utilizada em ML para avaliar a contribuição de cada variável de entrada

(recurso) no processo de tomada de decisão do modelo. Esse recurso é especialmente útil para entender quais variáveis são mais relevantes para a predição ou classificação de um determinado problema e pode ser considerado um método de interpretação *post-hoc*. Métodos *post-hoc* também podem ser aplicados a modelos intrinsecamente interpretáveis. Por exemplo, o recurso de permutação pode ser calculado para árvores de decisão (MOLNAR, 2022).

2.5.3 Escopo da Interpretabilidade

Um modelo de ML pode ser descrito como interpretável se for possível compreender o modelo inteiro de uma só vez (LIPTON, 2016).

A explicação da saída do **escopo global** requer a posse do modelo treinado, compreensão do algoritmo e conhecimento dos dados. Esse nível de interpretabilidade diz respeito à compreensão de como o modelo toma decisões, fundamentando-se em uma visão holística das variáveis ou atributos (que representam características do problema) e de cada componente aprendido, como pesos, outros parâmetros e estruturas. A interpretabilidade desse modelo global auxilia na compreensão da distribuição do resultado alvo com base nas suas variáveis ou atributos, o que se revela desafiador na prática. Qualquer espaço de características (variáveis ou atributos) com mais de três dimensões é considerado inimaginável para os seres humanos. Geralmente, quando as pessoas tentam compreender um modelo, elas consideram apenas partes dele, como os pesos em modelos lineares (MOLNAR, 2022).

É possível ampliar uma única instância, analisando as previsões do modelo para essa entrada e explicar o porquê. Ao focar em uma previsão individual, o comportamento do modelo complexo pode apresentar uma dinâmica mais compreensível e agradável (**escopo local**). Localmente, a previsão pode depender apenas linear ou monotonicamente de alguns recursos, em vez de ter uma dependência complexa deles. Explicações locais podem, portanto, ser mais precisas que explicações globais (MOLNAR, 2022).

2.5.4 Avaliação da Interpretabilidade

De acordo com Molnar (2022), não há consenso real sobre o que é interpretabilidade em ML. Também não está claro como medi-la. Entretano, na literatura, existem algumas pesquisas iniciais e uma tentativa de formular algumas abordagens para a avaliação da interpretabilidade.

No trabalho de Doshi-Velez e Kim (2017), são propostos três níveis principais para a avaliação da interpretabilidade:

- Avaliação do nível de aplicação (tarefa real): Coloque a explicação no produto e

faça com que seja testada pelo usuário final. Imagine um software de detecção de fraturas com um componente de ML que localiza e marca fraturas em raios-x. No nível da aplicação, os radiologistas testariam o software de detecção de fraturas diretamente para avaliar o modelo. Isso requer uma boa configuração experimental e um entendimento de como avaliar a qualidade. Uma boa linha de base para isso é sempre o quão bom um humano seria em explicar a mesma decisão.

- Avaliação do nível humano (tarefa simples): É uma avaliação simplificada do nível de aplicação. A diferença é que esses experimentos não são realizados com especialistas em domínio, mas com leigos. Isso torna os experimentos mais baratos e é mais fácil encontrar mais testadores. Um exemplo seria mostrar ao usuário explicações diferentes e ele escolheria a melhor.

- Avaliação do nível de função (tarefa de *proxy*): Não requer humanos. Isso funciona melhor quando a classe de modelo usada já foi avaliada por outra pessoa em uma avaliação no nível humano. Por exemplo, pode-se saber que os usuários finais entendem as árvores de decisão. Nesse caso, um *proxy* para a qualidade da explicação pode ser a profundidade da árvore. Árvores mais curtas obteriam uma melhor pontuação de explicação. Faria sentido acrescentar a restrição de que o desempenho preditivo da árvore permanece bom e não diminui muito em comparação com uma árvore maior.

2.6 Técnicas de Inteligência Artificial Explicável

As técnicas de XAI, ou técnicas interpretáveis, fornecem informações amigáveis ao ser humano com relação ao raciocínio por trás dos resultados e previsões dos modelos (AGARWAL; DAS, 2020). Vários métodos de interpretabilidade têm sido relatados na literatura. Os métodos explorados neste trabalho de pesquisa serão brevemente descritos nas próximas subseções.

2.6.1 CAM (*Class Activation Mapping*)

O CAM é uma técnica utilizada para interpretar e explicar as decisões tomadas por modelos de ML, especialmente em tarefas de classificação visual. Essa técnica é frequentemente aplicada em CNNs treinadas para classificar imagens, visando destacar as regiões das imagens mais relevantes para a classificação realizada pelo modelo.

O método CAM, proposto por Zhou, Khosla, Lapedriza, Oliva e Torralba (2016), produz mapas de ativação discriminativos a partir de camadas intermediárias da CNN utilizando a informação das últimas camadas convolucionais da rede para gerar um mapa de ativação específico para cada classe, destacando as regiões da imagem que mais con-

tribuíram para a classificação correspondente a essa classe. Os resultados experimentais mostram que o método CAM pode localizar com precisão as regiões discriminativas em imagens de diferentes tarefas de classificação, como reconhecimento de objetos e reconhecimento de cenas. Além disso, os autores mostram que a técnica CAM pode ser usada para melhorar a interpretabilidade dos modelos de CNN ao permitir a visualização das regiões mais relevantes para a classificação.

A Figura 7 exemplifica a aplicação do método CAM em uma imagem original (A), a qual apresenta tanto a presença de um cachorro quanto a de um gato. Neste exemplo, a técnica CAM foi aplicada com o objetivo de evidenciar as regiões da imagem que foram ativadas durante a classificação específica para cada categoria. Na imagem resultante (B), na qual a ativação se relaciona à categoria canina, nota-se uma sobreposição de áreas de interesse que correspondem à presença do cachorro na cena. Em contrapartida, na imagem (C), que representa a ativação do método para a categoria felina, as regiões destacadas concentram-se nas áreas associadas à presença do gato na fotografia. Este método proporciona uma visualização interpretável das partes específicas da imagem que influenciaram as decisões do modelo, oferecendo *insights* valiosos sobre o processo de classificação.

Figura 7 – Exemplo de aplicação da técnica CAM na imagem original (A) para classificação das categorias canina (B) e felina (C). Fonte: Adaptado de (MUHAMMAD; YEASIN, 2020).

2.6.2 Grad-CAM (*Gradient-weighted Class Activation Mapping*)

O Grad-CAM é uma extensão da técnica CAM e também é frequentemente aplicado em CNNs treinadas para classificar imagens. Destaca-se por oferecer interpretações mais refinadas e localizadas em comparação ao CAM original.

O método Grad-CAM, proposto por Selvaraju, Cogswell, Das, Vedantam, Parikh e Batra (2017), utiliza os gradientes das ativações da última camada convolucional da rede para gerar um mapa de ativação para cada classe. O Grad-CAM produz mapas de ativação mais precisos e localizados em relação a outros métodos existentes, que utilizam apenas as ativações da última camada convolucional, a exemplo do método CAM. Os resultados experimentais mostram que o Grad-CAM é capaz de gerar mapas de ativação precisos para

diferentes tarefas de classificação, como reconhecimento de objetos, detecção de faces e segmentação de imagens. Além disso, os autores mostram que o método Grad-CAM pode ser utilizado para detecção de objetos em imagens médicas, como mamografias.

A Figura 8 exemplifica a aplicação do método Grad-CAM em uma imagem original (A), retratando uma paisagem com a presença de dois elefantes. Neste caso, o Grad-CAM foi empregado para destacar regiões que evidenciam a presença dos animais. Diferentemente do CAM, o Grad-CAM utiliza gradientes para ponderar as ativações nas camadas convolucionais da CNN, resultando em uma representação mais refinada das áreas de interesse. A imagem resultante (B) apresenta uma máscara (mapa de calor) indicando as regiões cruciais para a previsão do modelo, enquanto a imagem (C) exibe a sobreposição do mapa de calor sobre a imagem original.

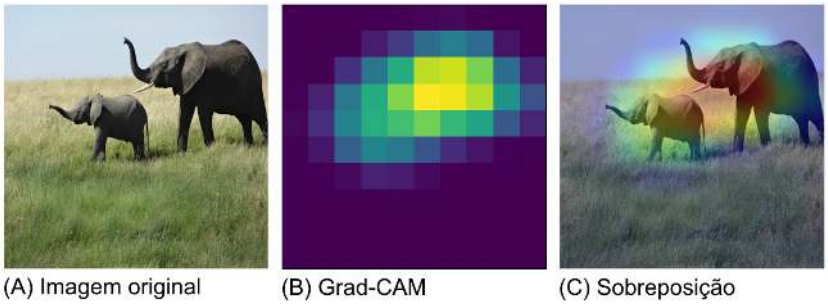


Figura 8 – Exemplo de aplicação da técnica Grad-CAM para identificação dos animais presentes na cena. Fonte: Disponível em https://keras.io/examples/vision/grad_cam/. Acessado em 06/03/2024.

2.6.3 LIME (*Local Interpretable Model-agnostic Explanations*)

O método LIME é uma técnica utilizada para interpretar e explicar as decisões tomadas por modelos de ML, independentemente do algoritmo subjacente. É um método de interpretabilidade local, o que significa que ele se concentra na explicação das previsões feitas para instâncias individuais de dados.

O método LIME, proposto por Ribeiro, Singh e Guestrin (2016), permite a geração de explicações para as predições de qualquer tipo de classificador. O método é baseado em uma abordagem de amostragem que permite a geração de explicações locais, mostrando quais características da entrada foram mais importantes para a tomada de decisão do modelo em cada caso específico. O método é modelo agnóstico, ou seja, pode ser aplicado em qualquer tipo de modelo de ML. A base do seu funcionamento é a geração de amostras perturbadas da entrada original e o treinamento de um modelo de interpretabilidade em cima dessas amostras. Em seguida, o modelo de interpretabilidade é utilizado para explicar a predição do modelo original em termos das características da entrada. Os resultados experimentais mostram que o método LIME é capaz de gerar explicações precisas e com-

preensíveis para diferentes tipos de modelos e tarefas de classificação, como classificação de texto e de imagens.

A Figura 9 exemplifica a aplicação do método LIME em uma imagem original (A), na qual são destacadas porções de pão assado em uma tigela. As imagens subsequentes (B) e (C) apresentam as explicações do LIME para as previsões associadas a dois rótulos, *"Bagel"* e "Morango". Na representação, a coloração verde indica que a presença dessa parte específica da imagem contribui para um aumento na probabilidade do rótulo, enquanto o vermelho indica uma redução.

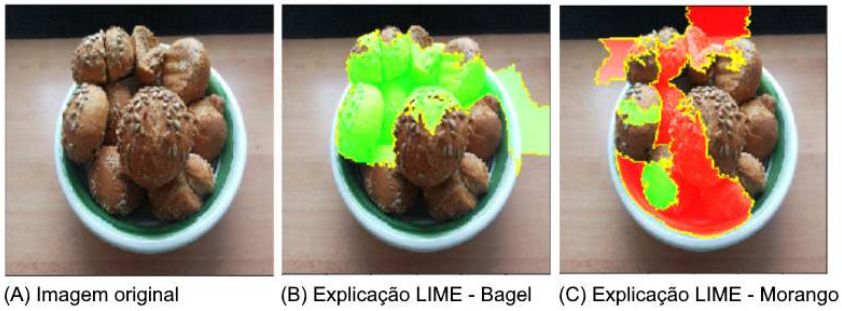


Figura 9 – Exemplo da técnica LIME aplicada na imagem orginal (A) para classificação das categorias *Bagel* (B) e Morango (C). Fonte: Adaptado de (MOLNAR, 2022).

2.6.4 SHAP (*SHapley Additive exPlanations*)

O método SHAP é uma técnica que visa explicar as previsões feitas por modelos de ML de forma global e/ou local. Ele se baseia no conceito de valores de *Shapley*, da teoria dos jogos, para atribuir importância às características (*features*) em uma previsão específica.

O método SHAP, proposto por Lundberg e Lee (2017), é uma abordagem modelo agnóstico, ou seja, é aplicável a diferentes tipos de modelos e é capaz de explicar as predições de forma global e/ou local, mostrando a importância de cada característica da entrada na tomada de decisão do modelo. Este método utiliza os valores *Shapley*, da teoria dos jogos, para calcular a contribuição de cada característica da entrada na predição do modelo. Os resultados experimentais mostram que o SHAP é computacionalmente eficiente e pode ser aplicado a modelos complexos, como redes neurais profundas. Ele é capaz de gerar explicações precisas e compreensíveis para diversos tipos de modelos e tarefas de classificação, podendo ainda ser utilizado para avaliar a importância de diferentes características de entrada e identificar possíveis vieses nos modelos.

A Figura 10 ilustra a aplicação da técnica SHAP em uma imagem (A) que representa um animal. As imagens subsequentes (B) e (C) exibem as explicações do SHAP para as previsões associadas a dois rótulos, "Suricata"e "Mangusto". Na representação, os pixels rosas indicam valores SHAP positivos que aumentam a probabilidade da classe,

enquanto os pixels azuis representam valores SHAP negativos, reduzindo a probabilidade da classe.

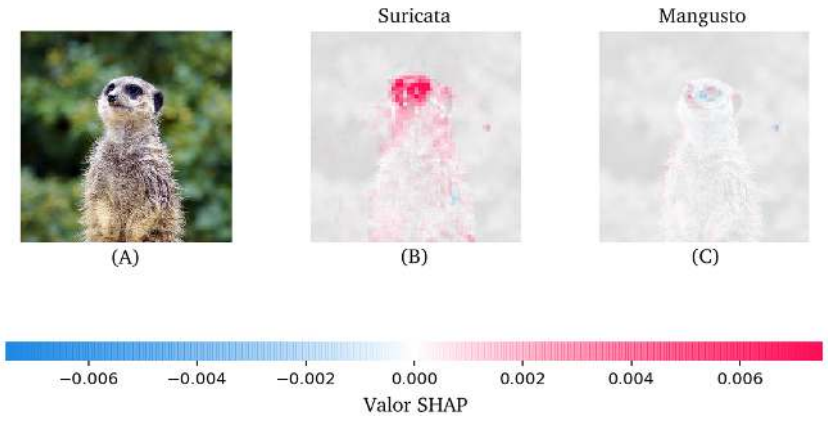


Figura 10 – Exemplo da técnica SHAP aplicada na imagem (A) para classificação das categorias Suricata (B) e Mangusto (C). Fonte: (Adaptado) Disponível em https://github.com/shap/. Acessado em 05/03/2024.

2.6.5 Vanilla Gradients

O método Vanilla Gradients é uma técnica que envolve a atualização iterativa dos pesos da rede proporcionalmente ao gradiente da função de perda. Apesar de sua simplicidade, pode ser sensível a hiperparâmetros e ter convergência lenta em cenários complexos, servindo como base para técnicas mais avançadas de otimização.

O método Vanilla Gradients, proposto por Simonyan, Vedaldi e Zisserman (2014), embora não seja originalmente desenvolvido para interpretabilidade visual, pode ser adaptado para fornecer *insights* sobre como o modelo toma decisões. Especificamente, ao aplicar essa técnica em uma tarefa de interpretabilidade visual, o objetivo é entender quais características específicas da entrada são mais influentes na decisão do modelo. Esse método é empregado com o propósito de elucidar as decisões do modelo, fornecendo *insights* sobre as características específicas dos dados de entrada que mais influenciam o processo de decisão. Em tal aplicação, a escolha de uma camada de interesse, geralmente associada à última camada antes da saída da rede, é crucial. Os gradientes são então calculados retropropagando a função de perda até a camada de interesse. Diferentemente da etapa de treinamento convencional, o foco recai sobre os gradientes em relação aos pixels da imagem de entrada. Esses gradientes são ponderados conforme sua importância, resultando em um mapa de saliência que destaca as regiões mais relevantes da imagem para as decisões do modelo. Assim, essa técnica proporciona uma abordagem inicial para interpretabilidade visual, embora ressalvas devam ser consideradas, especialmente em modelos complexos, onde relações intrincadas entre características podem dificultar a interpretabilidade direta.

A Figura 11 ilustra a aplicação da técnica Vanilla Gradients em uma imagem original (A) que retrata um cachorro da raça *Staffordshire Bull Terrier*. A imagem (B) exibe o mapa de gradientes, também conhecido como mapa de saliência, gerado pelo método Vanilla Gradients. Nesta representação, os pixels em tons avermelhados indicam as características distintivas para a previsão do modelo. A intensidade da tonalidade avermelhada está diretamente relacionada à força da característica, sendo que tonalidades mais intensas indicam características mais influentes para a predição do modelo.

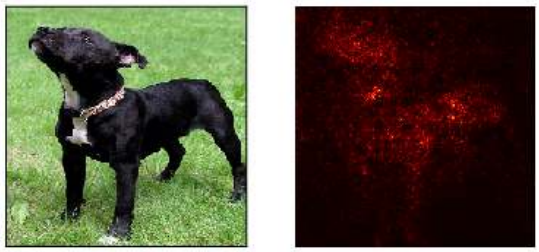


Figura 11 – Exemplo de aplicação da técnica Vanilla Gradients em uma imagem original (A) de um cachorro da raça *Staffordshire Bull Terrier*, evidenciando características distintivas em (B) relevantes para a predição do modelo. Fonte: Adaptado de (JERE, 2020).

2.6.6 SmoothGrad

O método SmoothGrad é uma abordagem mais avançada de interpretabilidade em ML. Ele suaviza os gradientes para proporcionar visualizações mais estáveis e confiáveis das regiões influentes nos dados de entrada, contribuindo para a transparência e explicabilidade em modelos complexos.

O método SmoothGrad, proposto por Smilkov, Thorat, Kim, Viégas e Wattenberg (2017), foi desenvolvido para melhorar a interpretabilidade dos modelos de ML, especialmente em redes neurais profundas, reduzindo a variabilidade nos mapas de calor de saliência gerados a partir dos gradientes. A ideia principal por trás deste método é introduzir um processo de suavização nos gradientes calculados. Isso é feito mediante a aplicação de uma técnica simples, mas eficaz: durante a geração do mapa de calor de saliência, múltiplas perturbações aleatórias são aplicadas ao dado de entrada e o gradiente é recalculado para cada versão perturbada. Em seguida, esses gradientes perturbados são combinados, geralmente por média, para criar um gradiente suavizado. Essa abordagem ajuda a reduzir o impacto de ruídos e flutuações nos gradientes, resultando em mapas de calor mais estáveis e consistentes. O SmoothGrad se mostra um método valioso em cenários nos quais é crucial compreender de forma confiável as características importantes que influenciam as decisões do modelo.

A Figura 12 ilustra a aplicação da técnica SmoothGrad em duas imagens originais (A), as quais representam um tubarão-branco e uma cobra-nariguda. Na imagem

(B), apresenta-se o mapa de gradientes, conhecido como mapa de saliência, gerado pelo método Vanilla Gradients. Já na imagem (C), exibe-se o mapa de gradientes gerado pelo método SmoothGrad. Nesta representação, destaca-se a notável redução de ruído e maior estabilidade proporcionadas pelo SmoothGrad em comparação com o Vanilla Gradients, revelando características mais consistentes e interpretações mais claras das regiões de influência na predição do modelo.

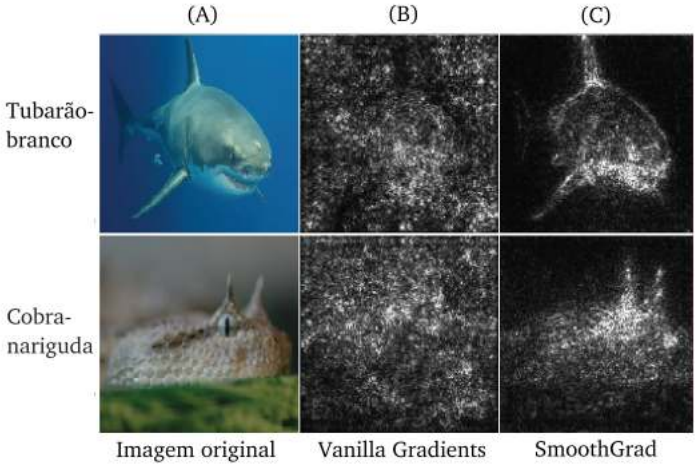


Figura 12 – Exemplo do método SmoothGrad aplicado em uma imagem de um tubarão-branco e uma cobra-nariguda em (A), com destaque para a notável redução de ruído na interpretação das regiões influentes do modelo, evidenciada nos mapas de gradiente gerado pelo SmoothGrad (C). Fonte: Adaptado de (SMILKOV, 2017).

3 Trabalhos Relacionados

Expostos os conceitos teóricos sobre a área de IA e sobre a anatomia do nervo óptico e sua relação com o glaucoma, este capítulo apresenta o conjunto de trabalhos mais relacionados à essa pesquisa, divididos em duas abordagens: (1) a utilização de CNNs aplicadas em imagens de fundo de retina para o diagnóstico do glaucoma; e (2) a aplicação de técnicas de XAI na análise de imagens médicas baseada no DL. As revisões de literatura apresentadas por Camara, Neto, Pires, Villasana, Zdravevski e Cunha (2022) e van der Velden, Kuijf, Gilhuijs e Viergever (2022) sintetizam os estudos relacionados a ambas as abordagens, orientando assim a condução deste trabalho. A seguir, são apresentadas as principais propostas de cada uma delas e seu relacionamento com o desenvolvimento desta pesquisa.

3.1 CNNs Aplicadas em Imagens de Fundo de Retina para Diagnóstico do Glaucoma

Diversos estudos têm empregado técnicas baseadas em CNNs para a classificação de imagens de fundo de retina no contexto do diagnóstico de glaucoma (SHYAMALEE; MEEDENIYA, 2022).

A proposta de Bajwa, Malik, Siddiqui, Dengel, Shafait, Neumeier e Ahmed (2019) consiste em uma estrutura de dois estágios. O primeiro estágio tem como objetivo detectar e localizar o disco óptico em uma imagem de fundo de retina. Já o segundo estágio utiliza uma CNN para classificar o disco óptico extraído como saudável ou glaucomatoso. O método foi avaliado em sete conjuntos de dados públicos para localização de discos, e no conjunto de dados ORIGA para a classificação de glaucoma. Os resultados da localização automática do disco óptico estabeleceram um novo estado da arte em seis conjuntos de dados, alcançando uma precisão de 100% em quatro deles. Quanto à classificação do glaucoma, a CNN obteve uma AUC de 87,4%, precisão de 77,97%, sensibilidade de 79,38% e *f1-score* de 77,88%.

Sreng, Maneerat, Hamamoto e Win (2020) propuseram um método de triagem para o glaucoma em dois estágios. Inicialmente, a segmentação do disco óptico é realizada por meio da arquitetura DeepLabv3+. Em seguida, uma rede neural profunda pré-treinada é utilizada para a classificação das imagens. O método foi avaliado em cinco conjuntos de dados publicamente disponíveis (RIM-ONE, ORIGA, DRISHTI-GS1, REFUGE e ACRIMA), totalizando 2787 imagens de retina. A combinação de DeepLabv3+ e MobileNet demonstrou ser a melhor opção para a segmentação do disco óptico. Para a

classificação do glaucoma, nos conjuntos de dados RIM-ONE, ORIGA, DRISHTI-GS1 e ACRIMA, foram alcançadas precisões de 97,37%, 90,00%, 86,84% e 99,53%, com AUCs de 100%, 92,06%, 91,67% e 99,98%, respectivamente. No conjunto de dados REFUGE, obteve-se uma precisão de 95,59% e AUC de 95,10%.

Em Aziz-ur-Rehman, Taj, Sajid e Karimov (2021), uma arquitetura de CNN profunda em dois estágios para a classificação do glaucoma foi proposta. No primeiro estágio, quatro arquiteturas de CNN diferentes, pré-treinadas pela ImageNet (AlexNet, Inceptionv3, InceptionResNetV2 e NasNet-Large), foram empregadas. NasNet-Large apresentou os melhores resultados, com sensibilidade, especificidade, precisão e AUC de 99,1%, 99,4%, 99,3% e 97,8%, respectivamente. Os conjuntos de dados públicos ACRIMA, ORIGA-Light, RIM-ONE e dois conjuntos de dados privados (AFIO e HMC) foram utilizados pelos autores. Na segunda etapa, os pesquisadores propuseram um classificador em conjunto com duas novas técnicas de montagem (votação ponderada com base na precisão e média ponderada com base na precisão/pontuação) para aprimorar ainda mais os resultados da classificação do glaucoma. Demonstrou-se que o conjunto com esquema baseado em precisão/pontuação melhora a precisão para diversos bancos de dados, atingindo 99,5%. Como resultado deste estudo, foi evidenciado que a arquitetura NasNet-Large é uma opção viável em termos de desempenho como um único classificador, enquanto o classificador em conjunto melhora ainda mais o desempenho generalizado para a classificação automática do glaucoma.

No estudo conduzido por Diaz-Pinto, Morales, Naranjo, Köhler, Mossi e Navea (2019), cinco modelos distintos de CNNs, treinados com os pesos do ImageNet (InceptionV3, Xception, VGG16, VGG19 e ResNet50), para a avaliação automática do glaucoma por meio de imagens de fundo de retina foram investigados. Utilizando cinco bancos de dados públicos, totalizando 1707 imagens, a arquitetura Xception alcançou uma AUC média de 96,05%, uma especificidade média de 85,80% e uma sensibilidade média de 93,46%. Adicionalmente, foi disponibilizado ao público um novo banco de dados clínico chamado ACRIMA, contendo 705 imagens rotuladas, sendo 396 imagens glaucomatosas e 309 imagens normais.

Em Chai, Liu e Xu (2018), os autores investigaram o uso de modelos de aprendizado profundo para integrar o conhecimento do domínio às imagens do fundo da retina, visando o diagnóstico automático de glaucoma. Esse conhecimento do domínio engloba medidas essenciais para o diagnóstico da condição e a identificação de regiões cruciais na imagem que contêm informações significativas. Para aproveitar plenamente esse conhecimento do domínio e extrair simultaneamente recursos ocultos da imagem, foi projetada uma rede neural multiramificada (MB-NN) com métodos para a extração automática de áreas importantes nas imagens e obtenção de recursos de conhecimento do domínio. A eficácia do modelo proposto foi avaliada em conjuntos de dados reais, registrando uma

precisão de 91,51%, sensibilidade de 92,33% e especificidade de 90,90%.

Gómez-Valverde, Antón, Fatti, Liefers, Herranz, Santos, Sánchez e Ledesma-Carbayo (2019) exploraram os modelos VGG19, GoogleNet, ResNet50 e DENet para a classificação automática do glaucoma. Neste estudo, os autores compararam o desempenho da TL e do treinamento do zero com esses modelos. Foram utilizadas 2313 imagens de fundo de retina provenientes de três bancos de dados distintos: RIM-ONE, DRISHTI-GS e Esperanza (um conjunto de dados privado). No banco de dados RIM-ONE, as imagens classificadas como suspeitas foram consideradas para o estudo como glaucomatosas. O melhor desempenho registrado foi alcançado utilizando a técnica de aprendizado por transferência com a arquitetura VGG19, atingindo uma AUC de 0,94.

No trabalho de Martins, Cardoso e Soares (2020), um pipeline de diagnóstico interpretável auxiliado por computador, capaz de diagnosticar o glaucoma por meio de imagens de fundo de retina, operando offline em dispositivos móveis é proposto. Diversos conjuntos de dados públicos de imagens de fundo de retina (ORIGA, DRISHTI-GS, iChallenge, RIM-ONE - r1, r2 e r3, e RIGA) foram combinados e utilizados para construir CNNs que desempenham tarefas de segmentação e classificação. Em relação à classificação, o modelo alcançou uma precisão de 0,87, uma sensibilidade de 0,85 e uma AUC de 0,93.

Em Bajwa, Singh, Neumeier, Malik, Dengel e Ahmed (2020), os autores conduziram experimentos para o diagnóstico automatizado de glaucoma, realizando a segmentação do disco óptico e copo óptico em diferentes conjuntos de dados, incluindo G1020 e ORIGA. Os métodos implementados foram a CNN baseada na região - *Region-based Convolutional Neural Network* (R-CNN), Restnet50 e InceptionV3, resultando em um *f1-score* de 88,6%.

Em Serener e Serte (2019), os modelos ResNet50 e GoogleNet foram selecionados e treinados com dois bancos de dados públicos: um banco de dados do *Kim's Eye Hospital* (um total de 1542 imagens, incluindo 786 fotos de pacientes normais e 756 de pacientes com glaucoma) e RIM-One-r3. Todas as imagens de fundo foram equalizadas com histograma, e o banco de dados do *Kim's Eye Hospital* foi utilizado para treinar os dois modelos. Para a avaliação de desempenho, os modelos foram testados com o banco de dados RIM-One-r3. Nesse trabalho, ficou demonstrado que o modelo GoogleNet supera o ResNet50 para a detecção precoce e avançada de glaucoma.

Norouzifard, Nemati, GholamHosseini, Klette, Nouri-Mahdavi e Yousefi (2018) empregaram uma rede de aprendizado profundo utilizando a TL com os pesos pré-treinados do ImageNet. Os dois modelos escolhidos foram o VGG19 e o InceptionResNetV2. Neste estudo, dois conjuntos de dados foram utilizados: um da Universidade da Califórnia em Los Angeles (UCLA) e outro, disponível ao público, o HRF. Os autores selecionaram aleatoriamente 70% das imagens do banco de dados da UCLA para treinamento, 25% para validação e os 5% restantes para teste. Os modelos foram, posteriormente, retestados com

o banco de dados HRF para reforçar os resultados. Notavelmente, o modelo Inception-ResNetV2 apresentou uma especificidade e sensibilidade superiores a 90% para o banco de dados da UCLA.

Os autores Li, Song, Chen, Xiong, Li, Zhong, Tang, Fan, Lam, Pan, Zheng, Li, Qu, He, Wang, Jin, Zhou, Song, Sun, Cheng, Yang, Fan, Li, Zhang, Yuan, Xu, Xiong, Jin, Lv, Niu, Liu, Li, Zhang, Zangwill, Frangi, Aung, Cheng, Qiao, Zhang e Ting (2020) avaliaram um sistema de DL baseado em aplicativos para smartphones (iGlaucoma) com o propósito de detectar alterações glaucomatosas no campo visual. O aplicativo móvel utiliza uma CNN com a arquitetura ResNet18 modificada para realizar a classificação do glaucoma. Os pesquisadores empregaram um conjunto de dados privado, no qual relataram uma AUC de 87,3%.

Em seu estudo, Camara, Neto, Pires, Villasana, Zdravevski e Cunha (2022) conclui que a análise de imagens de fundo de retina por meio de métodos de aprendizado profundo proporciona uma maior precisão no rastreamento do glaucoma. O autor ainda observa que as metodologias que utilizam bancos de imagens públicos para a triagem do glaucoma precisam de um aumento constante na disponibilidade de dados, de modo a abranger as diversas morfologias normais e patológicas associadas à doença.

A revisão bibliográfica realizada enfatiza o sucesso dos modelos baseados em DL, especialmente as CNNs, ao obter resultados notáveis em métricas de desempenho como acurácia, precisão, sensibilidade, *f1-score* e AUC. Os valores correspondentes a essas métricas estão resumidos na Tabela 1, consolidando as informações destacadas na revisão bibliográfica.

Trabalho/Autor	**A**	**P**	**S**	***F1-Score***	**AUC**
(CHAI, 2018)	-	91,51%	92,33%	-	-
(BAJWA, 2019)	91,51%	77,97%	79,38%	77,88%	87,4%
(DIAZ-PINTO, 2019)	-	-	96,05%	-	96,05%
(GóMEZ-VALVERDE, 2019)	-	-	-	-	94,00%
(SRENG, 2020)	-	99,59%	-	-	100,00%
(MARTINS, 2020)	-	87,00%	85,00%	-	93,00%
(BAJWA, 2020)	-	-	-	88,60%	-
(LI, 2020)	-	-	-	-	87,30%
(Aziz-ur-Rehman, 2021)	-	97,8%	99,1%	-	-

Tabela 1 – Síntese das métricas de desempenho destacadas na revisão da literatura, onde A, P e S representam Acurácia, Precisão e Sensibilidade, respectivamente.

Esses estudos se alinham ao escopo desta pesquisa ao treinar e avaliar diversos modelos de CNNs para a previsão do glaucoma por meio da classificação de imagens de retinografia. No entanto, identificou-se uma lacuna nesses trabalhos, uma vez que não abordam a questão da interpretabilidade, ou seja, a capacidade de compreender o processo de tomada de decisão pelos modelos. Portanto, a pesquisa aqui proposta busca

preencher essa lacuna, explorando não apenas a eficácia, mas também a interpretabilidade dos modelos.

3.2 Técnicas de XAI na Análise de Imagens Médicas com Base no DL

O trabalho conduzido por van der Velden, Kuijf, Gilhuijs e Viergever (2022) apresenta uma visão geral da XAI, a qual tem sido empregada na análise de imagens médicas com base no DL. Os autores discorrem sobre a importância da transparência e interpretabilidade em sistemas de IA aplicados à análise de imagens médicas. Inicialmente, apresentam os conceitos básicos de XAI e suas aplicações específicas na análise de imagens médicas. Em seguida, abordam a importância de se ter modelos explicáveis e interpretáveis na área médica, enfatizando a necessidade de os sistemas de IA fornecerem justificativas claras e compreensíveis para suas decisões. Uma revisão de literatura é realizada destacando quais técnicas XAI são empregadas na análise de imagens médicas classificando-as em três tipos: visual, textual e baseado em exemplos.

O estudo conduzido por Olden, Joy e Death (2004) compara diferentes métodos para quantificar a importância de variáveis em RNAs, utilizando dados simulados. Os autores geraram conjuntos de dados simulados com diferentes níveis de complexidade, incluindo dados com relações lineares e não lineares entre as variáveis. Em seguida, treinaram diferentes tipos de RNAs em cada conjunto de dados e avaliaram a importância das variáveis por meio de quatro métodos distintos: permutação de recurso, saliência de recurso, análise de sensibilidade global e decomposição SHAP. Os resultados indicaram que a permutação de recurso e a saliência de recurso produziram resultados semelhantes e consistentes em todos os conjuntos de dados simulados. Por outro lado, a análise de sensibilidade global e a decomposição SHAP mostraram diferenças significativas em sua capacidade de quantificar a importância das variáveis em diferentes tipos de dados. Os autores concluem que a permutação de recurso e a saliência de recurso são métodos confiáveis para avaliar a importância das variáveis em RNAs, enquanto a análise de sensibilidade global e a decomposição SHAP podem ser mais adequadas para tipos específicos de dados e tarefas de modelagem.

O estudo conduzido por Tsang, Cheng e Liu (2018) aborda a detecção de interações estatísticas entre variáveis em RNAs, realizando uma análise dos pesos das conexões entre as camadas. Neste trabalho, os autores propõem um método para identificar interações estatísticas entre pares de variáveis, baseado em uma medida de dependência mútua denominada *Distance Correlation* (DC). A partir dos pesos da RNA, o método estima a matriz de correlação de DC entre as variáveis e utiliza um teste de hipótese para identificar quais pares de variáveis possuem interações significativas. Os resultados experimentais

demonstram que o método proposto consegue detectar interações estatísticas com alta precisão em dados sintéticos e em dados reais de diversas áreas, como finanças e genômica. Além disso, os autores destacam que a detecção de interações estatísticas pode aprimorar a interpretabilidade e a precisão de modelos de redes neurais em tarefas de predição.

O trabalho de Fong e Vedaldi (2017) apresenta o método denominado *Meaningful Perturbation* (MMP), o qual possibilita a geração de explicações interpretáveis para modelos de ML opacos, conhecidos como caixas-pretas. O MMP consiste em um método de perturbação que modifica os dados de entrada do modelo, de modo a promover uma mudança significativa em relação à predição original. Essa abordagem permite testar o modelo com novas entradas semanticamente similares às originais, porém com diferenças específicas que viabilizam a análise da importância das características. O método proposto é aplicável a diferentes tipos de modelos de aprendizado de máquina, incluindo redes neurais profundas e árvores de decisão, e pode ser utilizado para gerar explicações tanto globais quanto locais. Os resultados experimentais evidenciam que o MMP é capaz de gerar explicações precisas e compreensíveis para uma variedade de modelos e tarefas de classificação, abrangendo desde a classificação de imagens até o diagnóstico médico.

No estudo conduzido por Meng, Hashimoto e Satoh (2020), os autores empregaram um mapa de ativação de classe ponderado por gradiente (Grad-CAM), mineraram a atenção aplicada - *Attention Mining* (AM) - com base nos resultados do Grad-CAM e implementaram a perda de dissimilaridade - Dissimilarity (DISSIM) - no processo de treinamento. Utilizando um conjunto de dados privado proveniente de 13 universidades, complementaram o estudo com o uso de CNNs profundas - *Deep Convolutional Neural Networks* (DCNNs) - e o modelo VGG19 para alcançar um reconhecimento preciso de glaucoma, atingindo uma precisão de 96,2%.

O estudo comparativo apresentado por Ahmad, Kasukurthi e Pande (2019) analisa diferentes arquiteturas de DL treinadas em um conjunto de dados proprietário de imagens de retinopatia diabética e testadas no conjunto de dados Messidor-2, disponível publicamente. Utilizando o CAM para a localização das anormalidades nas imagens, os autores realizaram uma comparação entre as diversas arquiteturas. Os resultados indicam que, para a tarefa de classificação, à medida que o número de parâmetros da arquitetura aumenta, os modelos apresentam um desempenho superior, com o NASNet alcançando maior precisão. No entanto, para a tarefa de localização, a arquitetura VGG19 superou todos os outros modelos.

No estudo conduzido por Jang, Son, Park, Park e Jung (2018), os autores apresentam uma análise das saídas de um modelo de CNN dedicado à classificação da lateralidade de imagens de fundo de retina[1]. O modelo proposto não apenas automatiza o processo

[1] A lateralidade refere-se à determinação da posição ou orientação de uma imagem de fundo de retina em relação à estrutura ocular ou anatômica do olho.

de classificação, resultando na redução da carga de trabalho dos médicos, mas também destaca as principais regiões da imagem por meio de mapas de ativação de classe baseados em gradiente (Grad-CAM). O modelo foi treinado e testado utilizando 25.911 imagens de fundo de retina, distribuídas da seguinte forma: 43,4% centradas na mácula e 28,3% para imagens superiores e nasais, respectivamente. Os resultados indicam que o modelo proposto alcançou uma precisão média de treinamento de 99%, com ativações significativas detectadas no local do disco óptico e nos vasos sanguíneos da retina ao redor do disco. Os autores concluem que a visualização de regiões informativas, juntamente com a apresentação do resultado da previsão, aumentaria a interpretabilidade do modelo neural, beneficiando os médicos no uso do sistema de classificação automática.

Em Costa, Araújo, Aresta, Galdran, Mendonça, Smailagic e Campilho (2019), os autores propõem o *EyeWeS*, um método que detecta a retinopatia diabética em imagens de fundo de retina e identifica as regiões da imagem que contêm lesões, sendo treinado apenas com os rótulos das imagens. O método proposto melhorou os resultados do InceptionV3 de 94,9% para 95,8%, com relação a AUC, mantendo apenas aproximadamente 5% do número de parâmetros do InceptionV3. Os autores concluem, relatando que o mesmo modelo é capaz de atingir 97,1% de AUC em um experimento com conjuntos de dados cruzados.

Em Kumar, Taylor e Wong (2019), os autores propõem o CLEAR-DR, um novo sistema CAD interpretável baseado na noção de radiômica de descoberta de resposta atenta aprimorada por classe, para fins de suporte à decisão clínica para retinopatia diabética. Além da classificação de doenças através do sequenciador radiômico profundo descoberto, o sistema CLEAR-DR também produz uma interpretação visual do processo de tomada de decisão para fornecer uma melhor compreensão do sistema. A eficácia e utilidade do sistema proposto para melhorar a interpretabilidade dos resultados da classificação diagnóstica para a aplicação da classificação da retinopatia diabética são demonstradas no conjunto de dados de retinopatia diabética da plataforma Kaggle. Os autores concluem que a abordagem proposta tem grande potencial para reduzir a variabilidade inter e intra-observador e acelerar o processo geral de triagem e diagnóstico, melhorando a consistência e a precisão em ambientes clínicos.

Em Thakoor, Li, Tsamis, Sajda e Hood (2019), modelos de CNN para a detecção de glaucoma com base em OCT e na camada de fibra nervosa da retina - *Retinal Nerve Fiber Layer* (RNFL) - são descritos e avaliados. Nos experimentos, todos os modelos demonstraram alta precisão na detecção de glaucoma, e a implementação de mapas de calor (Grad-CAM) baseados em atenção das regiões de interesse da CNN sugere que esses modelos podem ser aprimorados pela incorporação de informações de localização dos vasos sanguíneos. Os autores concluem que esses modelos de CNN têm o potencial de colaborar com especialistas humanos para manter a saúde ocular geral e agilizar a

detecção de doenças oculares causadoras de cegueira.

Zhou, Gao, Cheng, Gu, Fu, Tu, Yang, Zhao e Liu (2020) propõem uma nova estrutura para detecção de anomalias em imagens de OCT, denominada Rede Adversarial Generativa com Restrição ao *Sparsity* (Sparse-GAN), para triagem de doenças, onde apenas dados saudáveis estão disponíveis no conjunto de treinamento. Um mapa de ativação de anomalias, gerado através do CAM, exibe o mapa de calor das lesões. O método proposto é avaliado em um conjunto de dados disponível ao público, e os resultados demonstram superação em relação aos métodos de ponta.

Em Li, Xu, Liu, Li, Wang, Jiang, Wang, Fan e Wang (2020), os autores propõem uma CNN baseada na atenção para a detecção de glaucoma, denominada AG-CNN (*Attention Guided Convolutional Neural Network*). Nesse trabalho, estabelece-se um extenso banco de dados de glaucoma baseado em atenção em larga escala - *Large-scale Attention based Glaucoma* (LAG), contendo 11.760 imagens de fundo de retina, rotuladas como positivas para glaucoma (4.878 imagens) ou negativas para glaucoma (6.882 imagens). Os mapas de atenção de 5.824 imagens são obtidos por oftalmologistas por meio de um experimento simulado de rastreamento ocular. Uma nova estrutura do AG-CNN é projetada, incluindo uma sub-rede de previsão de atenção, uma sub-rede de localização de área patológica e uma sub-rede de classificação de glaucoma. Os mapas de ativação de classe são previstos na sub-rede de previsão de atenção, destacando as regiões mais importantes para a detecção de glaucoma. Os resultados dos testes realizados no conjunto de dados LAG e em outro banco de dados público, RIM-ONE, mostram que a abordagem AG-CNN proposta avança significativamente o estado da arte na detecção de glaucoma.

Conforme evidenciado na revisão de literatura, a explicação visual, também chamada de mapeamento de saliência, tem se destacado como a forma mais comum de XAI que tem sido aplicada na análise de imagens médicas (van der Velden, 2022). Uma síntese das técnicas de XAI utilizadas nesses estudos está apresentada na Tabela 2. Neste contexto, os trabalhos conduzidos por Meng, Hashimoto e Satoh (2020), Ahmad, Kasukurthi e Pande (2019), Li, Xu, Liu, Li, Wang, Jiang, Wang, Fan e Wang (2020) apresentam uma área de interseção com o escopo desta pesquisa, uma vez que empregam técnicas de XAI na análise de imagens de retinografia para o diagnóstico de glaucoma. No entanto, é relevante salientar que esses métodos de XAI são aplicados de maneira isolada em cada uma das soluções propostas, caracterizando uma abordagem individualizada em sua implementação. Dentro desse contexto, esta pesquisa visa preencher uma lacuna ao explorar diferentes técnicas de XAI aplicadas em diversas arquiteturas de CNNs, propondo uma análise comparativa para avaliar qual abordagem oferece os melhores recursos para interpretabilidade humana e apoio ao diagnóstico de glaucoma. Este estudo mais abrangente busca identificar sintonias entre as técnicas de XAI e as arquiteturas de CNNs, contribuindo para avanços significativos na interpretabilidade e confiabilidade dos modelos

aplicados à área médica.

Trabalho/Autor	**Técnicas de XAI**
(OLDEN, 2004)	Permutação de recurso, saliência de recurso, análise de sensibilidade global e decomposição SHAP.
(FONG; VEDALDI, 2017)	*Meaningful Perturbation* (MMP).
(TSANG, 2018)	Método para identificar interações estatísticas entre pares de variáveis - *Distance Correlation* (DC).
(JANG, 2018)	Grad-CAM aplicado à classificação da lateralidade de imagens de fundo de retina.
(AHMAD, 2019)	CAM para a localização das anormalidades nas imagens de retinopatia diabética.
(COSTA, 2019)	*EyeWeS*, um método que detecta a retinopatia diabética em imagens de fundo de retina e identifica as regiões da imagem que contêm lesões, sendo treinado apenas com os rótulos das imagens.
(KUMAR, 2019)	CLEAR-DR, um novo sistema CAD interpretável baseado na noção de radiômica de descoberta de resposta atenta aprimorada por classe, para fins de suporte à decisão clínica para retinopatia diabética.
(THAKOOR, 2019)	Grad-CAM baseados em atenção das regiões de interesse da CNN sugere o aprimoramento dos modelos pela incorporação de informações de localização dos vasos sanguíneos.
(MENG, 2020)	Grad-CAM, mineraram a atenção aplicada (AM) com base nos resultados do Grad-CAM e perda de dissimilaridade (DISSIM).
(ZHOU, 2020)	CAM aplicado em uma nova estrutura para detecção de anomalias em imagens de OCT denominado Sparse-GAN.
(LI, 2020)	CNN baseada na atenção para a detecção do glaucoma, denominada AG-CNN. Aplicação do CAM para destacar regiões importantes para a detecção da doença.

Tabela 2 – Síntese das técnicas de XAI destacadas na revisão da literatura.

4 Metodologia e Experimentos

Neste capítulo, apresenta-se a estratégia metodológica adotada para conduzir a pesquisa. Nas seções seguintes, são destacadas a síntese (resumo) da abordagem metodológica, o ambiente utilizado na realização dos experimentos, os conjuntos de dados empregados, bem como os instrumentos e técnicas aplicados ao longo do desenvolvimento deste estudo.

4.1 Síntese da Abordagem Metodológica

A abordagem metodológica deste trabalho foi dividida em quatro etapas principais, conforme ilustrado no fluxograma representado pela Figura 13.

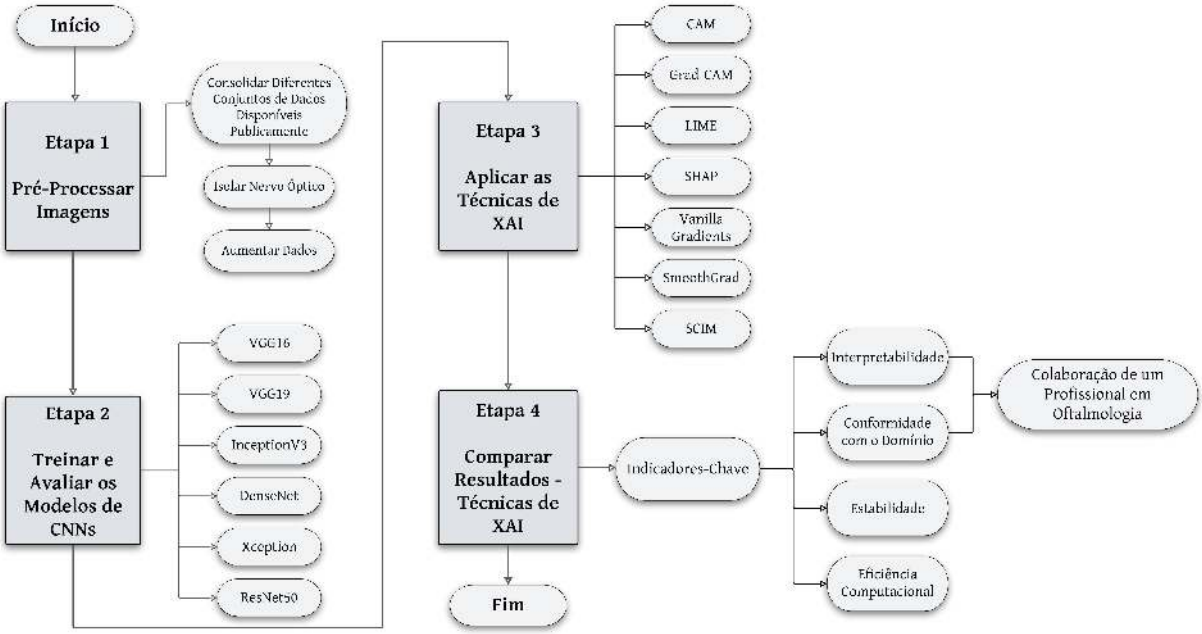


Figura 13 – Fluxograma da abordagem metodológica. Fonte: Elaborado pelo autor.

Na fase inicial, de pré-processamento das imagens (Etapa 1), diversos conjuntos de dados disponíveis publicamente foram consolidados em um único conjunto de dados. Foi realizado o isolamento do nervo óptico nas imagens e o conjunto de dados consolidado foi submetido à técnica de aumento de dados - *Data Augmentation* (DA). Para uma compreensão mais aprofundada, detalhes mais abrangentes sobre a Etapa 1 serão fornecidos na seção 4.4 (Pré-Processamento das Imagens).

Na fase subsequente, de treinamento e avaliação dos modelos de CNNs (Etapa 2), foram implementadas e avaliadas diferentes arquiteturas de CNNs, incluindo a VGG16,

VGG19, InceptionV3, Xception, DenseNet e ResNet50.

Na Etapa 3, dedicada à exploração e implementação de técnicas de XAI, cada modelo de CNN treinado foi submetido a diversas técnicas de interpretabilidade, incluindo as técnicas modelo-agnósticas locais: CAM, Grad-CAM, LIME, SHAP, Vanilla Gradients e SmoothGrad. Adicionalmente, foi concebida uma nova abordagem denominada SCIM, consistindo na integração das regiões confluentes e relevantes identificadas pelo SHAP e pelo CAM.

A escolha estratégica do SHAP e do CAM como base para a abordagem SCIM se fundamenta em suas características únicas para interpretação visual de modelos aplicados a imagens. O SHAP destaca-se pela capacidade de atribuir valores de contribuição a cada pixel, proporcionando insights detalhados sobre a influência de diferentes regiões na decisão do modelo. Com uma interpretação pixel a pixel, o SHAP oferece uma compreensão refinada das características visuais consideradas mais relevantes. Por sua vez, o CAM, ao visualizar espacialmente as regiões mais influentes na predição da classe, oferece uma abordagem intuitiva. Mapeando as áreas ativadas durante o processo decisório do modelo, o CAM facilita a identificação de padrões visuais cruciais para a classificação.

A origem dessa nova abordagem para interpretabilidade visual em XAI está fundamentada no raciocínio derivado da experiência do autor na análise gráfica de ativos financeiros, abrangendo bolsa de valores e criptomoedas. Inspirado pela prática de utilizar indicadores técnicos na análise gráfica para embasar decisões de compra ou venda de ativos, o autor reconheceu a importância da convergência de múltiplos indicadores para reforçar a confiança na tomada de decisão. Analogamente, na abordagem SCIM, buscou-se a confluência de regiões com o propósito de alcançar uma interpretabilidade mais robusta e confiável.

Assim como na análise gráfica, onde a convergência de indicadores proporciona maior segurança ao analista, a abordagem SCIM propõe a convergência de regiões interpretativas dessas técnicas (SHAP e CAM), visando potencializar a confiabilidade das conclusões obtidas. Esse racional inovador oferece uma perspectiva promissora para aprimorar a interpretabilidade em XAI, contribuindo para a confiança e uma compreensão mais aprofundada dos modelos.

Por fim, na fase de comparação dos resultados obtidos mediante a aplicação das técnicas de XAI (Etapa 4), conduziu-se uma análise comparativa das interpretações geradas por essas técnicas nas diferentes arquiteturas de CNNs. Para fins de comparação, foram estabelecidos e examinados indicadores específicos, conforme descritos em Molnar (2022): Interpretabilidade, Conformidade com o Domínio, Estabilidade e Eficiência Computacional.

Para a avaliação dos indicadores de Interpretabilidade e Conformidade com o Do-

mínio, contou-se com a *expertise* de um profissional da área de oftalmologia, que possui mais de 7 anos de experiência no diagnóstico e tratamento do glaucoma. Este especialista foi selecionado para realizar uma avaliação de nível humano (tarefa simples), conforme descrito por Doshi-Velez e Kim (2017), utilizando um questionário específico, o qual está disponível no Apêndice A desta dissertação.

Nas próximas seções, são apresentados detalhes mais aprofundados acerca dos procedimentos realizados em cada uma das etapas da metodologia.

4.2 Ambiente

O ambiente empregado na realização dos experimentos possui a seguinte configuração:

- Processador Intel(R) Core(TM) i5 - 6500 CPU @ 3.20 GHz;
- 8GB de memória ram;
- Ssd de 512gb;
- Sistema operacional Windows 10 Pro de 64 bits; e
- Visual Studio Code Versão 1.77.3 (Python 3.10.4).

O autor utilizou um monitor HP 22mp55py-b V225hz LCD *Wide* 21.5 polegadas para as análises visuais durante a pesquisa. Este monitor possui resolução *widescreen*, destacando-se pela taxa de atualização de 225Hz, crucial para a reprodução fluida de imagens dinâmicas. A tecnologia LCD (*Liquid Crystal Display*) garante reprodução de cores vibrante e fiel, enquanto a configuração *widescreen* favorece a exibição simultânea de informações. Ressalta-se que a configuração do monitor do profissional em oftalmologia durante a sua avaliação não foi determinada.

4.3 Conjunto de Dados

A escassez de grandes conjuntos de dados de imagens de fundos de retina disponíveis ao público para detecção automatizada de glaucoma tem sido um gargalo para a aplicação bem-sucedida de IA em direção a diagnósticos práticos auxiliados por computador (*Computer Aided Diagnosis* - CAD). Alguns pequenos conjuntos de dados disponíveis para a comunidade de pesquisa geralmente sofrem com condições impraticáveis de captura de imagens e critérios rigorosos de inclusão. Essas deficiências na escolha já limitada dos conjuntos de dados existentes tornam difícil amadurecer um sistema CAD para que ele possa funcionar no ambiente do mundo real (BAJWA, 2020).

Com a finalidade de desenvolver um conjunto de dados robusto, diversos conjuntos de dados públicos, enumerados na Tabela 3, foram consolidados em uma única base de dados para a condução dos experimentos nesta pesquisa.

Conjunto de dados	**G**	**NG**	**Total**	**Obs.**
Acrima	396	309	705	-
Beh	171	463	634	-
Drishti-gs	70	31	101	-
Hrf	15	15	45	15 imagens de retinopatia diabética
Keh	756	786	1542	-
Lag-part-1	1711	3143	4854	Dataset completo com 11760 imagens
Origa	168	482	650	-
Refuge	120	1080	1200	-
Rim-one dl	172	313	485	-

Tabela 3 – Contagem de imagens por conjunto de dados onde G e NG representam Glaucoma e Não-Glaucoma, respectivamente.

O conjunto de dados ACRIMA consiste em um total de 705 imagens de fundo de retina sendo 396 imagens com glaucoma e 309 imagens normais, tiradas com o disco óptico centralizado. Este conjunto de dados não fornece nenhuma anotação para segmentação do disco óptico e copa óptica (DIAZ-PINTO, 2019).

Com proporção relativamente equilibrada de imagens normais e glaucomatosas, o conjunto de dados ACRIMA se torna particularmente adequado para treinamento de classificadores baseados em DL (BAJWA, 2020).

O conjunto de dados BEH (*Bangladesh Eye Hospital*) contém um total de 634 imagens de fundo de retina sendo 171 imagens com glaucoma e 463 imagens normais. As imagens foram capturadas por uma câmera Topcon TRC-50DX, considerada o padrão-ouro para imagens da retina entre os anos de 2019 a 2020, de vários pacientes do Hospital Ocular de Bangladesh, em Dhaka, Bangladesh com idades entre 35 e 80 anos. Neste conjunto de dados, o diagnóstico de glaucoma na cabeça do nervo óptico (rótulo das imagens) foi realizado por dois profissionais da doença ocular, sendo um oftalmologista pediátrico e um cirurgião refrativo (ISLAM, 2022).

O conjunto de dados DRISHTI-GS é um conjunto de dados disponível publicamente para avaliação de glaucoma com segmentações de disco óptico e escavação. Este conjunto de dados consiste em 101 imagens de fundo de retina sendo 70 imagens de glaucoma e 31 imagens de olhos normais, divididas em diretórios de treinamento e teste, com quatro segmentações especializadas do disco e copo para o conjunto de treinamento (SIVASWAMY, 2014).

O conjunto de dados HRF foi estabelecido por um grupo de pesquisa colaborativo para apoiar estudos comparativos em algoritmos de segmentação automática em imagens

de fundo da retina. Este conjunto de dados, público, contém 15 imagens de pacientes saudáveis, 15 imagens de pacientes com retinopatia diabética e 15 imagens de pacientes glaucomatosos (BUDAI, 2013). Entretanto, para os propósitos deste estudo, optou-se por não incluir as 15 imagens referentes aos pacientes com retinopatia diabética.

O conjunto de dados Keh (*Kim's Eye Hospital*), publicado em 2018, contém um total de 1542 imagens de fundo de retina. Deste total, 786 são rotuladas como saudáveis (sem glaucoma), 289 são rotuladas com glaucoma inicial e 467 rotuladas com glaucoma avançado (KIM, 2018).

O conjunto de dados LAG, em sua versão completa, contém um total de 11760 imagens de fundo de retina. Deste total, 4878 são referentes a imagens com suspeita de glaucoma e 6882 imagens de olhos saudáveis. Todas as imagens são rotuladas com os resultados do diagnóstico (0 refere-se ao glaucoma negativo e 1 refere-se glaucoma suspeito). Neste conjunto de dados, 5824 imagens de fundo de retina são rotuladas com regiões de atenção baseadas em um método alternativo para rastreamento ocular, no qual 2392 são referentes a glaucoma positivo e o restante, 3432 imagens são referentes a glaucoma negativo (LI, 2019). Foi realizado contato com os autores, através de correio eletrônico, e parte do conjunto de dados LAG (Lag_part_1) foi fornecido para este trabalho. A parte fornecida contém um total de 4854 imagens de fundo de retina sendo 1711 imagens rotuladas como glaucoma e 3143 imagens rotuladas como olhos saudáveis (sem glaucoma).

O conjunto de dados ORIGA é um dos maiores e comumentemente um dos mais utilizados conjuntos de dados para detecção do glaucoma. Este conjunto de dados está público desde 2010 e consiste em 650 imagens de fundo de retina coletados pela Singapore Eye Research Institute entre os anos de 2004 e 2007. O conjunto de dados contém 168 imagens rotuladas como glaucomatosas e 482 imagens rotuladas como normais. Além dos rótulos para o glaucoma, o conjunto de dados fornece os valores de CDR para cada imagem (ZHANG, 2010).

O conjunto de dados REFUGE é um dos maiores conjuntos de dados disponível publicamente para detecção de glaucoma. Foi tornado público no ano de 2018 e consiste em 1200 imagens de fundo de retina contendo as marcações segmentadas do disco óptico e copa óptica, além dos rótulos clínicos de glaucoma. Apesar do grande tamanho deste conjunto de dados, seu conteúdo é altamente desequilibrado em relação à classe saudável, pois contém apenas 120 imagens de glaucoma (ORLANDO, 2019).

O conjunto de dados de imagens RIM-ONE DL consiste em 313 retinografias de indivíduos normais e 172 retinografias de pacientes com glaucoma. Estas imagens foram captadas em três hospitais espanhóis: Hospital Universitario de Canarias (HUC), em Tenerife, Hospital Universitario Miguel Servet (HUMS), em Saragoça, e Hospital Clínico Universitario San Carlos (HCSC), em Madrid (BATISTA, 2020).

De acordo com Batista, Diaz-Aleman, Sigut, Alayon, Arnay e Angel-Pereira (2020), este conjunto de dados, RIM-ONE DL, foi dividido em conjuntos de treinamento e teste, com duas variantes:

- Particionado aleatoriamente: os conjuntos de treinamento e teste são construídos aleatoriamente a partir de todas as imagens do conjunto de dados.
- Particionado por hospital: as imagens feitas no HUC são usadas para o conjunto de treinamento, enquanto as imagens feitas no HUMS e HCSC são usadas para teste.

Segundo Claro, Vogado, Santos e Veras (2020), a quantidade de amostras rotuladas disponíveis no conjunto de dados e os resultados gerados pelo algoritmo de ML estão totalmente relacionados, onde quanto maior a quantidade de amostras, maior a capacidade de predição dos classificadores. No entanto, em situações reais, possuir um conjunto de dados com uma grande quantidade de amostras rotuladas nem sempre é uma tarefa fácil, e muitas vezes, custosa. Desse modo, podemos utilizar abordagens de aumento de dados - *Data Augmentation* (DA), que é um conceito fundado por técnicas computacionais com o objetivo de aumentar a quantidade de amostras rotuladas em um conjunto de dados e assim, melhorar os resultados obtidos (TAYLOR; NITSCHKE, 2017).

Dados sintéticos, relacionados com as amostras originais do conjunto de dados, são gerados por meio de transformações. Essa prática envolve a aplicação de transformações como rotação, variação de cor, ajustes na luminosidade e/ou outras modificações que não sejam invasivas e degradem a natureza da imagem original. Essas perturbações acrescentam mais variabilidade à entrada, contribuindo assim para uma potencial redução na probabilidade de *overfitting* (PEREZ; WANG, 2017; CUBUK, 2018). Neste contexto, foram adicionados dados sintéticos ao conjunto de dados consolidado por meio da aplicação de transformações não intrusivas nas imagens originais. O procedimento de DA, empregado no conjunto de dados consolidado, será detalhado na próxima seção.

4.4 Pré-processamento das Imagens

Durante a fase de pré-processamento de imagens, foi realizada a consolidação das imagens provenientes dos diversos conjuntos de dados públicos mencionados anteriormente, na Tabela 3 (Seção 4.3 - Conjunto de Dados). Cada conjunto de dados contribuiu para a formação de uma base de dados unificada, englobando uma variabilidade representativa de casos clínicos.

O conjunto de dados consolidado apresenta uma heterogeneidade quanto à presença de imagens com o nervo óptico isolado. Enquanto alguns conjuntos de dados públicos disponibilizam imagens com o nervo óptico isolado, outros não o fazem, tornando

necessário, nesse último caso, realizar o isolamento. Para superar essa disparidade e assegurar consistência nos dados, foi utilizado o algoritmo proposto por Oliveira, Vieira, Filippo, Leles, Dias, Guimarães, Tuler e Rocha (2023) para isolar o nervo óptico em todas as imagens que não continham essa segmentação prévia. No método proposto por Oliveira, Vieira, Filippo, Leles, Dias, Guimarães, Tuler e Rocha (2023), o processo de isolamento do nervo óptico consiste em reduzir a imagem original, melhorar o contraste local e detalhes da imagem por meio de uma técnica específica (CLAHE) e isolar o nervo a partir do ponto mais brilhante em uma área delimitada. O algoritmo ajusta dinamicamente o tamanho da área de busca até encontrar o nervo óptico e realizar o seu recorte/isolamento.

Realizar o isolamento do disco óptico tem uma razão clínica, já que o glaucoma afeta principalmente o disco óptico e seus arredores (DIAZ-PINTO, 2019). Conforme apontado por Orlando, Prokofyeva, Fresno e Blaschko (2017), a estratégia de recorte das imagens ao redor do disco óptico mostrou-se mais eficiente do que a utilização das imagens completas ao empregar a CNN na avaliação do glaucoma.

Na Figura 14, estão ilustrados alguns exemplos de imagens com o nervo óptico isolado pertencentes ao conjunto de dados consolidado. Essas imagens estão rotuladas como casos de glaucoma, sendo caracterizadas pela presença de uma escavação óptica mais proeminente na região central do copo óptico, uma marcante característica que se destaca nas imagens positivas para glaucoma.

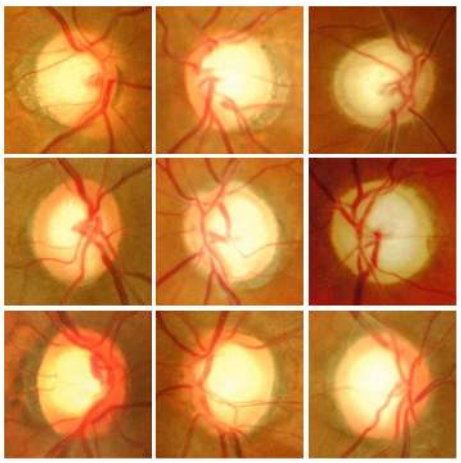


Figura 14 – Exemplo de imagens do conjunto de dados consolidado rotuladas com glaucoma. Fonte: Elaborado pelo autor.

Na Figura 15, estão ilustrados alguns exemplos de imagens com o nervo óptico isolado pertencentes ao conjunto de dados consolidado. Essas imagens estão rotuladas como casos de olhos saudáveis, ou seja, sem glaucoma, evidenciando a ausência da característica de escavação óptica acentuada na região central do copo óptico, que é distintiva nas imagens positivas para glaucoma.

Com o objetivo de aumentar a quantidade de amostras rotuladas no conjunto de

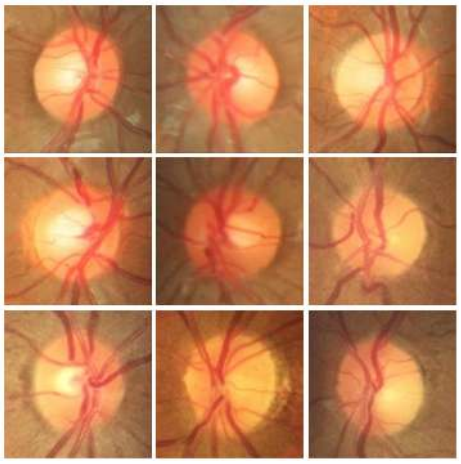


Figura 15 – Exemplo de imagens do conjunto de dados consolidado rotuladas como saudáveis (sem glaucoma). Fonte: Elaborado pelo autor.

dados, dar mais variedade e melhorar os resultados obtidos reduzindo a probabilidade de *overfitting* (TAYLOR; NITSCHKE, 2017), o processo de DA foi realizado através da aplicação de operações de transformações não invasivas no conjunto de dados consolidado. Essas transformações envolvem operações de rotação, zoom e espelhamento horizontal e vertical nas imagens do conjunto de dados consolidado (imagens com o nervo óptico isolado). Para cada imagem presente no conjunto de dados consolidado, foram geradas mais duas imagens sintéticas mediante a aplicação, de forma aleatória, das transformações especificadas na Tabela 4, cujos valores, foram escolhidos empiricamente, visando otimizar a diversidade e preservar as características fundamentais das imagens originais.

Transformação	**Valor**
Rotação	20 graus
Zoom	20%
Espelhamento Horizontal	True
Espelhamento Vertical	True

Tabela 4 – Transformações aplicadas no conjunto de dados consolidado para o processo de DA

Ao término da fase de pré-processamento, o conjunto de dados consolidado passou a contar com um total de 28.539 imagens, das quais 18.363 são rotuladas como saudáveis (sem glaucoma) e 10.176 são rotuladas com glaucoma. Nenhum outro metadado adicional, como CDR ou diâmetro do disco, por exemplo, foi utilizado. O conjunto de dados consolidado está disponível no seguinte endereço: https://shre.ink/unified-dataset-glaucoma.

4.5 Treinamento e Validação das Arquiteturas de CNNs

Durante a etapa de treinamento e validação das arquiteturas de CNNs, foram treinados seis modelos de diferentes arquiteturas de CNNs: VGG16, VGG19, InceptionV3,

DenseNet, Xception e ResNet50.

Todos os modelos foram treinados utilizando a aprendizagem por transferência - *Transfer Learning* (TL) - com os pesos pré-treinados do ImageNet (DENG, 2009). Os parâmetros utilizados para a compilação de todos os modelos foram: *batch_size*: 8, *learning_rate*: 0,0001 e *epochs*: 10. A escolha desses parâmetros buscou um equilíbrio entre eficiência computacional (devido a limitação de hardware) e a prevenção do sobreajuste. Optou-se por um menor número de épocas a fim de prevenir treinamento excessivo, considerando a potencialidade de que um número maior de épocas poderia resultar em sobreajuste aos dados de treinamento, prejudicando a capacidade de generalização dos modelos.

As imagens foram redimensionadas para o tamanho de entrada padrão de cada arquitetura de CNN. Para as arquiteturas VGG16, VGG19, DenseNet e ResNet50 as imagens foram redimensionadas para 224x224 e para as arquiteturas InceptionV3 e Xception as imagens foram redimensionadas para 229x229.

Para a avaliação de desempenho das CNNs (VGG16, VGG19, InceptionV3, DenseNet, Xception e ResNet50), foi utilizada a técnica de validação cruzada estratificada (HASTIE, 2009). A validação cruzada estratificada é uma técnica de avaliação de modelos na qual os dados são divididos em k partes, garantindo que cada subconjunto preserve a mesma proporção de classes que o conjunto de dados original. Essa abordagem assegura uma avaliação mais robusta e representativa do desempenho do modelo em conjuntos de dados desbalanceados. Para lidar com limitações de hardware e mitigar possíveis restrições no número de épocas, permitindo uma avaliação mais robusta do desempenho dos modelos em diferentes subconjuntos de dados, optou-se pela aplicação da validação cruzada estratificada com k igual a cinco ($k = 5$). Neste contexto, foram obtidos cinco valores correspondentes às métricas de acurácia, precisão, sensibilidade, *f1-score* e AUC.

Essas métricas fazem uso de uma matriz de confusão, que é uma matriz quadrada e visa comparar os valores previstos pelo modelo com os valores esperados (valores corretos). A diagonal principal contém os acertos do modelo e as posições restantes denotam erros. Há quatro tipos de rótulos em uma matriz de confusão. Verdadeiro negativo (*True Negative* - TN) representa o número de amostras rotuladas corretamente para a classe negativa. Verdadeiro positivo (*True Positive* - TP) indica o número de amostras rotuladas corretamente para a classe positiva. Falso negativo (*False Negative* - FN) é o número de amostras rotuladas incorretamente para a classe positiva. E falso positivo (*False Positive* - FP) é o número de amostras rotuladas incorretamente para a classe negativa (MILANI, 2023).

A acurácia é uma métrica fundamental que expressa a proporção de predições corretas em relação ao total de previsões realizadas pelo modelo. O cálculo da acurácia é feito dividindo o número de predições corretas pelo total de predições. A fórmula é:

$$\text{Acurácia} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4.1}$$

A precisão mede a proporção de instâncias positivas corretamente classificadas em relação ao total de instâncias classificadas como positivas pelo modelo. O cálculo da precisão é dado pela fórmula:

$$\text{Precisão} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4.2}$$

A sensibilidade avalia a capacidade do modelo em identificar todas as instâncias positivas. O cálculo é feito dividindo o número de verdadeiros positivos pelo total de positivos reais. A fórmula é:

$$\text{Sensibilidade} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.3}$$

O *f1-score* é uma métrica que combina precisão e sensibilidade em um único valor, sendo útil quando ambas as métricas são importantes. É calculado pela média harmônica entre precisão e sensibilidade, usando a fórmula:

$$\textit{F1-Score} = 2 \times \frac{\text{Precisão} \times \text{Sensibilidade}}{\text{Precisão} + \text{Sensibilidade}} \tag{4.4}$$

A AUC é utilizada para avaliar a capacidade discriminativa de um modelo em relação às classes positivas e negativas. Representa a área sob a curva ROC (*Receiver Operating Characteristic*) e varia de 0 a 1, sendo 1 indicativo de um modelo perfeito. A curva ROC compara a taxa de verdadeiros positivos com a taxa de falsos positivos em diferentes pontos de corte de probabilidade (LING, 2003). Quanto maior a AUC, melhor o desempenho do modelo.

O desempenho final de cada modelo/arquitetura é representado pela média desses valores (acurácia, precisão, sensibilidade, *f1-score* e AUC). Os resultados obtidos durante a fase de treinamento e validação estão registrados na Tabela 5. Os códigos-fonte utilizados para o desenvolvimento dos modelos estão disponíveis no repositório do GitHub, podendo ser acessados através do seguinte endereço: `https://shre.ink/cnns-glaucoma`.

4.6 Aplicação das Técnicas de XAI

Durante a fase de implementação das técnicas de XAI, foram empregadas seis técnicas distintas para interpretação visual em cada modelo de CNN desenvolvido. As técnicas de XAI adotadas incluíram LIME, SHAP, CAM, Grad-CAM, Vanilla Gradients e SmoothGrad. Adicionalmente, uma estratégia inovadora, denominada SCIM, foi desenvolvida, fundamentando-se na confluência das regiões importantes, destacadas pelos

CNN	Acurácia	Precisão	Sensibilidade	*F1-Score*	AUC
VGG16	0.97864	0.97168	0.96858	0.97008	0.9764
VGG19	0.9744	0.9774	0.9493	0.9626	0.9688
InceptionV3	0.79512	0.76414	0.6959	0.69624	0.77304
DenseNet	0.8622	0.79426	0.84394	0.81372	0.85812
Xception	0.80818	0.75038	0.75836	0.72348	0.79708
ResNet50	0.9615	0.97424	0.91634	0.94208	0.95146

Tabela 5 – Desempenho de cada arquitetura de CNN no conjunto de dados consolidado utilizando a validação cruzada estratificada com cinco dobras

métodos SHAP e CAM, e posteriormente aplicada nos modelos treinados. Os códigos-fonte utilizados para o desenvolvimento das técnicas de XAI estão disponíveis no repositório do GitHub, podendo ser acessados através do seguinte endereço: https://shre.ink/xai-glaucoma.

A aplicação das técnicas de XAI restringiu-se a uma amostra representativa do conjunto de dados, com foco exclusivo nos casos classificados pelas CNNs como glaucoma. A seleção dessas amostras, totalizando vinte e quatro, seguiu dois critérios: primeiro, as seis CNNs deveriam concordar na classificação das imagens como casos positivos de glaucoma; segundo, as imagens foram escolhidas de modo a abranger as diversas variações de tonalidade presentes no conjunto de dados, incluindo tonalidades avermelhadas, amareladas, mais escuras e mais claras.

Nas próximas subseções, serão delineados os principais aspectos relacionados à implementação de cada técnica de XAI adotada neste estudo.

4.6.1 LIME (*Local Interpretable Model-Agnostic Explanations*)

Para a implementação do LIME foi utilizado o pacote LIME na versão 0.2.0.1[1].

A implementação realizada carrega o modelo de rede neural convolucional treinado para a classificação de imagens, carrega uma imagem de exemplo, segmenta a imagem usando o algoritmo SLIC (*Simple Linear Iterative Clustering*), e em seguida executa uma explicação local da classificação da imagem.

O SLIC é um algoritmo de segmentação de imagens que agrupa pixels com características similares em uma mesma região (segmento). É um método rápido e eficiente, que utiliza uma abordagem de clusterização de *K-means* em um espaço de cores em que cada dimensão corresponde a uma informação da imagem, como cor, intensidade, textura, etc. Os parâmetros utilizados na função *slic* definem como a segmentação será realizada. O parâmetro *n_segments* indica o número de segmentos desejados. O parâmetro *compactness* é um parâmetro que controla a suavização dos limites entre segmentos. Quanto maior o

[1] Disponível em: https://pypi.org/project/lime/. Acesso em 04/11/2023.

valor de *compactness*, mais suaves são os limites. O parâmetro *sigma* controla a variação espacial. Um valor baixo de *sigma* indica que pixels próximos têm maior peso no cálculo das diferenças entre os clusters. Já o parâmetro *start_label* define a etiqueta do primeiro segmento.

Os parâmetros utilizados na implementação foram *n_segments=100*, *compactness=10*, *sigma=1* e *startlabel=1*. A Figura 16 representa a segmentação realizada pelo algoritmo SLIC com os parâmetros utilizados na implementação.

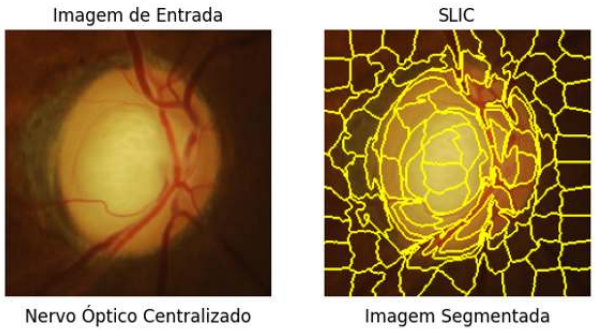


Figura 16 – Representação da imagem de entrada e da imagem segmentada pelo algoritmo SLIC. Fonte: Elaborado pelo autor.

A explicação local, utilizando o LIME, consiste em gerar uma máscara que destaca as regiões importantes da imagem, de acordo com a segmentação SLIC, que mais influenciaram na classificação da CNN.

A classe *LimeImageExplainer*, do pacote LIME, é utilizada para gerar uma explicação local da classificação da imagem de entrada. A explicação local tenta explicar a previsão de classe da rede neural, mostrando quais partes da imagem (segmentos) são (ou foram) mais importantes para a classificação.

O método *explain_instance* da classe *LimeImageExplainer* é chamado com os seguintes parâmetros:

- *img.astype('double')*: é a imagem de entrada que será explicada, convertida para o tipo de dados double;
- *model.predict*: é a função de previsão da rede neural que será explicada;
- *top_labels = 2*: são as duas classes previstas pelo modelo para a imagem de entrada;
- *hide_color = 0*: é a cor de fundo para as partes da imagem que não são importantes para a classificação;
- *num_samples = 1000*: é o número de amostras geradas para aproximar a distribuição de pesos da classe prevista pelo modelo;

- *segmentation_fn = slic*: é a função de segmentação (algoritmo) usada para segmentar a imagem de entrada em regiões.

Como saída da explicação, apresenta-se uma figura contendo cinco imagens: a imagem de entrada, a imagem referente a segmentação SLIC, a máscara das regiões importantes (explicação LIME), a imagem de entrada com a sobreposição das regiões importantes geradas pela explicação LIME e um mapa de calor que representa a importância de cada segmento SLIC na classificação realizada pelo modelo.

As Figuras 17, 18, 19, 20, 21 e 22 ilustram um exemplo de glaucoma para a saída do LIME nas arquiteturas VGG16, VGG19, InceptionV3, Xception, DenseNet e ResNet50, respectivamente.

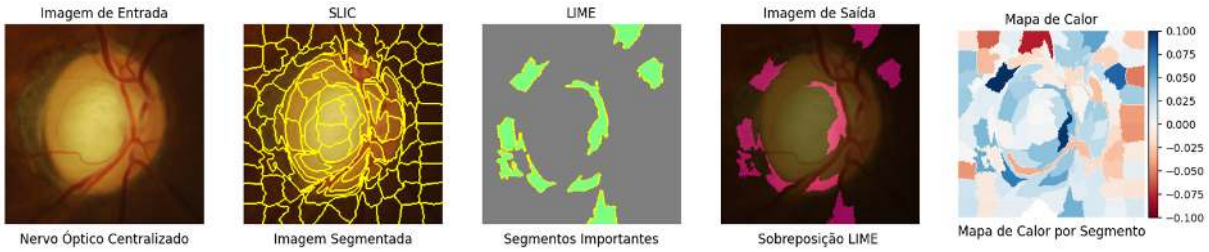


Figura 17 – Aplicação da técnica LIME na arquitetura VGG16. Fonte: Elaborado pelo autor.

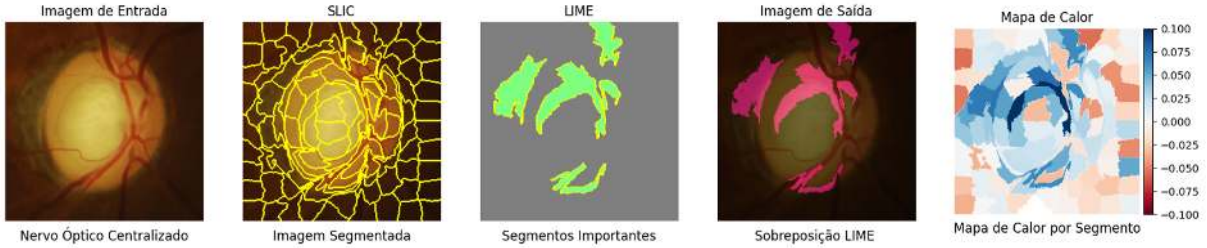


Figura 18 – Aplicação da técnica LIME na arquitetura VGG19. Fonte: Elaborado pelo autor.

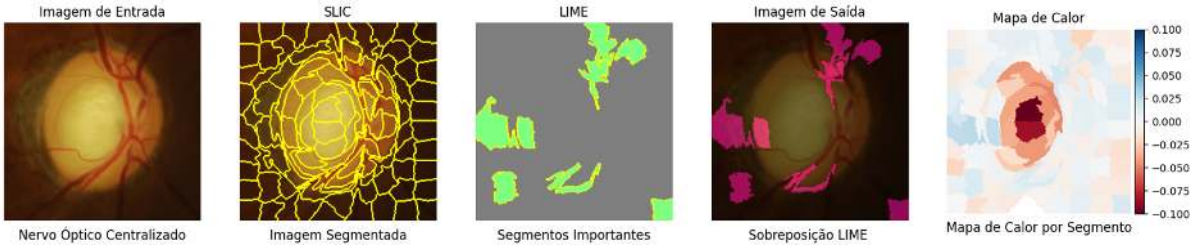


Figura 19 – Aplicação da técnica LIME na arquitetura InceptionV3. Fonte: Elaborado pelo autor.

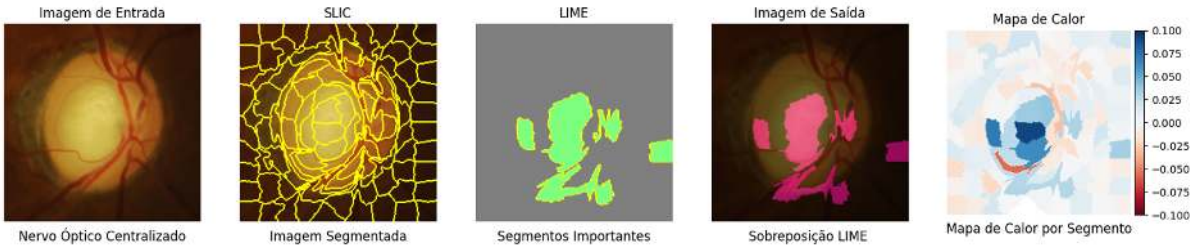


Figura 20 – Aplicação da técnica LIME na arquitetura Xception. Fonte: Elaborado pelo autor.

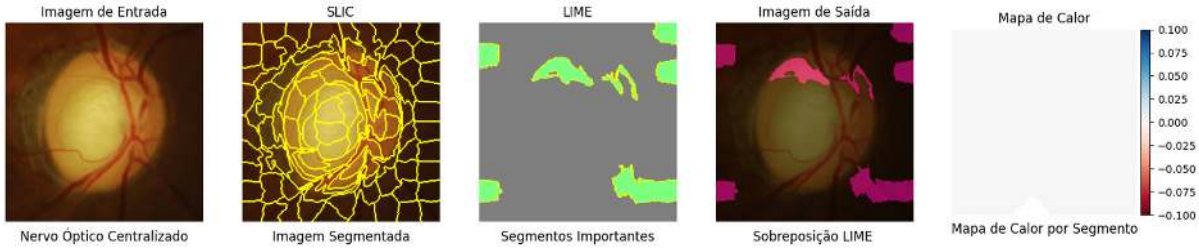


Figura 21 – Aplicação da técnica LIME na arquitetura DenseNet. Fonte: Elaborado pelo autor.

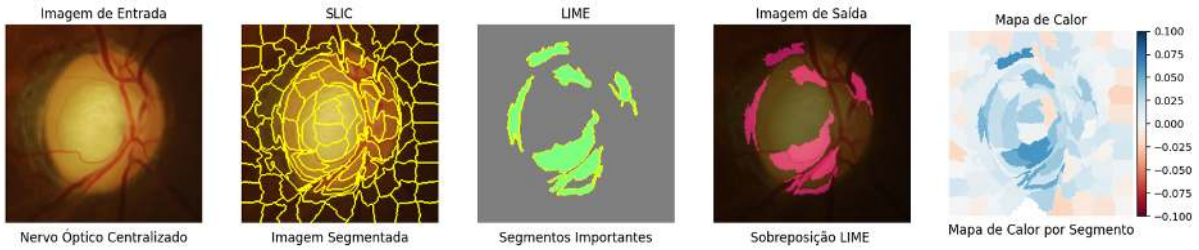


Figura 22 – Aplicação da técnica LIME na arquitetura ResNet50. Fonte: Elaborado pelo autor.

4.6.2 SHAP (*SHapley Additive exPlanations*)

Para a implementação do SHAP foi realizada a instalação do pacote python SHAP na versão 0.41.0[2].

O modelo de CNN treinado é carregado utilizando o método *load_model* do pacote Keras, a imagem de entrada é carregada e o valor de *Shapley* é calculado sobre os dados do conjunto de treinamento. O objeto de explicação (*explainer*) do tipo *shap.GradientExplainer* é inicializado com dois parâmetros: o modelo treinado e o conjunto de dados de treinamento. Esse objeto *explainer* é responsável por calcular os valores de SHAP para a imagem de entrada. Ele utiliza o modelo fornecido para calcular os gradientes em relação às características da imagem. Esses gradientes medem o quanto cada pixel contribui para a previsão final do modelo.

Em seguida, a função *shap_values* do objeto *explainer* é chamada, passando a imagem de entrada como argumento. Essa função retorna os valores de SHAP (*shap_values*) calculados para cada pixel da imagem de entrada representando a contribuição de cada pixel para a diferença entre a previsão do modelo e um valor de referência (a média das previsões calculadas com os dados de treinamento).

Os *shap_values* são interpretados visualmente para entender quais regiões da imagem são mais importantes para a previsão feita pelo modelo. Valores positivos (pixels na tonalidade rosa) indicam que a presença da característica aumenta a probabilidade da classe prevista, enquanto valores negativos (pixels na tonalidade azul) indicam que a presença da característica diminui a probabilidade da classe prevista. A magnitude dos valores de SHAP reflete a importância relativa das características.

Como saída da explicação, apresenta-se uma figura contendo duas imagens: a imagem de entrada, com o nervo óptico centralizado e a máscara representando a importância dos pixels para a previsão na saída do modelo. Uma barra indicando a magnitude dos valores SHAP (valores máximos e mínimos do *shap_values*) também é exibida.

As Figuras 23, 24, 25, 26, 27 e 28 ilustram um exemplo de glaucoma para a saída do SHAP nas arquiteturas VGG16, VGG19, InceptionV3, Xception, DenseNet e ResNet50, respectivamente.

[2] Disponível em: https://pypi.org/project/shap/0.41.0/. Acesso em 04/03/2023.

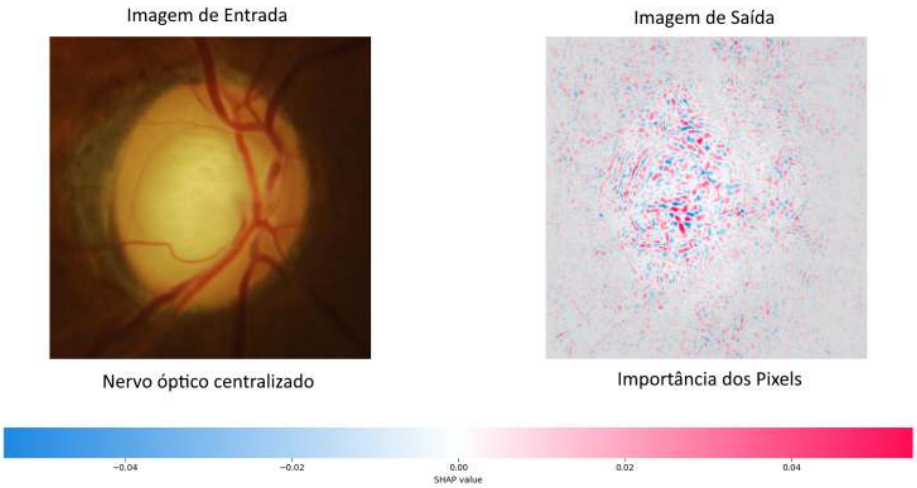


Figura 23 – Aplicação da técnica SHAP na arquitetura VGG16. Fonte: Elaborado pelo autor.

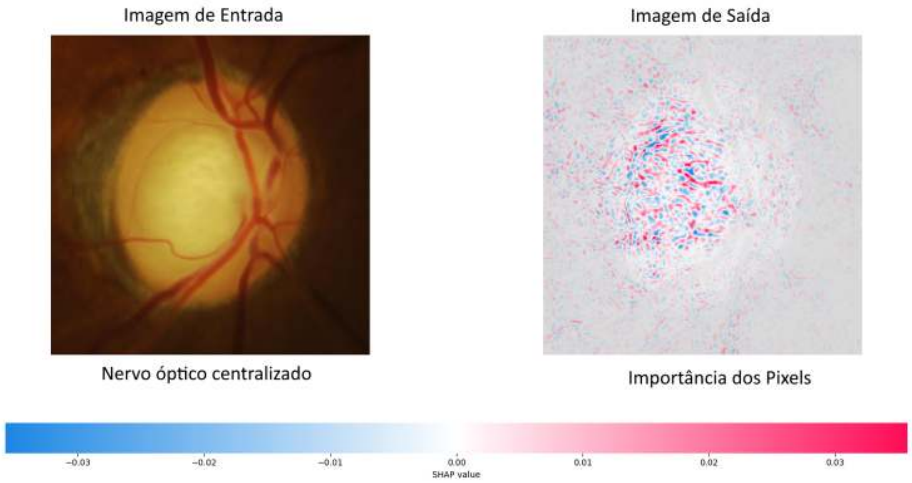


Figura 24 – Aplicação da técnica SHAP na arquitetura VGG19. Fonte: Elaborado pelo autor.

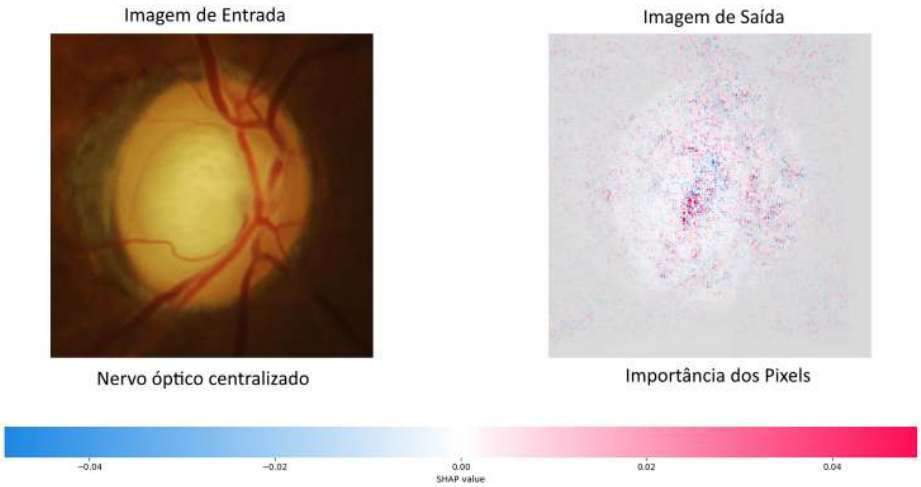


Figura 25 – Aplicação da técnica SHAP na arquitetura InceptionV3. Fonte: Elaborado pelo autor.

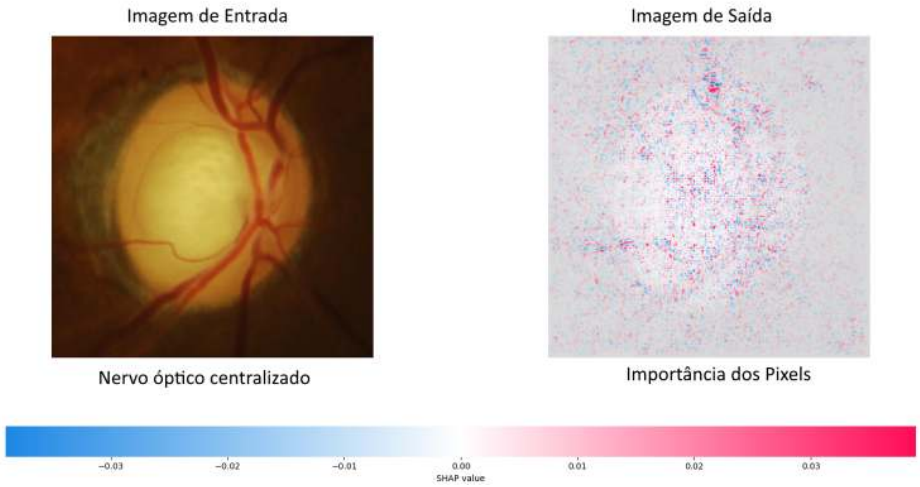


Figura 26 – Aplicação da técnica SHAP na arquitetura Xception. Fonte: Elaborado pelo autor.

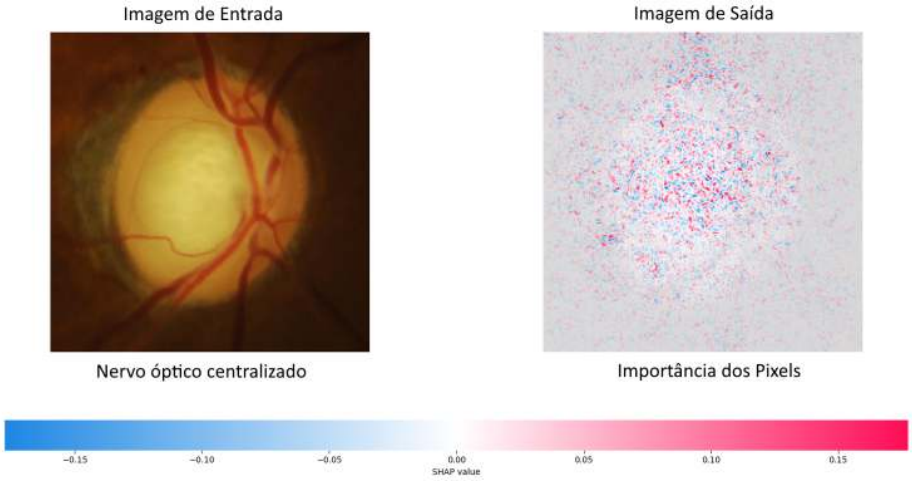


Figura 27 – Aplicação da técnica SHAP na arquitetura DenseNet. Fonte: Elaborado pelo autor.

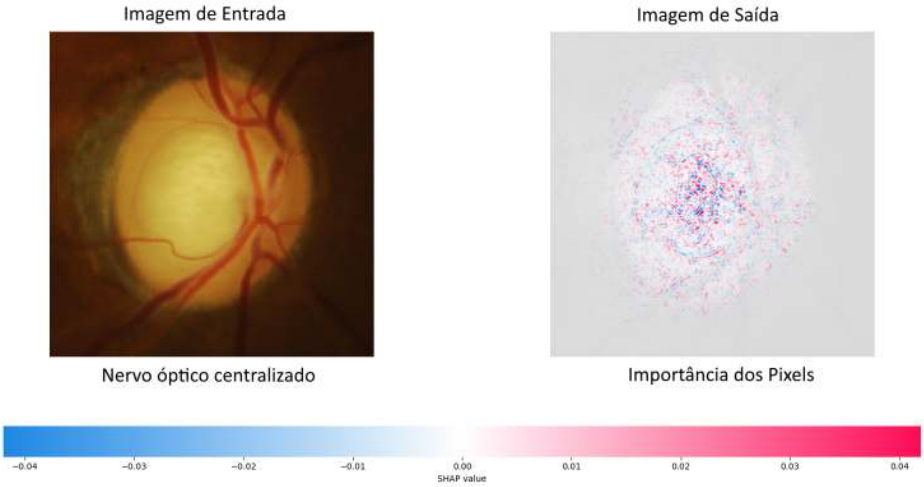


Figura 28 – Aplicação da técnica SHAP na arquitetura ResNet50. Fonte: Elaborado pelo autor.

4.6.3 CAM (*Class Activation Mapping*)

A técnica CAM foi implementada utilizando as bibliotecas TensorFlow (versão 2.10.1)[3], NumPy (versão 1.26.2)[4], Keras (versão 2.10.0)[5], Matplotlib (versão 3.8.2)[6] e OpenCV (versão 4.8.1.78)[7].

O modelo de CNN treinado é carregado utilizando o método *load_model* do pacote Keras e a imagem de entrada é carregada utilizando o pacote OpenCV. A imagem de entrada é convertida para o espaço de cores RGB e redimensionada para o padrão esperado pela arquitetura de CNN (224x224 ou 229x229). Em seguida a imagem de entrada é transformada em um vetor utilizando o método *array* do pacote NumPy e uma dimensão extra é adicionada utilizando o método *expand_dims*.

Os gradientes são calculados utilizando o método *GradientTape* do pacote TensorFlow. A média dos gradientes ao longo dos eixos (0, 1, 2) é calculada utilizando o método *reduce_mean* também do pacote Tensorflow. Esse cálculo é armazenado em uma variável denominanda *pooled_grads*. O mapa de ativação da classe (*heatmap*) é calculado multiplicando os gradientes pela saída da camada convolucional final e acrescentando uma nova dimensão.

Por fim, para um aprimoramento na precisão do resultado, o processo de normalização do mapa de ativação é realizado. Primeiro, é realizado um ajuste de valores negativos para zero através do método *maximum(heatmap, 0)*, onde o *heatmap* representa o mapa de calor da ativação de classe gerado anteriormente. Em seguida, o mapa de ativação é dividido pelo valor máximo presente no mapa através do método *heatmap / math.reduce_max(heatmap)* para garantir que os valores estejam no intervalo entre 0 e 1, proporcionando uma representação mais precisa e escalonada do impacto das características ativadas.

Como saída da explicação, apresenta-se uma figura contendo três imagens: a imagem de entrada, com o nervo óptico centralizado, o mapa de calor com as ativações dos gradientes indicando as regiões mais fortes e importantes e a imagem de saída, com o mapa de calor sobreposto na imagem de entrada.

As Figuras 29, 30, 31, 32, 33 e 34 ilustram um exemplo de glaucoma para a saída do CAM nas arquiteturas VGG16, VGG19, InceptionV3, Xception, DenseNet e ResNet50, respectivamente. Para a arquitetura ResNet50 (Figura 34), a técnica CAM não produziu resultados (*heatmap* e, consequentemente, a imagem de saída com sobreposição do mapa de calor na imagem de entrada).

[3] Disponível em: https://pypi.org/project/tensorflow/2.10.1/. Acesso em 17/11/2023.
[4] Disponível em: https://pypi.org/project/numpy/1.26.2/. Acesso em 17/11/2023.
[5] Disponível em: https://pypi.org/project/keras/2.10.0/. Acesso em 17/11/2023.
[6] Disponível em: https://pypi.org/project/matplotlib/3.8.2/. Acesso em 17/11/2023.
[7] Disponível em: https://pypi.org/project/opencv-python/4.8.1.78/. Acesso em 17/11/2023.

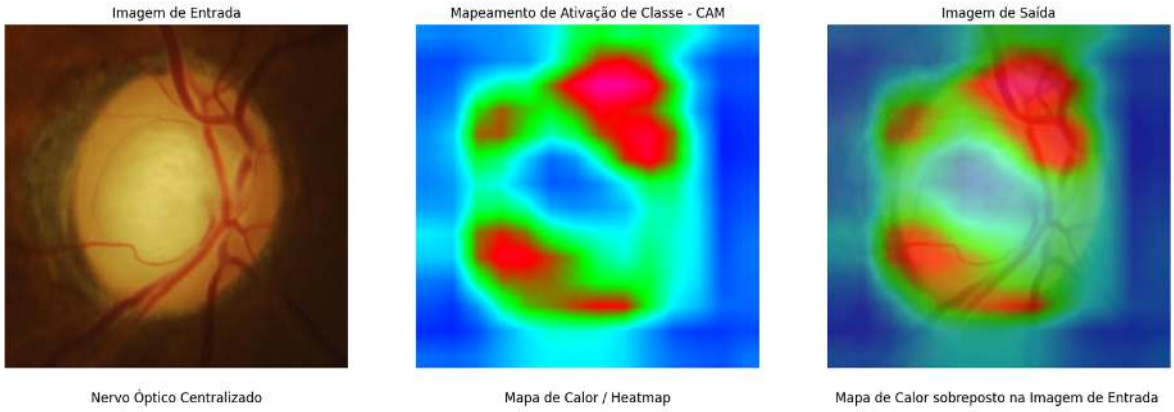


Figura 29 – Aplicação da técnica CAM na arquitetura VGG16. Fonte: Elaborado pelo autor.

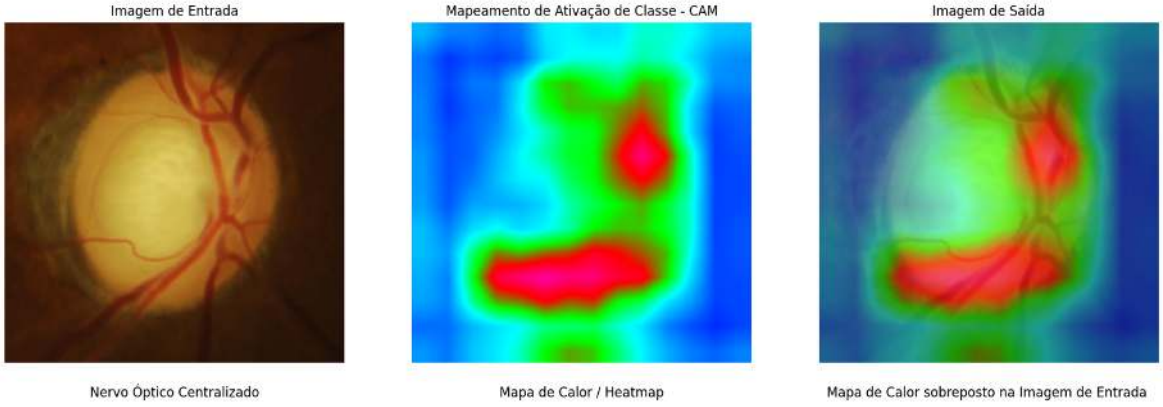


Figura 30 – Aplicação da técnica CAM na arquitetura VGG19. Fonte: Elaborado pelo autor.

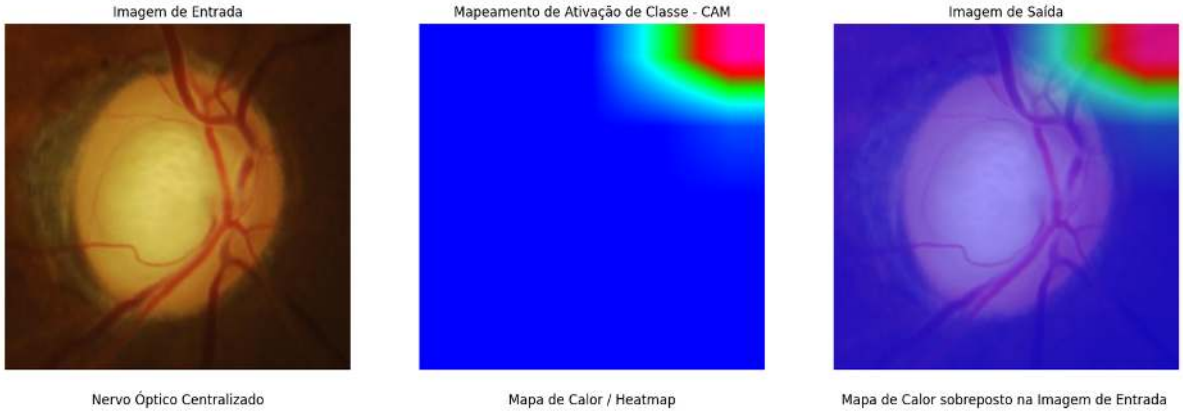


Figura 31 – Aplicação da técnica CAM na arquitetura InceptionV3. Fonte: Elaborado pelo autor.

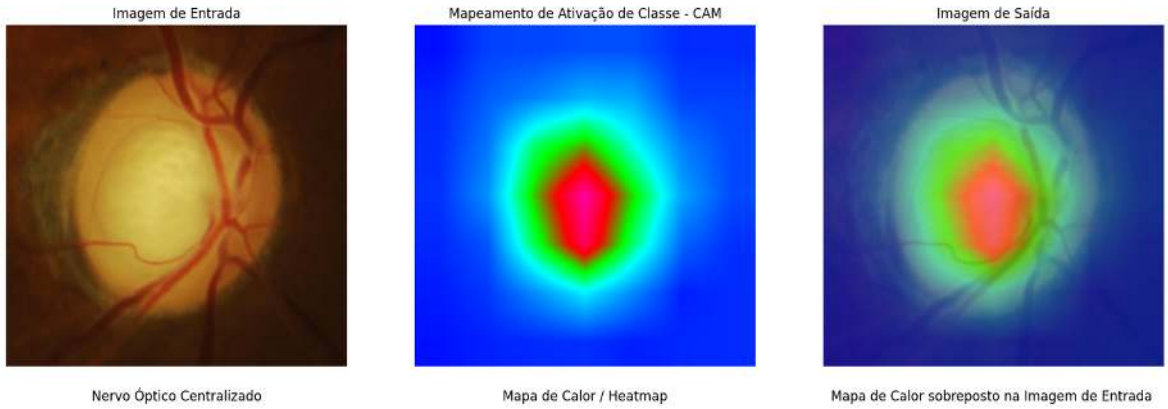


Figura 32 – Aplicação da técnica CAM na arquitetura Xception. Fonte: Elaborado pelo autor.

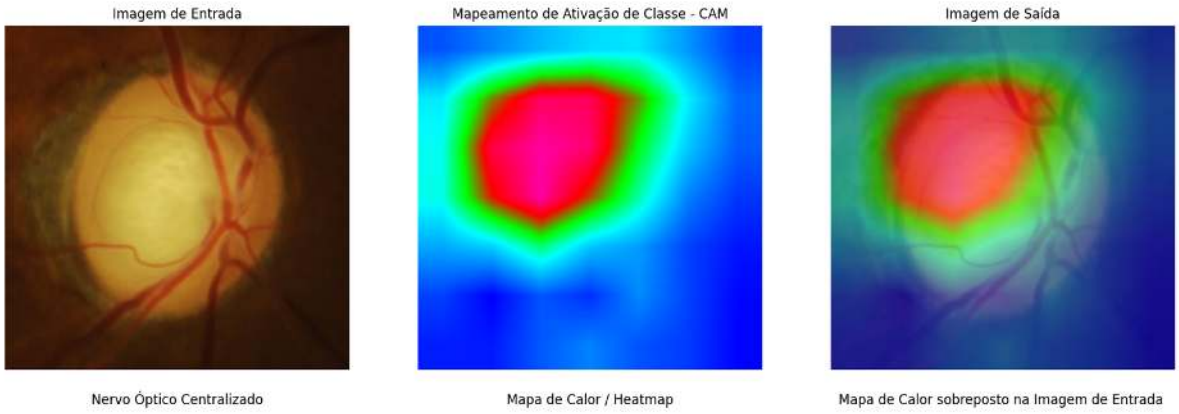


Figura 33 – Aplicação da técnica CAM na arquitetura DenseNet. Fonte: Elaborado pelo autor.

Figura 34 – Aplicação da técnica CAM na arquitetura ResNet50. Fonte: Elaborado pelo autor.

4.6.4 Grad-CAM (*Gradient-weighted Class Activation Mapping*)

Para implementação da técnica Grad-CAM foram utilizadas as bibliotecas TensorFlow (versão 2.10.1), NumPy (versão 1.26.2), Keras (versão 2.10.0), Matplotlib (versão 3.8.2) e OpenCV (versão 4.8.1.78).

O modelo de CNN treinado é carregado utilizando o método *load_model* do pacote Keras e a imagem de entrada é carregada utilizando o pacote OpenCV. A imagem de entrada é convertida para o espaço de cores RGB e redimensionada para o padrão esperado pela arquitetura de CNN (224x224 ou 229x229). Em seguida a imagem de entrada é transformada em um vetor utilizando o método *array* do pacote NumPy e uma dimensão extra é adicionada utilizando o método *expand_dims*.

Da mesma forma que o método CAM, os gradientes são calculados utilizando o método *GradientTape* do pacote TensorFlow. A média dos gradientes ao longo dos eixos (0, 1, 2) é calculada utilizando o método *reduce_mean* também do pacote Tensorflow. Esse cálculo é armazenado em uma variável denominanda *pooled_grads*. O mapa de ativação da classe (*heatmap*) é calculado multiplicando as ativações da última camada convolucional pelos gradientes médios ponderados e somando em todos os canais.

Por fim, para um aprimoramento na precisão do resultado, o processo de normalização do mapa de ativação é realizado garantindo que os valores estejam no intervalo entre 0 e 1, proporcionando uma representação mais precisa e escalonada do impacto das características ativadas.

Como saída da explicação, apresenta-se uma figura contendo três imagens: a imagem de entrada, com o nervo óptico centralizado, o mapa de calor com as ativações do gradientes ponderados indicando as regiões mais fortes e importantes (região amarelada) e a imagem de saída, com o mapa de calor sobreposto na imagem de entrada.

As Figuras 35, 36, 37, 38, 39 e 40 ilustram um exemplo de glaucoma para a saída do Grad-CAM nas arquiteturas VGG16, VGG19, InceptionV3, Xception, DenseNet e ResNet50, respectivamente. Na arquitetura Xception (Figura 38), a técnica Grad-CAM não produziu resultados (*heatmap* e, consequentemente, a imagem de saída com sobreposição do mapa de calor na imagem de entrada).

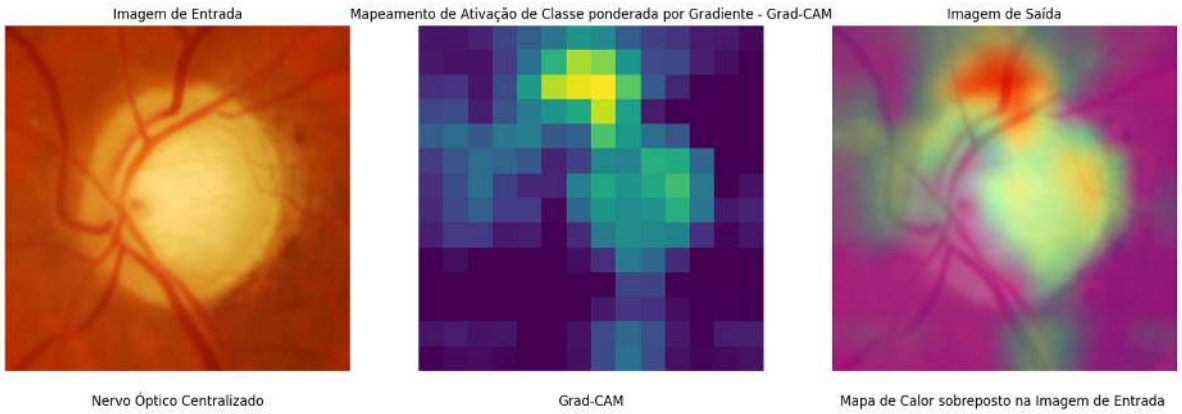


Figura 35 – Aplicação da técnica Grad-CAM na arquitetura VGG16. Fonte: Elaborado pelo autor.

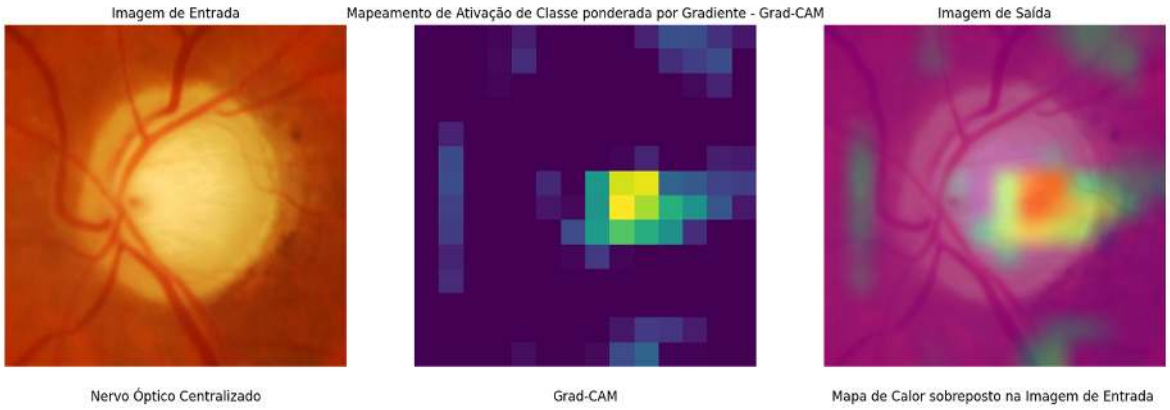


Figura 36 – Aplicação da técnica Grad-CAM na arquitetura VGG19. Fonte: Elaborado pelo autor.

Figura 37 – Aplicação da técnica Grad-CAM na arquitetura InceptionV3. Fonte: Elaborado pelo autor.

Figura 38 – Aplicação da técnica Grad-CAM na arquitetura Xception. Fonte: Elaborado pelo autor.

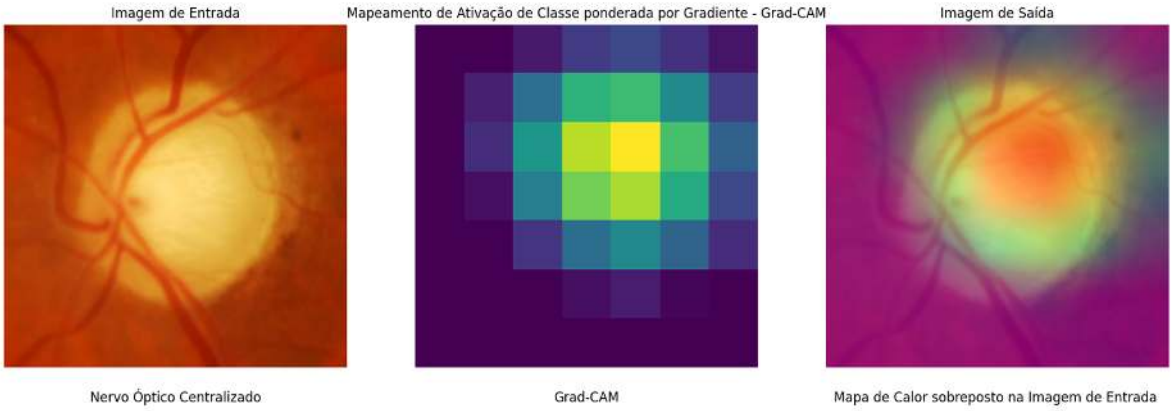


Figura 39 – Aplicação da técnica Grad-CAM na arquitetura DenseNet. Fonte: Elaborado pelo autor.

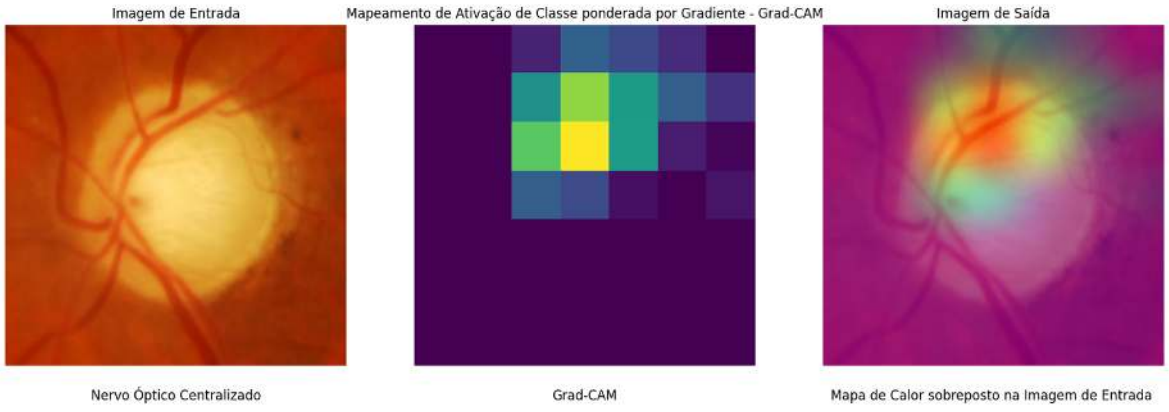


Figura 40 – Aplicação da técnica Grad-CAM na arquitetura ResNet50. Fonte: Elaborado pelo autor.

4.6.5 Vanilla Gradients

Para implementação da técnica Vanilla Gradients foram utilizadas as bibliotecas TensorFlow (versão 2.10.1), NumPy (versão 1.26.2), Keras (versão 2.10.0), OpenCV (versão 4.8.1.78), Pillow (versão 10.1.0)[8], Matplotlib (versão 3.8.2) e o pacote SALIENCY (versão 0.2.0)[9].

O modelo de CNN treinado é carregado utilizando o método *load model* do pacote Keras e a imagem de entrada é carregada utilizando o pacote OpenCV. A imagem de entrada é convertida para o espaço de cores RGB e redimensionada para o padrão esperado pela arquitetura de CNN (224x224 ou 229x229). Em seguida a imagem de entrada é transformada em um vetor utilizando o método *array* do pacote NumPy e uma dimensão extra é adicionada utilizando o método expand dims.

Inicialmente, uma instância do objeto *GradientSaliency* da biblioteca *saliency* é criada. Em seguida, esse objeto é utilizado para calcular a máscara de interpretabilidade chamada de *vanilla_mask_3d* através do método *GetMask*. Esse método recebe como entrada a imagem e duas funções que representam a função do modelo (*call_model_function*) a ser interpretado e seus argumentos (argumentos adicionais).

Uma máscara tridimensional obtida é convertida para uma representação mais compreensível visualmente (imagem 2D em tons de cinza) através da função *VisualizeImageGrayscale* também do pacote *saliency*. O resultado final é armazenado em uma variável chamada *vanilla_mask_grayscale*. Essa etapa de visualização é essencial para interpretar e entender quais partes da imagem são mais influentes na decisão do modelo.

Como saída da explicação, apresenta-se uma figura contendo duas imagens: a imagem de entrada, com o nervo óptico centralizado e a imagem de saída representando o mapa de gradientes (ou mapa de saliência). Quanto maior a intensidade da cor branca no mapa de gradientes, maior a importância dessa região da imagem para a predição do modelo.

As Figuras 41, 42, 43, 44, 45 e 46 ilustram um exemplo de glaucoma para a saída do Vanilla Gradients nas arquiteturas VGG16, VGG19, InceptionV3, Xception, DenseNet e ResNet50, respectivamente. Na arquitetura VGG19 (Figura 42), a técnica Vanilla Gradients não produziu resultado (mapa de gradientes).

[8] Disponível em: https://pypi.org/project/pillow/10.1.0/. Acesso em 17/11/2023.
[9] Disponível em: https://pypi.org/project/saliency/0.2.0/. Acesso em 17/11/2023.

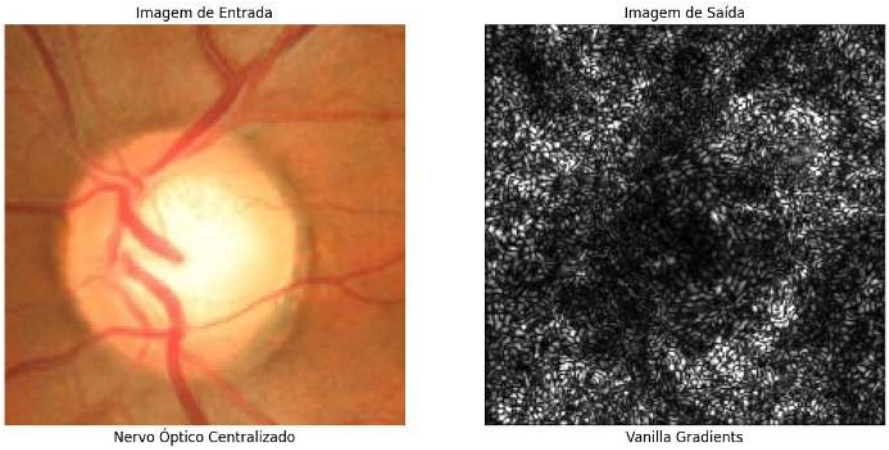


Figura 41 – Aplicação da técnica Vanilla Gradients na arquitetura VGG16. Fonte: Elaborado pelo autor.

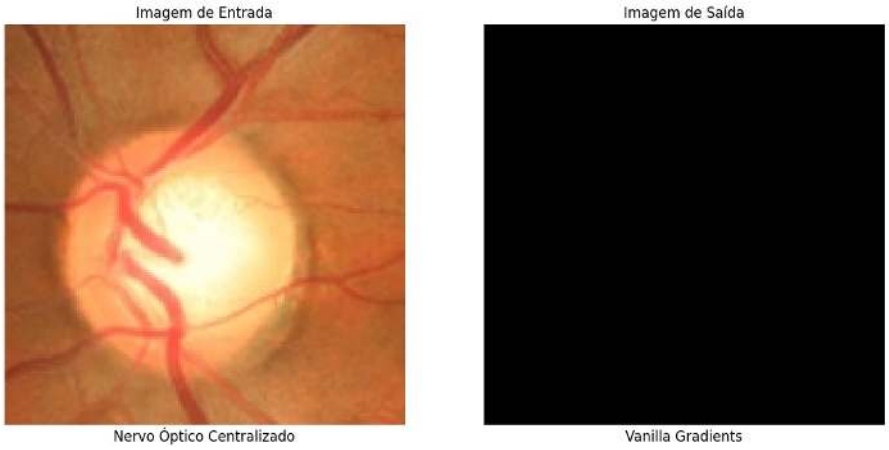


Figura 42 – Aplicação da técnica Vanilla Gradients na arquitetura VGG19. Fonte: Elaborado pelo autor.

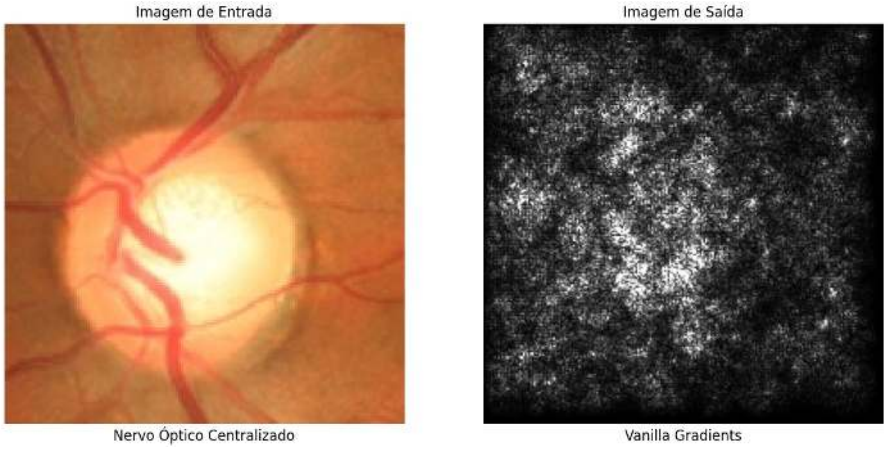


Figura 43 – Aplicação da técnica Vanilla Gradients na arquitetura InceptionV3. Fonte: Elaborado pelo autor.

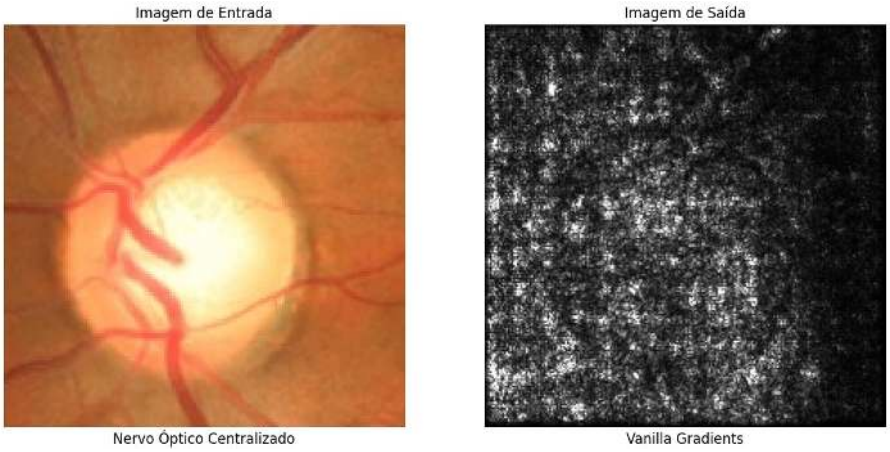


Figura 44 – Aplicação da técnica Vanilla Gradients na arquitetura Xception. Fonte: Elaborado pelo autor.

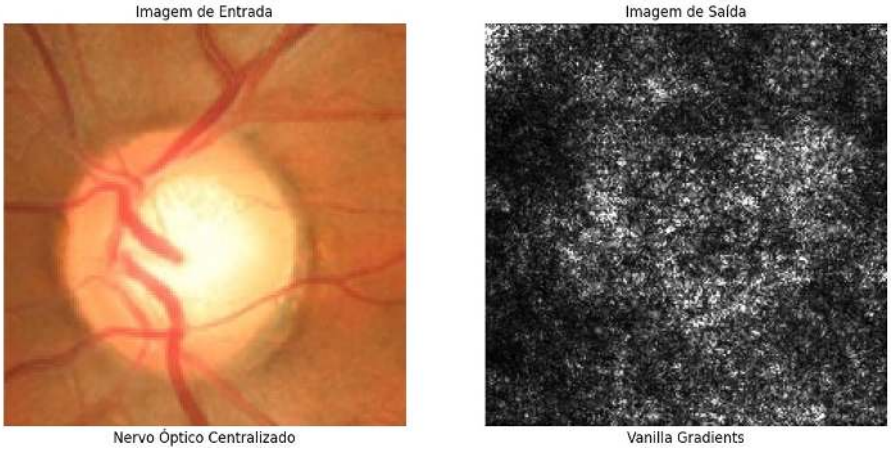


Figura 45 – Aplicação da técnica Vanilla Gradients na arquitetura DenseNet. Fonte: Elaborado pelo autor.

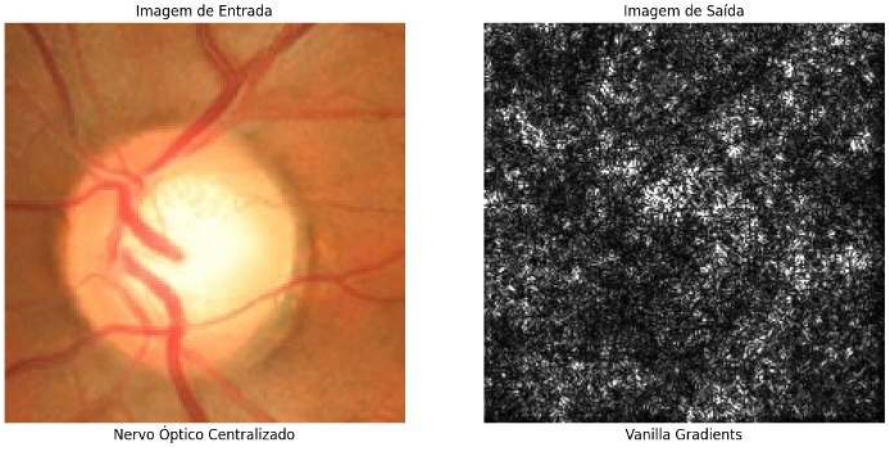


Figura 46 – Aplicação da técnica Vanilla Gradients na arquitetura ResNet50. Fonte: Elaborado pelo autor.

4.6.6 SmoothGrad

Para implementação da técnica SmoothGrad foram utilizadas as bibliotecas TensorFlow (versão 2.10.1), NumPy (versão 1.26.2), Keras (versão 2.10.0), OpenCV (versão 4.8.1.78), Pillow (versão 10.1.0), Matplotlib (versão 3.8.2) e o pacote SALIENCY (versão 0.2.0).

O modelo de CNN treinado é carregado utilizando o método *load model* do pacote Keras e a imagem de entrada é carregada utilizando o pacote OpenCV. A imagem de entrada é convertida para o espaço de cores RGB e redimensionada para o padrão esperado pela arquitetura de CNN (224x224 ou 229x229). Em seguida a imagem de entrada é transformada em um vetor utilizando o método *array* do pacote NumPy e uma dimensão extra é adicionada utilizando o método expand dims.

Inicialmente, uma instância do objeto *GradientSaliency* da biblioteca *saliency* é criada. Em seguida, esse objeto é utilizado para calcular a máscara de interpretabilidade chamada de *smoothgrad_mask_3d* através do método *GetSmoothedMask*. Esse método recebe como entrada a imagem e duas funções que representam a função do modelo (*call_model_function*) a ser interpretado e seus argumentos (argumentos adicionais).

Uma máscara tridimensional obtida é convertida para uma representação mais compreensível visualmente (imagem 2D em tons de cinza) através da função *VisualizeImageGrayscale* também do pacote *saliency*. O resultado final é armazenado em uma variável chamada *smoothgrad_mask_grayscale*. Essa etapa de visualização é essencial para interpretar e entender quais partes da imagem são mais influentes na decisão do modelo.

Como saída da explicação, apresenta-se uma figura contendo duas imagens: a imagem de entrada, com o nervo óptico centralizado e a imagem de saída representando o mapa de gradientes (mapa de saliência) suavizado. Quanto maior a intensidade da cor branca no mapa de gradientes, maior a importância dessa região da imagem para a predição do modelo.

As Figuras 47, 48, 49, 50, 51 e 52 ilustram um exemplo de glaucoma para a saída do SmoothGrad nas arquiteturas VGG16, VGG19, InceptionV3, Xception, DenseNet e ResNet50, respectivamente.

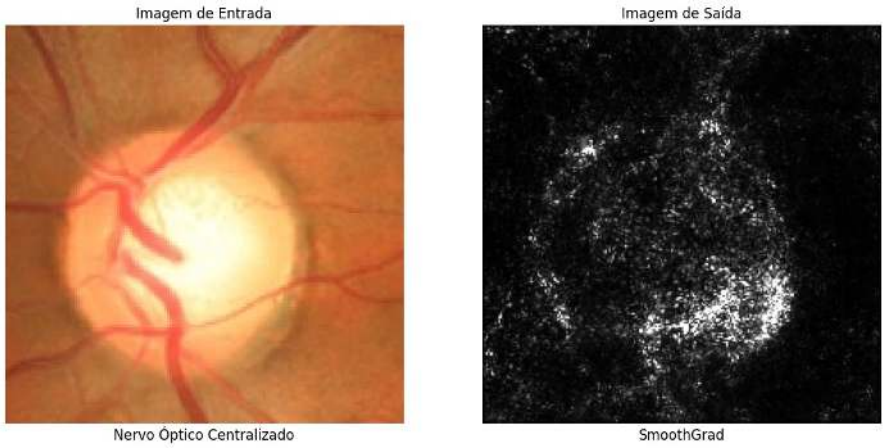


Figura 47 – Aplicação da técnica SmoothGrad na arquitetura VGG16. Fonte: Elaborado pelo autor.

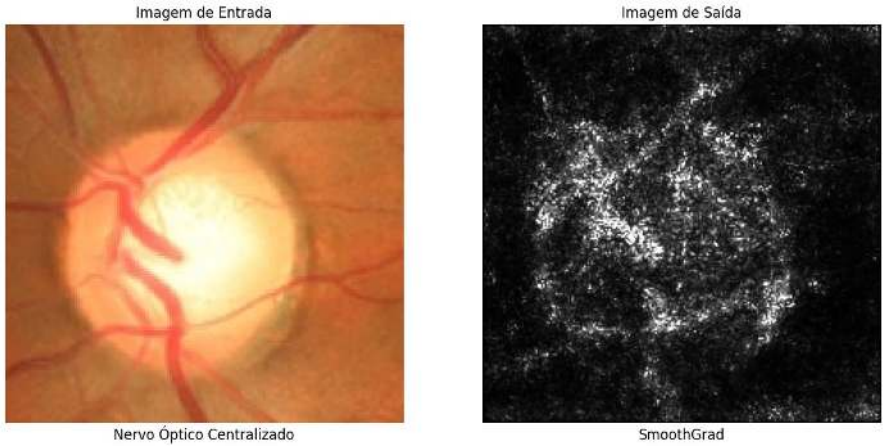


Figura 48 – Aplicação da técnica SmoothGrad na arquitetura VGG19. Fonte: Elaborado pelo autor.

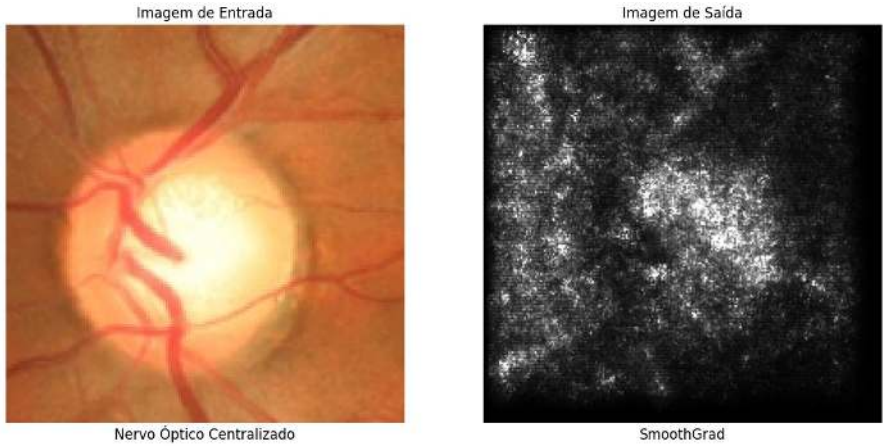


Figura 49 – Aplicação da técnica SmoothGrad na arquitetura InceptionV3. Fonte: Elaborado pelo autor.

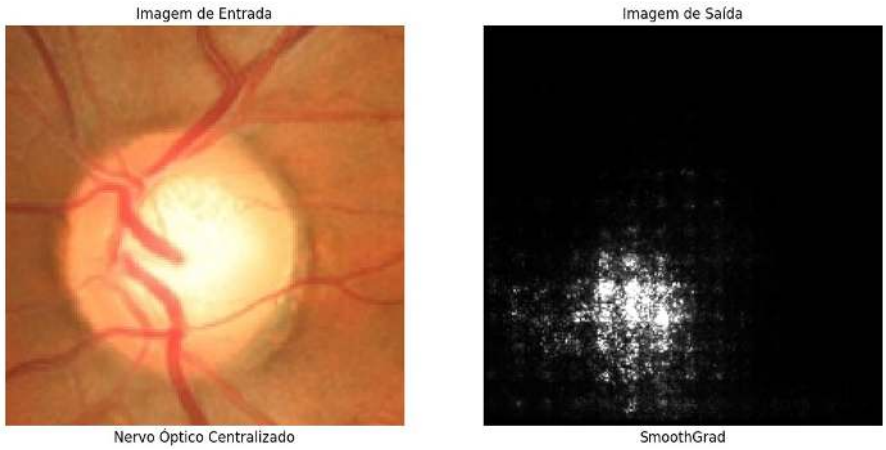


Figura 50 – Aplicação da técnica SmoothGrad na arquitetura Xception. Fonte: Elaborado pelo autor.

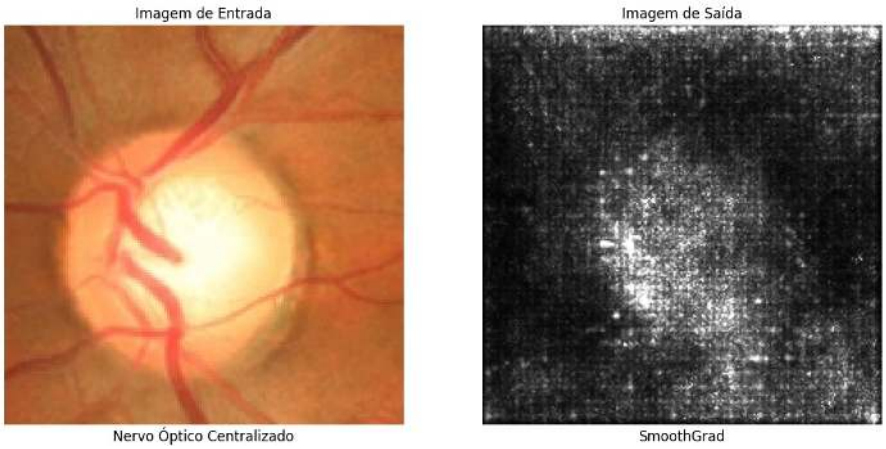


Figura 51 – Aplicação da técnica SmoothGrad na arquitetura DenseNet. Fonte: Elaborado pelo autor.

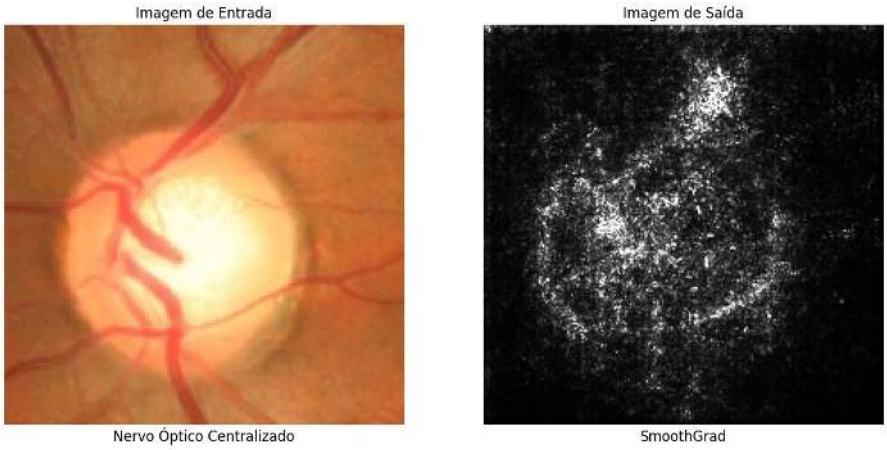


Figura 52 – Aplicação da técnica SmoothGrad na arquitetura ResNet50. Fonte: Elaborado pelo autor.

4.6.7 SCIM (*SHAP-CAM Interpretable Mapping*)

A abordagem SCIM propõe uma estratégia inovadora para interpretar as CNNs, utilizando a confluência entre duas técnicas reconhecidas: o SHAP e o CAM. O SHAP destaca-se por atribuir importância a cada pixel da imagem de entrada, proporcionando uma análise minuciosa da contribuição de cada pixel na classificação final da CNN. De forma complementar, o CAM gera mapas de ativação de classe, identificando regiões importantes na imagem que influenciam a predição para uma classe específica. Essas técnicas são combinadas na SCIM, visando criar um mapeamento interpretável das CNNs.

No contexto da SCIM, o SHAP é aplicado às camadas intermediárias da CNN, atribuindo importâncias locais a cada pixel. Estas importâncias (*shap_values* positivos) são então confrontadas com as ativações do CAM, revelando a região de confluência entre esses dois métodos. Esse processo resulta na geração de um mapa interpretável, denominado máscara de confluência, que destaca a área da imagem considerada importante para a predição do modelo, representando uma fusão significativa e coesa das informações do SHAP e do CAM.

A implementação prática da abordagem SCIM segue uma série de passos bem definidos, conforme ilustrado na Figura 53.

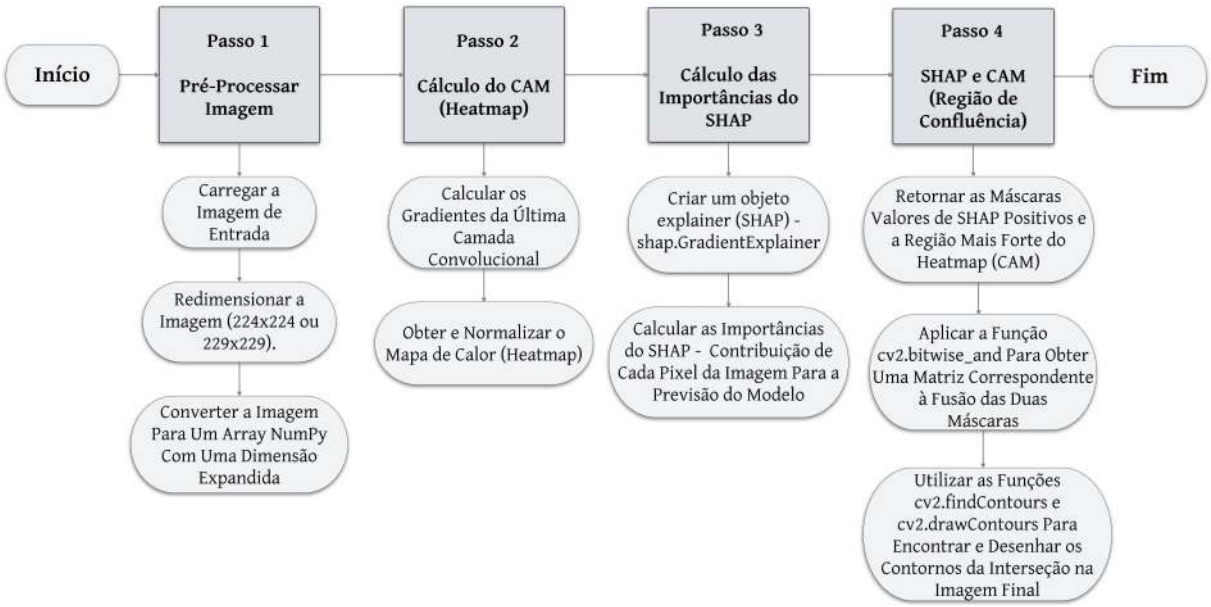


Figura 53 – Fluxograma da abordagem SCIM. Fonte: Elaborado pelo autor.

Inicialmente, no Passo 1, uma imagem de entrada é carregada, convertida para o espaço de cores RGB, redimensionanda para o tamanho esperado pelo modelo (224x224 ou 229x229) e convertida em um *array* NumPy com uma dimensão expandida representando o lote de entrada. Em seguida, no Passo 2, O CAM é calculado através de um modelo intermediário mapeando a imagem de entrada para as ativações da última ca-

mada convolucional e as previsões de saída. Os gradientes são calculados com relação às ativações da última camada convolucional e o mapa de calor (*heatmap*) é obtido e normalizado. Na próxima fase, Passo 3, o objeto *explainer*, do SHAP, é criado usando o método *shap.GradientExplainer*. As importâncias do SHAP, fornecendo uma medida de contribuição de cada pixel da imagem para a previsão do modelo, são calculadas com o *explainer.shap_values* passando a imagem de entrada como parâmetro. Com o *shap_values* (SHAP) e o *heatmap* (CAM) calculados, no Passo 4, um filtro é realizado para retornar os *shap_values* positivos, ou seja, maiores que zero e a região mais forte do *heatmap*. A função *cv2.inRange* é usada para criar uma máscara e a função *cv2.bitwise_and* para retornar uma matriz que corresponde à imagem resultante da fusão das duas máscaras, ou seja, os pontos de confluência. A função *cv2.findContours* e *cv2.drawContours* são utilizadas para encontrar os contornos da interseção e desenhá-los na imagem final.

A saída da explicação consiste em uma figura contendo cinco imagens: a imagem de entrada, a máscara com os pixels positivos do SHAP, a máscara com a região ativada pelo CAM, a máscara da região de confluência entre SHAP e CAM, e a imagem de saída com os contornos destacados representando a área mais relevante para a predição do modelo.

As Figuras 54, 55, 56, 57, 58 e 59 exemplificam a aplicação da abordagem SCIM em diferentes arquiteturas de CNNs, ilustrando o resultado obtido para um exemplo de glaucoma. Nota-se que em algumas arquiteturas, como InceptionV3 (Figura 56) e ResNet50 (Figura 59), a ausência de regiões de interseção entre SHAP e CAM compromete o resultado na imagem de saída.

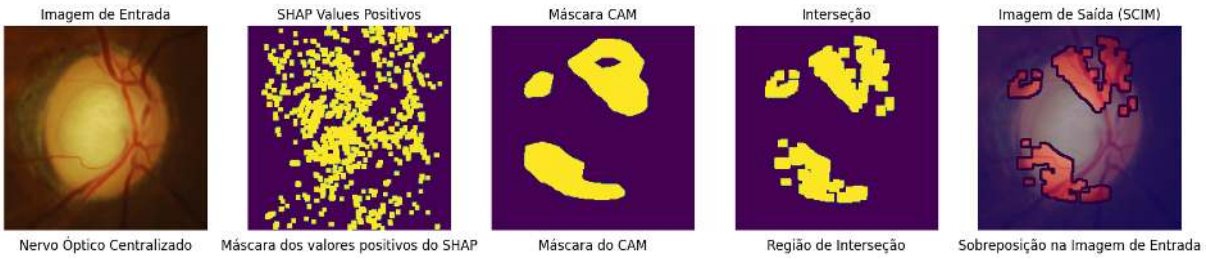


Figura 54 – Aplicação da abordagem SCIM na arquitetura VGG16. Fonte: Elaborado pelo autor.

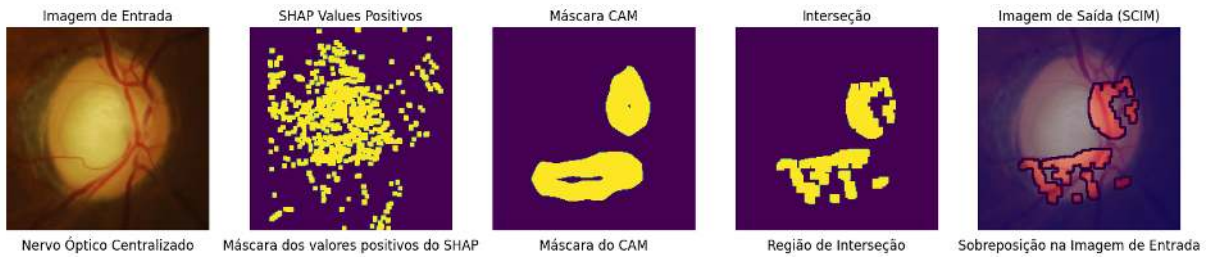


Figura 55 – Aplicação da abordagem SCIM na arquitetura VGG19. Fonte: Elaborado pelo autor.

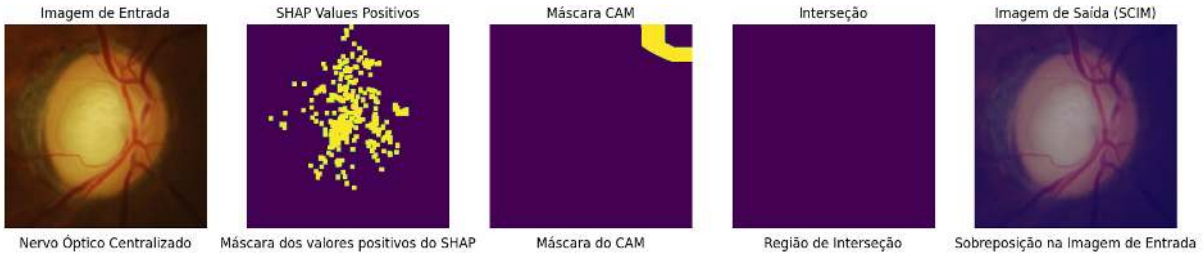


Figura 56 – Aplicação da abordagem SCIM na arquitetura InceptionV3. Fonte: Elaborado pelo autor.

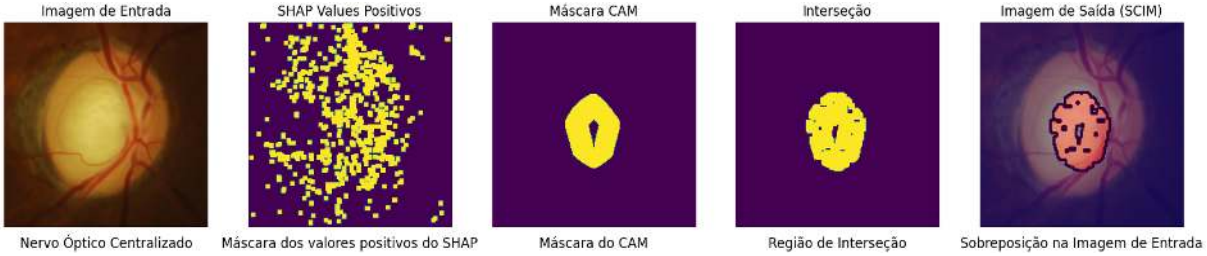


Figura 57 – Aplicação da abordagem SCIM na arquitetura Xception. Fonte: Elaborado pelo autor.

Figura 58 – Aplicação da abordagem SCIM na arquitetura DenseNet. Fonte: Elaborado pelo autor.

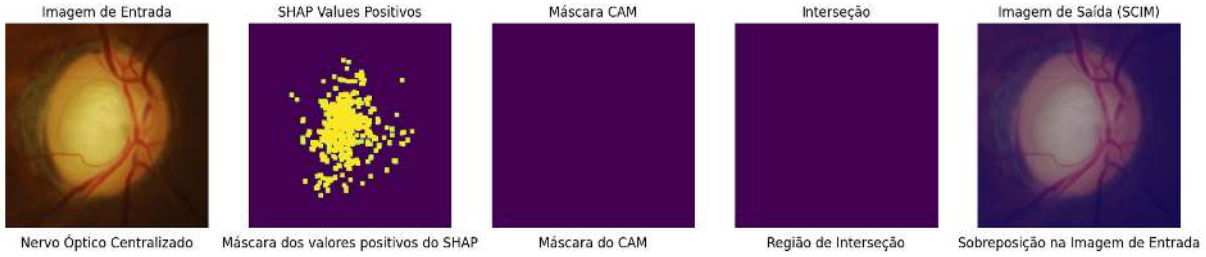


Figura 59 – Aplicação da abordagem SCIM na arquitetura ResNet50. Fonte: Elaborado pelo autor.

5 Análise dos Resultados e Discussões

Neste capítulo, são apresentados os resultados derivados da exploração e aplicação das técnicas de XAI nas diferentes arquiteturas de CNNs, proporcionando *insights* significativos em relação aos objetivos estabelecidos neste trabalho de pesquisa. Nas próximas seções, são delineadas a análise comparativa, discussão e uma avaliação crítica dos resultados obtidos.

5.1 Análise Comparativa e Discussão

Para efeito de avaliação e comparação entre as técnicas XAI foram definidos alguns indicadores conforme descritos em Molnar (2022):

- Interpretabilidade: Capacidade de entender e interpretar as decisões do modelo classificando sua qualidade e utilidade;
- Conformidade com o Domínio: Adequação das explicações geradas às expectativas e regras estabelecidas no domínio específico;
- Estabilidade: Consistência das explicações geradas pelas técnicas de XAI, considerando estável ao produzir explicações consistentes para as mesmas entradas repetidamente; e
- Eficiência Computacional: Refere-se ao tempo e aos recursos necessários para gerar as explicações. Basicamente, esse indicador se refere ao tempo de processamento para execução da técnica (algoritmo).

Para a avaliação dos indicadores de Interpretabilidade e Conformidade com o Domínio, contamos com a expertise de um profissional da área de oftalmologia, com mais de 7 anos de experiência no diagnóstico e tratamento do glaucoma. Este especialista, cujo grau de familiaridade com sistemas computacionais de apoio à decisão é indeterminado, foi selecionado para uma avaliação de nível humano (tarefa simples), conforme descrito por Doshi-Velez e Kim (2017) e teve a oportunidade de comparar cada técnica de XAI, aplicada nas diferentes arquiteturas de CNNs. A análise foi conduzida sob a perspectiva médica, permitindo uma avaliação da interpretabilidade visual gerada por cada método e/ou abordagem.

O instrumento de coleta de dados utilizado nesta avaliação com o especialista compreende um questionário composto por 24 perguntas, o qual está disponível no Apêndice A desta dissertação. A última indagação, identificada como Pergunta 24, é de natureza

opcional e apresenta um formato de resposta aberta, permitindo que o avaliador forneça comentários e/ou sugestões específicas em relação às observações sobre cada técnica de XAI. Para as 23 perguntas anteriores, é solicitado ao avaliador a atribuição das respostas em uma escala de 1 a 5, onde (1) indica discordância total, (2) discordância, (3) neutralidade, (4) concordância e (5) concordância total. Cada uma dessas perguntas é acompanhada de uma representação visual, na forma de figura, que ilustra a aplicação de cada técnica de XAI em suas respectivas arquiteturas de CNNs, exceto para os casos em que as técnicas de XAI não apresentaram resultados em arquiteturas específicas, a exemplo do CAM aplicado na arquitetura ResNet50 e do Grad-CAM na arquitetura Xception.

Para a análise dos indicadores de Estabilidade e Eficiência Computacional, foram conduzidos testes empíricos que incluíram a mensuração do tempo de processamento utilizando o módulo *time* do Python e a avaliação da consistência nos resultados obtidos por meio de múltiplas execuções. A fim de avaliar o indicador de Estabilidade, foi analisado o resultado de cada técnica de XAI, aplicada em cada arquitetura de CNN, por meio de três execuções consecutivas. Para a análise da Eficiência Computacional, o módulo *time* proporcionou uma abordagem que permitiu a medição do tempo médio de processamento dedicado à execução de cada técnica de XAI em cada arquitetura de CNN, proporcionando uma visão clara sobre o desempenho.

A Tabela 6 apresenta um resumo referente à análise comparativa realizada entre as técnicas XAI de acordo com os indicadores-chave previamente estabelecidos.

Técnica de XAI	**I**	**E**	**CD**	**EC**
LIME	Boa	Baixa	Baixa	Média
SHAP	Boa	Boa	Boa	Média
CAM	Alta	Boa	Alta	Alta
Grad-CAM	Baixa	Boa	Baixa	Alta
Vanilla Gradients	Baixa	Boa	Baixa	Alta
SmoothGrad	Baixa	Boa	Baixa	Alta
SCIM	Boa	Boa	Boa	Média

Tabela 6 – Análise comparativa entre as técnicas XAI em relação aos indicadores-chave, onde I, E, CD e EC representam Interpretabilidade, Estabilidade, Conformidade com o Domínio e Eficiência Computacional, respectivamente.

Os rótulos "Baixa", "Média", "Boa"e "Alta" foram utilizados para permitir uma avaliação clara e relativa das técnicas em relação aos quatro indicadores. Esses rótulos proporcionam uma escala gradativa que permite diferenciar entre as técnicas com base nos níveis de desempenho em cada indicador, oferecendo uma maneira intuitiva de comparar e comunicar as características das técnicas, permitindo o entendimento de como cada uma se posiciona em relação aos indicadores-chave.

A Tabela 7 sintetiza os tempos médios de processamento das técnicas de XAI em

todas as imagens do conjunto amostral. Os dados, ordenados em forma crescente, refletem as médias dos tempos de execução das técnicas em cada uma das seis arquiteturas de CNNs investigadas neste estudo, proporcionando uma análise abrangente do desempenho temporal relativo.

Técnica de XAI	**Tempo Médio (em segundos)**
CAM	14.347031
Vanilla Gradients	15.168250
Grad-CAM	16.833995
Smoothgrad	62.318061
LIME	517.019131
SHAP	956.836887
SCIM	971.183918

Tabela 7 – Tempos médios de processamento das técnicas de XAI em todas as imagens do conjunto amostral, considerando as seis arquiteturas de CNNs exploradas neste estudo.

Ao aplicar a técnica LIME, observou-se, em todas as arquiteturas de CNNs, superpixels fora da área de interesse, ou seja, distante dos arredores do disco óptico. Este é um aspecto prejudicial no quesito Conformidade com o Domínio. As segmentações que foram determinantes para o diagnóstico do glaucoma se espalharam por regiões fora do disco óptico apresentando uma métrica de baixa fidelidade.

Outro aspecto importante observado nos experimentos é a instabilidade das explicações geradas pelo LIME. Ao executar o LIME algumas vezes, com os mesmos parâmetros e a mesma imagem de entrada, são geradas saídas diferentes, o que compromete a consistência das explicações. As Figuras 60, 61 e 62 ilustram a execução da técnica LIME para a mesma imagem de entrada, na mesma arquitetura de CNN (VGG19, neste caso) e com os mesmos parâmetros.

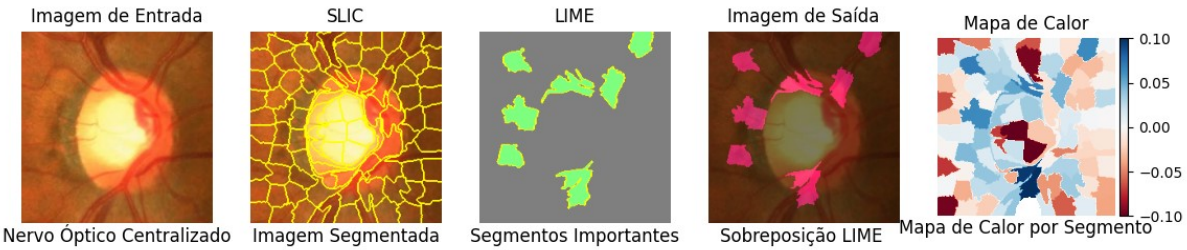


Figura 60 – Aplicação da técnica LIME na arquitetura VGG19 (primeira execução). Fonte: Elaborado pelo autor.

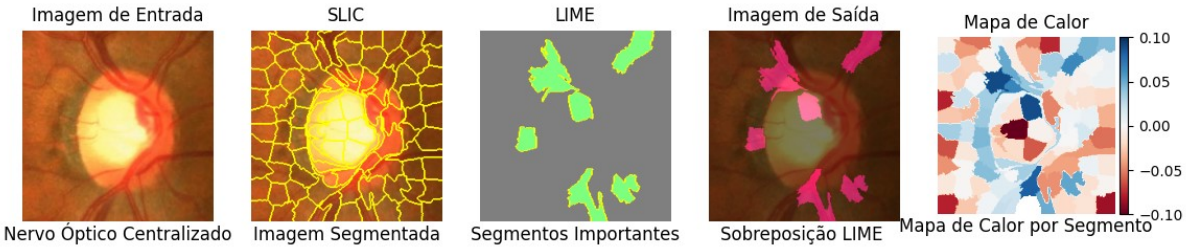


Figura 61 – Aplicação da técnica LIME na arquitetura VGG19 (segunda execução). Fonte: Elaborado pelo autor.

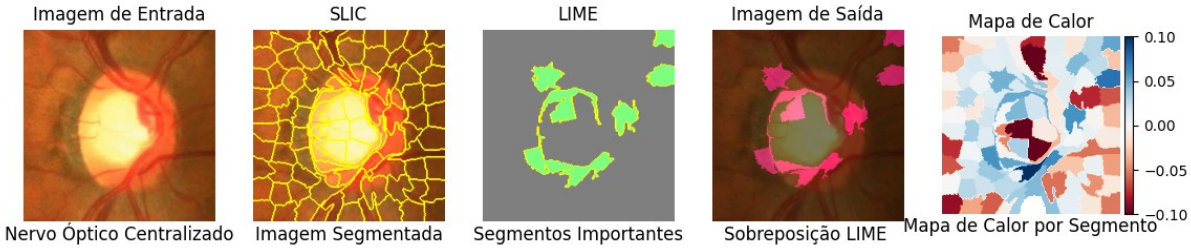


Figura 62 – Aplicação da técnica LIME na arquitetura VGG19 (terceira execução). Fonte: Elaborado pelo autor.

Observou-se que, para cada execução, a imagem de saída gera um resultado diferente (diferentes regiões destacadas). A variação nos resultados pode ser atribuída à natureza estocástica do processo de geração de explicações pelo LIME. O LIME opera introduzindo perturbações locais nos dados de entrada e observando como o modelo de DL reage a essas perturbações. A geração dessas perturbações é realizada de forma aleatória, o que leva a diferentes conjuntos de instâncias perturbadas a cada execução. Portanto, as variações nos resultados do LIME em execuções consecutivas são inerentes à abordagem estocástica do método, refletindo a sensibilidade da interpretação às pequenas variações nos dados de entrada. O tempo de processamento médio registrado para a explicação com o LIME foi de 517.019131 segundos.

As técnicas CAM e Grad-CAM apresentaram destaque notável em termos de eficiência computacional. O CAM obteve um desempenho ligeiramente superior ao Grad-CAM nesse quesito, com um tempo médio de processamento de 14.347031 segundos, enquanto o Grad-CAM registrou um tempo médio de processamento de 16.833995 segundos. As duas técnicas apresentaram falhas na geração do *heatmap* em alguma arquitetura. O CAM, por exemplo, não gerou o *heatmap* para a arquitetura ResNet50, conforme podemos observar nas Figuras 63 e 64. Uma hipótese para a não geração de resultados pode estar na complexidade da estrutura e as conexões residuais da ResNet50, que podem dificultar a correlação precisa entre características importantes e as camadas mais profundas da CNN. Além disso, a ResNet50 tem uma natureza não linear que pode complicar a inter-

pretação direta das ativações, dificultando a geração de mapas de calor significativos em áreas específicas das imagens de fundo de retina. Para superar essa limitação, estudos no sentido de modificar parâmetros ou procedimentos específicos da técnica CAM para melhor se adaptar à arquitetura ResNet50 são necessários. Isso pode envolver a otimização de pesos, a incorporação de informações específicas da própria arquitetura no processo de geração de mapas de calor, introduzir camadas adicionais ou modificar as camadas existentes para enfatizar características importantes. Outra alternativa seria realizar um pré-processamento mais específico nas imagens de entrada antes de aplicar a técnica CAM. Isso pode envolver o realce de características relevantes ou a normalização das imagens para melhorar a interpretação na arquitetura ResNet50.

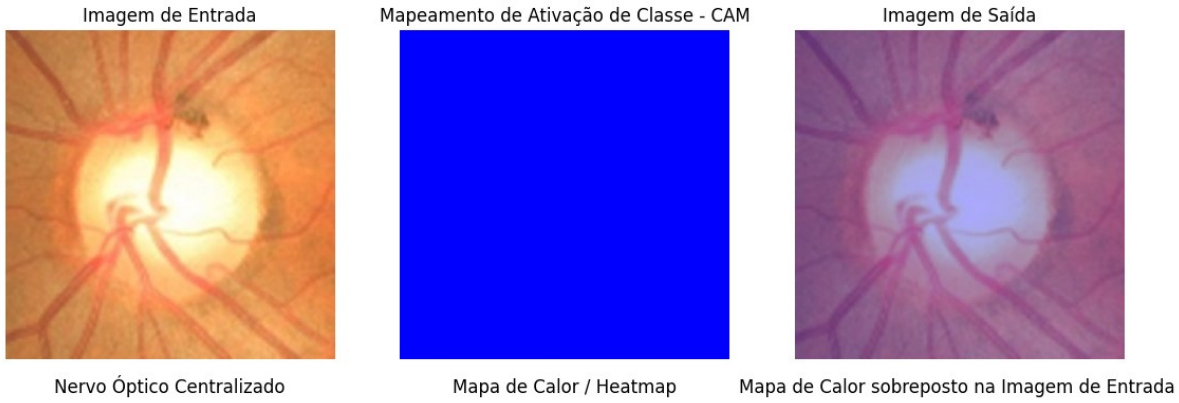


Figura 63 – Técnica CAM sem resultado na arquitetura ResNet50 (exemplo 1). Fonte: Elaborado pelo autor.

Figura 64 – Técnica CAM sem resultado na arquitetura ResNet50 (exemplo 2). Fonte: Elaborado pelo autor.

O Grad-CAM não gerou resultado (*heatmap*) para a arquitetura Xception, conforme podemos observar nas Figuras 65 e 66. Neste caso, uma hipótese para a não geração de resultados seria que a arquitetura Xception, ao explorar uma estrutura baseadas em blocos de convolução separável, pode introduzir caminhos de informação altamente intrincados, ou seja difíceis de seguir e compreender devido à complexidade da própria arquitetura, dificultando a identificação clara das características relevantes para uma interpretação precisa. Para superar essa limitação ao empregar a técnica Grad-CAM na arquitetura Xception, são necessários estudos exploratórios, envolvendo ajustes nos cálculos de gradientes ou na ponderação das ativações para realçar regiões de interesse em imagens de fundo de retina. Adicionalmente, propõe-se uma análise minuciosa da arquitetura Xception, identificando caminhos críticos e contemplando modificações específicas para facilitar a interpretação visual, levando em consideração a complexa estrutura dos blocos de convolução separável.

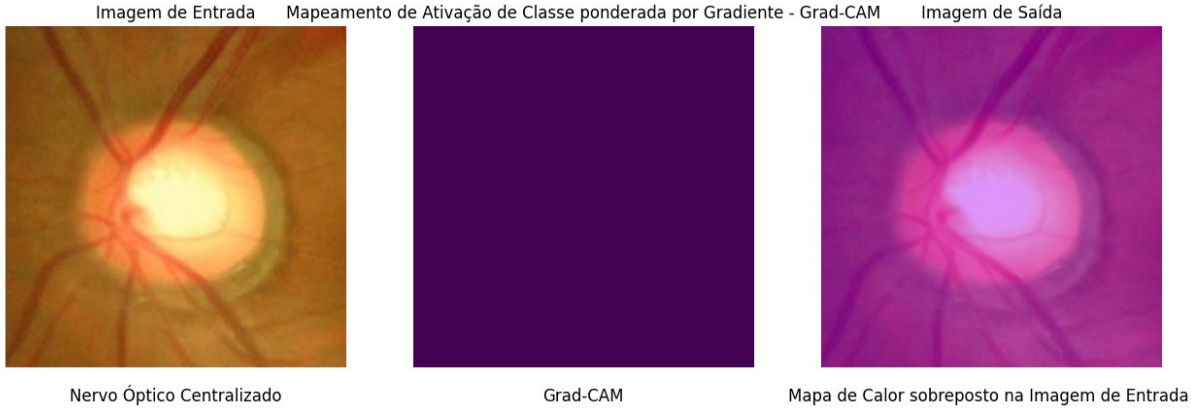


Figura 65 – Técnica Grad-CAM sem resultado na arquitetura Xception (exemplo 1). Fonte: Elaborado pelo autor.

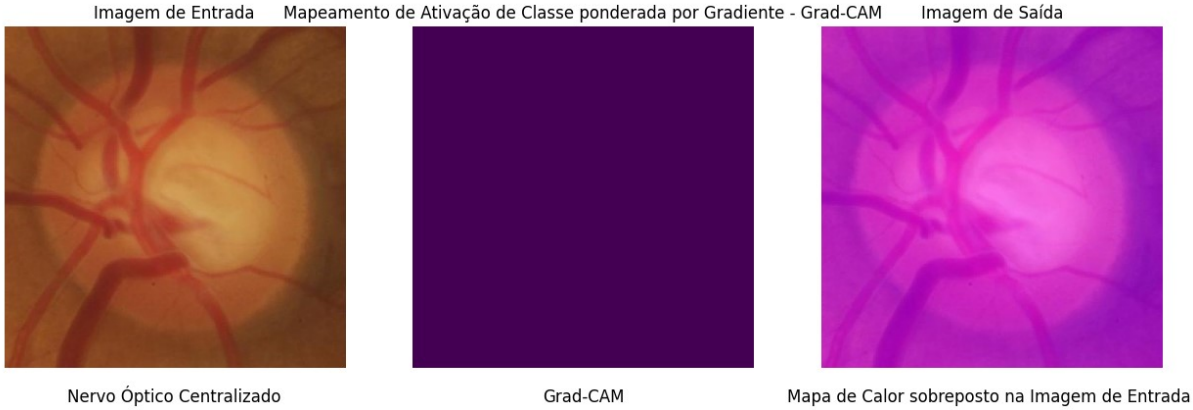


Figura 66 – Técnica Grad-CAM sem resultado na arquitetura Xception (exemplo 2). Fonte: Elaborado pelo autor.

Os métodos Vanilla Gradients e SmoothGrad apesar de estáveis, não demonstraram bons resultados, do ponto de vista clínico, conforme avaliação do especialista médico. O profissional, ao avaliar a interpretabilidade dessas técnicas, se posicionou em um viés neutro, indicando que os métodos não auxiliaram no entendimento da previsão dos modelos. O tempo médio de processamento registrado para Vanilla Gradients e SmoothGrad foi de 15.168250 segundos e 62.318061 segundos, respectivamente.

O SHAP se mostrou uma técnica estável, mantendo a consistência das explicações geradas para as mesmas imagens de entrada e apresentou uma boa interpretabilidade do ponto de vista clínico, de acordo com a avaliação do especialista. Para o SHAP aplicado em todas as arquiteturas de CNNs, o avaliador se posiciou em um viés de concordância parcial. O SHAP apresentou um desempenho inferior com relação a eficiência computacional registrando um tempo médio de processamento de 956.836887 segundos. Esse fato pode ser atribuído à natureza intrínseca e à metodologia de cálculo da própria técnica. O SHAP realiza uma abordagem de atribuição de valores baseada nos valores de *Shapley*, que exigem a avaliação de todas as combinações possíveis de características no conjunto de dados de treinamento. Esta característica computacionalmente intensiva, embora possa ser valiosa para uma interpretabilidade mais robusta, tende a aumentar significativamente o tempo de processamento.

A abordagem SCIM, ao empregar a convergência de duas técnicas distintas, revelou um tempo médio de processamento de 971.183918 segundos. Esta abordagem demonstrou boa estabilidade, uma adaptação eficaz nas arquiteturas VGG16, VGG19 e Xception devido as técnicas subjacentes e um viés de concordância parcial, de acordo com a avaliação do especialista. No caso de enfrentar desafios na geração dos *shap_values* positivos (SHAP) ou *heatmap* (CAM), conforme ilustrado na Figura 67 (ResNet50), o resultado pode ser comprometido. Quando não há regiões ou áreas de convergência entre as técnicas subjacentes, a imagem resultante não exibe qualquer marcação, indicando ausência de concordância entre os métodos base sobre as áreas cruciais para a predição, como evidenciado na Figura 68 (InceptionV3).

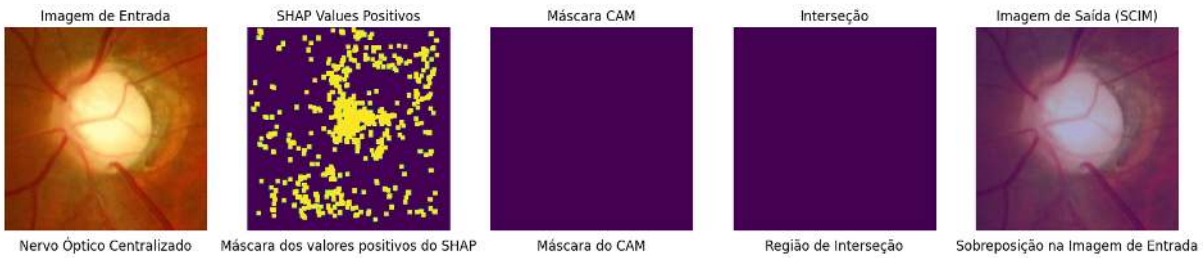


Figura 67 – Abordagem SCIM sem *heatmap* na arquitetura ResNet50. Fonte: Elaborado pelo autor.

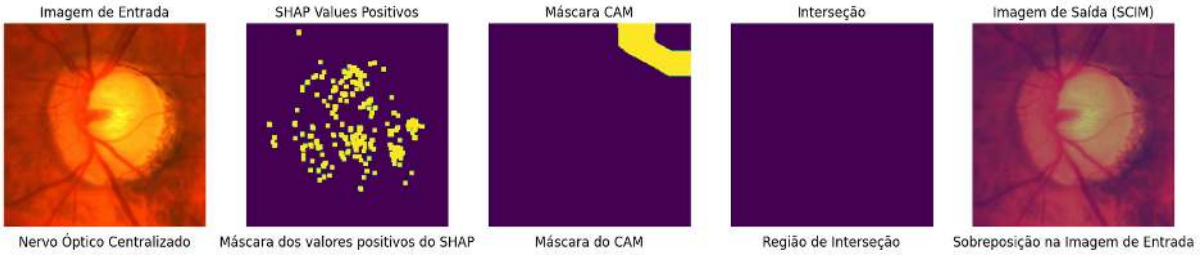


Figura 68 – Abordagem SCIM sem região de confluência na arquitetura InceptionV3. Fonte: Elaborado pelo autor.

5.2 Avaliação Crítica dos Resultados

Os resultados obtidos nesta pesquisa levantam uma hipótese intrigante acerca da eficácia diferenciada da interpretabilidade nas arquiteturas VGG16 e VGG19, em comparação com as outras arquiteturas de CNN investigadas. A aplicação da técnica de mapeamento de ativação de classe (CAM) nestes dois modelos destacou-se de maneira notável, recebendo a pontuação máxima, de concordância total, por parte do especialista durante o processo de avaliação. A análise de desempenho fortalece a robustez dessas arquiteturas, com a VGG16 alcançando uma média harmônica entre precisão e sensibilidade (*f1-score*) de 97,00% e a VGG19 registrando 96,26%. Esses resultados sugerem que a incorporação da interpretabilidade nas arquiteturas VGG16 e VGG19 pode resultar em um aumento significativo da confiabilidade na classificação do glaucoma em imagens de retinografia, contribuindo para o apoio ao diagnóstico.

A utilização da técnica CAM nas arquiteturas VGG16 e VGG19 revelou uma notável capacidade em destacar regiões clinicamente relevantes, especialmente as disparidades no padrão ISNT, uma característica importante na identificação da condição glaucomatosa. Essa técnica proporcionou uma interpretação visual mais clara e intuitiva, alinhando-se de forma coesa com as normativas estabelecidas no âmbito clínico.

Apesar de apresentar um desempenho inferior em relação às arquiteturas VGG16 e VGG19, com *f1-score* registrado de 72,34%, a arquitetura Xception revelou-se relevante no contexto da interpretabilidade. A aplicação da técnica CAM nesta arquitetura ressaltou a região da escavação óptica aumentada, caracterizada pelo afinamento da borda neurorretiniana, aumento da proporção copa-disco (CDR) e alongamento vertical da escavação. Estes achados sugerem que, mesmo diante de resultados menos precisos, a arquitetura Xception oferece *insights* valiosos para compreender o raciocínio subjacente que orientou o modelo na predição diagnóstica.

A arquitetura ResNet50 apresentou um desempenho considerável com *f1-score* de 94,20%, no entanto, a eficácia em termos de interpretabilidade revelou-se limitada. Isso se evidencia pela incapacidade da aplicação da técnica CAM em gerar mapas de calor que

pudessem proporcionar orientação clara e intuitiva ao raciocínio da CNN para a predição do modelo.

A abordagem SCIM, proposta no âmbito desta pesquisa, torna-se um elo promissor para a aceitação clínica. Ao fornecer uma representação visual clara dos fatores que influenciam as decisões dos modelos, esta abordagem permite que os profissionais de saúde compreendam não apenas o resultado final, mas também o raciocínio subjacente. Este aspecto não apenas promove um ambiente de confiança, mas também contribui significativamente para o aumento da confiabilidade dos resultados, uma vez que os médicos podem verificar e validar as conclusões do modelo de forma transparente.

6 Conclusão

Modelos de aprendizado de máquina têm demonstrado um potencial significativo em diversas áreas, incluindo a saúde, onde são utilizados para aprimorar o diagnóstico de doenças, como o glaucoma. Contudo, a falta de explicabilidade desses modelos tem sido uma limitação para sua adoção generalizada na prática clínica. Este trabalho abordou essa questão ao explorar técnicas de XAI em CNNs para a classificação do glaucoma. Os experimentos conduzidos, com a colaboração de um especialista com mais de sete anos de experiência na área de oftalmologia, revelaram a eficácia da técnica baseada em mapeamento de ativação de classe (CAM) aplicada nas arquiteturas VGG16 e VGG19. Isso sugere que ao incorporar interpretabilidade nessas arquiteturas, é possível aumentar significativamente a confiabilidade na classificação do glaucoma em imagens de retinografia para apoio ao diagnóstico. A interpretabilidade facilitada pela abordagem SCIM, proposta neste trabalho, torna-se um elo promissor para a aceitação clínica. Ao fornecer uma representação visual clara dos fatores que influenciam as decisões dos modelos, por meio dessa abordagem, os profissionais de saúde podem entender não apenas o resultado final, mas também o raciocínio subjacente, promovendo um ambiente de confiança, pois os médicos podem verificar e validar as conclusões do modelo de forma transparente.

6.1 Contribuições e Limitações

Esta seção destaca as principais contribuições e identifica as limitações inerentes à pesquisa, fornecendo uma visão abrangente do impacto e alcance do estudo em questão. Algumas das contribuições essenciais deste trabalho de pesquisa são delineadas a seguir:

- A exploração e aplicação de técnicas de XAI no contexto do diagnóstico automatizado do glaucoma têm o potencial de avançar as técnicas de interpretabilidade em modelos de ML. Isso possibilita o desenvolvimento de abordagens mais confiáveis e compreensíveis, permitindo que os médicos entendam e confiem nas decisões tomadas pelos modelos de CNNs.
- O estudo comparativo das arquiteturas de CNNs aplicadas à classificação do glaucoma pode fornecer *insights* valiosos para o aprimoramento dessas redes, identificando características e camadas relevantes para a detecção da doença.
- A proposta de uma nova abordagem, a SCIM, para a interpretação visual dos resultados dos modelos de CNNs representa uma contribuição tecnológica inovadora, com potencial de aplicação não apenas no diagnóstico do glaucoma, mas também em outras áreas de aplicação das CNNs.

- Com os resultados obtidos nesse estudo, destacando a eficácia do CAM nas arquiteturas VGG16 e VGG19 para a interpretabilidade e apoio ao diagnóstico, é possível fornecer uma direção clara para futuras implementações clínicas.
- Ao abordar a falta de explicabilidade das decisões dos modelos e fornecer um estudo mais aprofundado para essa limitação, o trabalho contribui para o avanço do uso de técnicas de ML na medicina, permitindo a adoção mais ampla da IA na prática clínica e impulsionando o campo da saúde computacional.

A seguir, são enumeradas algumas das limitações que foram consideradas no contexto desta pesquisa:

- A disponibilidade de dados de imagens de fundo de retina de alta qualidade e em quantidade suficiente pode representar um desafio, assim como a escassez de dados rotulados de pacientes com glaucoma. Essa limitação pode afetar a capacidade de treinamento e avaliação adequada dos modelos de ML, impactando os resultados do estudo.
- Os resultados obtidos com as técnicas de interpretabilidade e as arquiteturas de CNNs utilizadas podem ser específicos para o contexto do diagnóstico do glaucoma, dificultando sua generalização para outras doenças oculares ou áreas de aplicação da IA na medicina.
- As técnicas de interpretabilidade empregadas, como CAM e a abordagem proposta SCIM, embora tenham apresentado resultados promissores, podem ter suas próprias limitações, resultando em explicações que podem não ser completamente compreensíveis ou confiáveis para médicos e pacientes.
- A escolha das arquiteturas de CNNs e dos hiperparâmetros pode influenciar significativamente os resultados, tornando as conclusões específicas para as configurações utilizadas e limitando a generalização para outras arquiteturas ou configurações diferentes.
- A interpretabilidade, embora tenha sido avaliada com o apoio de especialista, ainda está sujeita a variações na interpretação humana. A subjetividade nesse processo pode impactar a validade das conclusões obtidas.
- A avaliação completa da utilidade dos métodos propostos na prática clínica pode ser limitada devido a restrições de tempo, recursos e complexidade do próprio ambiente clínico. Esses fatores podem afetar a extensão e a abrangência dos estudos clínicos a serem realizados, dificultando uma avaliação completa da viabilidade e eficácia dos métodos propostos.

6.2 Trabalhos Futuros

Esta seção destaca possíveis direções para pesquisas futuras, visando ampliar e aprofundar as contribuições desta dissertação, bem como superar as limitações identificadas. Embora os experimentos tenham alcançado resultados significativos e promissores na aplicação de técnicas de XAI para a classificação do glaucoma, existem algumas direções futuras que podem e devem ser exploradas. Algumas dessas direções futuras são:

- Necessidade de realizar experimentos com outras arquiteturas de CNN e de explorar outras técnicas de XAI. A análise comparativa entre outras arquiteturas de CNNs e outras técnicas de XAI permitirá aumentar o grau de compreensão de como diferentes estruturas respondem à interpretabilidade e contribuem para o apoio ao diagnóstico clínico;
- Investir em pesquisas que busquem reduzir a subjetividade na avaliação humana da interpretabilidade. Desenvolver métricas mais objetivas e protocolos padronizados para a análise visual pode contribuir para uma avaliação mais consistente;
- Explorar a integração de outras técnicas de XAI com a abordagem SCIM. Essa integração pode aprimorar a confiabilidade do diagnóstico ao combinar informações de diferentes técnicas promovendo mais transparência e entendimento no processo de decisão dos modelos de DL;
- A condução de validações clínicas requer uma ampliação na participação de profissionais especializados, particularmente na área de oftalmologia. A colaboração com um número significativamente maior de especialistas de domínio proporcionará uma análise mais abrangente, enriquecendo a validação clínica e fortalecendo as bases para a aplicação prática desses modelos na prática médica; e
- Possibilidade de incorporar um método de análise qualitativa complementar. Isso se justifica pela dificuldade em obter avaliadores humanos (profissionais em oftalmologia) para a interpretabilidade dos resultados.

Essas sugestões para futuras pesquisas têm como finalidade aprofundar a compreensão prática das técnicas interpretáveis no diagnóstico médico assistido por IA. Dessa forma, busca-se impulsionar avanços significativos, contribuindo para a constante evolução desse campo de estudo.

6.3 Trabalhos Submetidos e Aceitos para Publicação

Durante o desenvolvimento desta pesquisa, foram submetidos alguns trabalhos a periódicos científicos e/ou conferências renomadas, buscando a disseminação dos resulta-

dos obtidos. Neste contexto, destacam-se os trabalhos aceitos para publicação:

- ***Automatic features extraction from the optic cup and disc segmentation for Glaucoma classification***: Este trabalho foi aceito para publicação no *ICCSA 2023 - International Conference on Computational Science and Its Applications*. Nesse artigo, foi proposto um método computacional de baixo custo para extrair características da anatomia do nervo óptico (escavação óptica e segmentação do disco) por meio do processamento de imagens de fundo de retina, que é utilizado em conjunto com algoritmos de classificação de baixo custo computacional (máquina de vetor de suporte - SVM) demonstrando a capacidade de realizar diagnósticos precisos. Os atributos mais dominantes foram identificados por meio de análises de explicações adaptativas modeladas (SHAP) e de explicações agnósticas de modelos interpretáveis locais (LIME). Autores: Marcus Oliveira, Cleverson Vieira, Ana Paula De Filippo, Michel Carlo Rodrigues Leles, Diego Dias, Marcelo Guimarães, Elisa Tuler e Leonardo Rocha. O artigo encontra-se disponível em https://link.springer.com/chapter/10.1007/978-3-031-36805-9_36
- ***Applied Explainable Artificial Intelligence (XAI) in the classification of retinal images for support in the diagnosis of Glaucoma***: Este trabalho foi submetido e aceito para publicação no WebMedia 2023 - Simpósio Brasileiro de Sistemas Multimídia e Web. Neste trabalho, foi explorada a aplicação de técnicas de XAI em diferentes arquiteturas de CNNs para a classificação de glaucoma em imagens de retinografia e realizada uma comparação em relação a quais métodos fornecem os melhores recursos para a análise e interpretação humana, servindo de apoio no diagnóstico do glaucoma. Uma abordagem baseada na confluência de outras duas técnicas, para interpretação visual, denominada SCIM é proposta demonstrando resultados promissores. Os experimentos indicam, em um olhar não clínico, que a técnica de interpretabilidade baseada em mapeamento de ativação de classe ponderada por gradiente (Grad-CAM), bem como a abordagem proposta (SCIM), aplicadas à arquitetura VGG19, fornecem os melhores recursos para a interpretabilidade humana e apoio ao diagnóstico do glaucoma. Autores: Cleverson Vieira, Marcus Oliveira, Marcelo Guimarães, Leonardo Rocha e Diego Dias. O artigo encontra-se disponível em https://dl.acm.org/doi/abs/10.1145/3617023.3617026

Os trabalhos aceitos para publicação representam não apenas os resultados alcançados, mas também a relevância e inovação que esta pesquisa proporcionou ao panorama científico.

Referências

BARROS, D. M. S.; MOURA, J. C. C.; FREIRE, C. R.; TALEB, A. C.; VALENTIM, R. A. M.; MORAIS, P. S. G. Machine learning applied to retinal image processing for glaucoma detection: review and perspective. *BioMedical Engineering OnLine*, v. 19, n. 1, p. 20, Apr 2020. ISSN 1475-925X. Disponível em: <https://doi.org/10.1186/s12938-020-00767->. Citado 2 vezes nas páginas 9 e 23.

OLIVEIRA, M.; VIEIRA, C.; FILIPPO, A. P. D.; LELES, M. C. R.; DIAS, D.; GUIMARÃES, M.; TULER, E.; ROCHA, L. Automatic features extraction from the optic cup and disc segmentation for glaucoma classification. In: GERVASI, O.; MURGANTE, B.; TANIAR, D.; APDUHAN, B. O.; BRAGA, A. C.; GARAU, C.; STRATIGEA, A. (Ed.). *Computational Science and Its Applications – ICCSA 2023*. Cham: Springer Nature Switzerland, 2023. p. 550–563. ISBN 978-3-031-36805-9. Citado 3 vezes nas páginas 9, 23 e 54.

HARIZMAN, N.; OLIVEIRA, C.; CHIANG, A.; TELLO, C.; MARMOR, M.; RITCH, R.; LIEBMANN, J. M. The ISNT rule and differentiation of normal from glaucomatous eyes. *Arch Ophthalmol*, United States, v. 124, n. 11, p. 1579–1583, nov. 2006. Citado 2 vezes nas páginas 9 e 24.

MONARD, M. C.; BARANAUSKAS, J. A. Aplicações de inteligência artificial: uma visão geral. In: *Congresso de Lógica Aplicada à Tecnologia*. [S.l.]: Faculdade SENAC de Ciências Exatas e Tecnologia, 2000. Citado 2 vezes nas páginas 9 e 25.

HIDAKA, A.; KURITA, T. Consecutive dimensionality reduction by canonical correlation analysis for visualization of convolutional neural networks. *Proceedings of the ISCIE International Symposium on Stochastic Systems Theory and its Applications*, v. 2017, p. 160–167, 2017. Citado 2 vezes nas páginas 9 e 28.

MILLER, T. Explanation in artificial intelligence: Insights from the social sciences. *CoRR*, abs/1706.07269, 2017. Disponível em: <https://arxiv.org/abs/1706.0726>. Citado 2 vezes nas páginas 9 e 28.

MUHAMMAD, M. B.; YEASIN, M. Eigen-cam: Class activation map using principal components. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020. Disponível em: <http://dx.doi.org/10.1109/IJCNN48605.2020-.920662>http://dx.doi.org/10.1109/IJCNN48605.2020.920662. Citado 2 vezes nas páginas 9 e 33.

MOLNAR, C. *Interpretable Machine Learning*: A guide for making black box models explainable. 2. ed. [s.n.], 2022. Disponível em: <https://christophm.github-.io/interpretable-ml-boo>. Citado 7 vezes nas páginas 9, 29, 30, 31, 35, 49 e 80.

JERE, M.; KUMAR, M.; KOUSHANFAR, F. *A Singular Value Perspective on Model Robustness*. 12 2020. Citado 2 vezes nas páginas 9 e 37.

SMILKOV, D.; THORAT, N.; KIM, B.; VIÉGAS, F. B.; WATTENBERG, M. Smoothgrad: removing noise by adding noise. *ArXiv*, abs/1706.03825, 2017. Disponível em: <https://api.semanticscholar.org/CorpusID:1169587>. Citado 2 vezes nas páginas 9 e 38.

AGARWAL, N.; DAS, S. Interpretable machine learning tools: A survey. In: *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. [S.l.: s.n.], 2020. p. 1528–1534. Citado 3 vezes nas páginas 18, 28 e 32.

NAZIR, S.; DICKSON, D. M.; AKRAM, M. U. Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. *Computers in Biology and Medicine*, v. 156, p. 106668, 2023. ISSN 0010-4825. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0010482523001333>. Citado 2 vezes nas páginas 18 e 28.

KUMAR, D.; TAYLOR, G. W.; WONG, A. Discovery radiomics with clear-dr: Interpretable computer aided diagnosis of diabetic retinopathy. *IEEE Access*, v. 7, p. 25891–25896, 2019. Citado 3 vezes nas páginas 18, 45 e 47.

DIAZ-PINTO, A.; MORALES, S.; NARANJO, V.; KÖHLER, T.; MOSSI, J. M.; NAVEA, A. Cnns for automatic glaucoma assessment using fundus images: an extensive validation. *BioMedical Engineering OnLine*, v. 18, n. 1, p. 29, Mar 2019. ISSN 1475-925X. Disponível em: <https://doi.org/10.1186/s12938-019-0649->. Citado 6 vezes nas páginas 18, 19, 40, 42, 51 e 54.

SARHAN, M. H.; NASSERI, M. A.; ZAPP, D.; MAIER, M.; LOHMANN, C. P.; NAVAB, N.; ESLAMI, A. Machine learning techniques for ophthalmic data processing: A review. *IEEE Journal of Biomedical and Health Informatics*, v. 24, n. 12, p. 3338–3350, 2020. Citado 2 vezes nas páginas 18 e 22.

STEFAN, A.-M.; PARASCHIV, E.-A.; OVREIU, S.; OVREIU, E. *A Review of Glaucoma Detection from Digital Fundus Images using Machine Learning Techniques*. 2020. 1-4 p. Citado 2 vezes nas páginas 18 e 22.

NOROUZIFARD, M.; NEMATI, A.; GHOLAMHOSSEINI, H.; KLETTE, R.; NOURI-MAHDAVI, K.; YOUSEFI, S. Automated glaucoma diagnosis using deep and transfer learning: Proposal of a system for clinical testing. In: *2018 International Conference on Image and Vision Computing New Zealand (IVCNZ)*. [S.l.: s.n.], 2018. p. 1–6. Citado 2 vezes nas páginas 19 e 41.

GóMEZ-VALVERDE, J. J.; ANTóN, A.; FATTI, G.; LIEFERS, B.; HERRANZ, A.; SANTOS, A.; SáNCHEZ, C. I.; LEDESMA-CARBAYO, M. J. Automatic glaucoma classification using color fundus images based on convolutional neural networks and transfer learning. *Biomed. Opt. Express*, Optica Publishing Group, v. 10, n. 2, p. 892–913, Feb 2019. Disponível em: <https://opg.optica.org/boe/abstract.cfm?URI=boe-10-2-89>. Citado 3 vezes nas páginas 19, 41 e 42.

SERENER, A.; SERTE, S. Transfer learning for early and advanced glaucoma detection with convolutional neural networks. In: *2019 Medical Technologies Congress (TIPTEKNO)*. [S.l.: s.n.], 2019. p. 1–4. Citado 2 vezes nas páginas 19 e 41.

MARTINS, J.; CARDOSO, J. S.; SOARES, F. Offline computer-aided diagnosis for glaucoma detection using fundus images targeted at mobile devices. *Computer Methods and Programs in Biomedicine*, v. 192, p. 105341, 2020. ISSN 0169-2607. Disponível em: <https://www.sciencedirect.com/science/article/pii/S016926071931201>. Citado 3 vezes nas páginas 19, 41 e 42.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. v. 30. Disponível em: <https://proceedings-.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pd>. Citado 2 vezes nas páginas 19 e 35.

ZHOU, B.; KHOSLA, A.; LAPEDRIZA, A.; OLIVA, A.; TORRALBA, A. Learning deep features for discriminative localization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2016. Citado 2 vezes nas páginas 19 e 32.

SIMONYAN, K.; ZISSERMAN, A. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. Citado na página 19.

SZEGEDY, C.; VANHOUCKE, V.; IOFFE, S.; SHLENS, J.; WOJNA, Z. *Rethinking the Inception Architecture for Computer Vision*. 2015. Citado na página 20.

HUANG, G.; LIU, Z.; MAATEN, L. van der; WEINBERGER, K. Q. *Densely Connected Convolutional Networks*. 2018. Citado na página 20.

CHOLLET, F. Xception: Deep learning with depthwise separable convolutions. *CoRR*, abs/1610.02357, 2016. Disponível em: <https://arxiv.org/abs/1610.0235>. Citado na página 20.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. Disponível em: <https://arxiv.org/abs/1512.0338>. Citado na página 20.

SELVARAJU, R. R.; COGSWELL, M.; DAS, A.; VEDANTAM, R.; PARIKH, D.; BATRA, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. [S.l.: s.n.], 2017. Citado 2 vezes nas páginas 20 e 33.

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*. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 1135–1144. ISBN 9781450342322. Disponível em: <https://doi.org/10.1145/2939672.293977>. Citado 2 vezes nas páginas 20 e 34.

SIMONYAN, K.; VEDALDI, A.; ZISSERMAN, A. *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. 2014. Citado 2 vezes nas páginas 20 e 36.

SMILKOV, D.; THORAT, N.; KIM, B.; VIéGAS, F.; WATTENBERG, M. *SmoothGrad: removing noise by adding noise*. 2017. Citado 2 vezes nas páginas 20 e 37.

MANASSAKORN, A.; AUETHAVEKIAT, S.; SA-ING, V.; CHANSANGPETCH, S.; RATANAWONGPHAIBUL, K.; URAMPHORN, N.; TANTISEVI, V. Glaunet: Glaucoma diagnosis for octa imaging using a new cnn architecture. *IEEE Access*, v. 10, p. 95613–95622, 2022. Citado na página 22.

ZANGALLI, C.; GUPTA, S. R.; SPAETH, G. L. The disc as the basis of treatment for glaucoma. *Saudi J Ophthalmol*, India, v. 25, n. 4, p. 381–387, jul. 2011. Citado na página 22.

KUMAR, B. N.; CHAUHAN, R. P.; DAHIYA, N. Detection of glaucoma using image processing techniques: A review. In: *2016 International Conference on Microelectronics, Computing and Communications (MicroCom)*. [S.l.]: IEEE, 2016. p. 1–6. Citado na página 23.

ZHAO, R.; CHEN, X.; LIU, X.; CHEN, Z.; GUO, F.; LI, S. Direct cup-to-disc ratio estimation for glaucoma screening via semi-supervised learning. *IEEE Journal of Biomedical and Health Informatics*, v. 24, n. 4, p. 1104–1113, 2020. Citado na página 23.

VESSANI, R. M. *Comparação entre diversas técnicas de imagem para diagnóstico do glaucoma*. [S.l.]: Doctoral Thesis in Oftalmologia, 2008. Faculdade de Medicina, Universidade de São Paulo. Citado na página 24.

NORVIG, P.; RUSSELL, S. *Inteligência Artificial*. ELSEVIER EDITORA, 2013. ISBN 9788535237016. Disponível em: <https://books.google.com.br-/books?id=KhUQvgAACAA>. Citado na página 24.

HAUGELAND, J. *Artificial Intelligence: The Very Idea*. [S.l.]: Cambridge: MIT Press, 1985. Citado na página 24.

KURZWEIL, R. *The Age of Intelligent Machines*. [S.l.]: MIT Press, 1990. Citado na página 24.

CHARNIAK, E.; MCDERMOTT, D. *Introduction to Artificial Intelligence*. [S.l.]: Addison Wesley, 1985. Citado na página 24.

POOLE, D.; MACKWORTH, A.; GOEBEL, R. *Computational Intelligence: A Logical Approach*. [S.l.: s.n.], 1998. ISBN 978-0-19-510270-3. Citado na página 24.

FLECK, L.; TAVARES, M. H. F.; EYNG, E.; HELMANN, A. C.; ANDRADE, M. A. d. M. Redes neurais artificiais: Princípios básicos. *Revista Eletrônica Científica Inovação e Tecnologia*, v. 1, n. 13, p. 47–57, 2016. Citado 2 vezes nas páginas 25 e 26.

HAYKIN, S. *Redes Neurais: Princípios e Prática*. Bookman Editora, 2001. ISBN 9788577800865. Disponível em: <https://books.google.com.br-/books?id=bhMwDwAAQBA>. Citado na página 25.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. https://www.deeplearningbook.org. Citado 2 vezes nas páginas 26 e 27.

NIELSEN, M. A. misc, *Neural Networks and Deep Learning*. Determination Press, 2018. Disponível em: <https://neuralnetworksanddeeplearning.com>. Citado na página 26.

BAJWA, M. N.; MALIK, M. I.; SIDDIQUI, S. A.; DENGEL, A.; SHAFAIT, F.; NEUMEIER, W.; AHMED, S. Two-stage framework for optic disc localization and glaucoma classification in retinal fundus images using deep learning. *BMC Medical Informatics and Decision Making*, v. 19, n. 1, p. 136, Jul 2019. ISSN 1472-6947. Disponível em: <https://doi.org/10.1186/s12911-019-0842->. Citado 3 vezes nas páginas 27, 39 e 42.

LALONDE, R.; TORIGIAN, D.; BAGCI, U. Encoding visual attributes in capsules for explainable medical diagnoses. In: MARTEL, A. L.; ABOLMAESUMI, P.; STOYANOV, D.; MATEUS, D.; ZULUAGA, M. A.; ZHOU, S. K.; RACOCEANU, D.; JOSKOWICZ, L. (Ed.). *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Cham: Springer International Publishing, 2020. p. 294–304. ISBN 978-3-030-59710-8. Citado na página 28.

GILL, P. *Introduction to Machine Learning Interpretability*. O'Reilly Media, Incorporated, 2018. ISBN 9781492033158. Disponível em: <https://books.google.com-.br/books?id=4CyVtgEACAA>. Citado na página 29.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. *Model-Agnostic Interpretability of Machine Learning*. 2016. Citado na página 29.

LIPTON, Z. C. The mythos of model interpretability. *CoRR*, abs/1606.03490, 2016. Disponível em: <https://arxiv.org/abs/1606.0349>. Citado 2 vezes nas páginas 30 e 31.

ARRIETA, A. B.; RODRÍGUEZ, N. D.; SER, J. D.; BENNETOT, A.; TABIK, S.; BARBADO, A.; GARCÍA, S.; GIL-LOPEZ, S.; MOLINA, D.; BENJAMINS, R.; CHATILA, R.; HERRERA, F. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *CoRR*, abs/1910.10045, 2019. Disponível em: <https://arxiv.org/abs/1910.1004>. Citado na página 30.

DOSHI-VELEZ, F.; KIM, B. *Towards A Rigorous Science of Interpretable Machine Learning*. 2017. Citado 3 vezes nas páginas 31, 50 e 80.

CAMARA, J.; NETO, A.; PIRES, I. M.; VILLASANA, M. V.; ZDRAVEVSKI, E.; CUNHA, A. Literature review on artificial intelligence methods for glaucoma screening, segmentation, and classification. *Journal of Imaging*, v. 8, n. 2, 2022. ISSN 2313-433X. Disponível em: <https://www.mdpi.com/2313-433X/8/2/1>. Citado 2 vezes nas páginas 39 e 42.

van der Velden, B. H.; KUIJF, H. J.; GILHUIJS, K. G.; VIERGEVER, M. A. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, v. 79, p. 102470, 2022. ISSN 1361-8415. Disponível em: <https://www.sciencedirect.com/science/article/pii/S136184152200117>. Citado 3 vezes nas páginas 39, 43 e 46.

SHYAMALEE, T.; MEEDENIYA, D. Cnn based fundus images classification for glaucoma identification. In: *2022 2nd International Conference on Advanced Research in Computing (ICARC)*. [S.l.: s.n.], 2022. p. 200–205. Citado na página 39.

SRENG, S.; MANEERAT, N.; HAMAMOTO, K.; WIN, K. Y. Deep learning for optic disc segmentation and glaucoma diagnosis on retinal images. *Applied Sciences*, v. 10, n. 14, 2020. ISSN 2076-3417. Disponível em: <https://www.mdpi.com/2076-3417/10/14/491>. Citado 2 vezes nas páginas 39 e 42.

Aziz-ur-Rehman; TAJ, I. A.; SAJID, M.; KARIMOV, K. S. An ensemble framework based on deep cnns architecture for glaucoma classification using fundus photography. *Mathematical Biosciences and Engineering*, v. 18, n. 5, p. 5321–5346, 2021. ISSN 1551-0018. Disponível em: <https://www.aimspress.com/article/doi/10.3934/mbe.202127>. Citado 2 vezes nas páginas 40 e 42.

CHAI, Y.; LIU, H.; XU, J. Glaucoma diagnosis based on both hidden features and domain knowledge through deep learning models. *Knowledge-Based Systems*, v. 161, p. 147–156, 2018. ISSN 0950-7051. Disponível em: <https://www.sciencedirect-.com/science/article/pii/S095070511830394>. Citado 2 vezes nas páginas 40 e 42.

BAJWA, M. N.; SINGH, G. A. P.; NEUMEIER, W.; MALIK, M. I.; DENGEL, A.; AHMED, S. G1020: A benchmark retinal fundus image dataset for computer-aided glaucoma detection. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. [S.l.: s.n.], 2020. p. 1–7. Citado 2 vezes nas páginas 41 e 42.

LI, F.; SONG, D.; CHEN, H.; XIONG, J.; LI, X.; ZHONG, H.; TANG, G.; FAN, S.; LAM, D. S. C.; PAN, W.; ZHENG, Y.; LI, Y.; QU, G.; HE, J.; WANG, Z.; JIN, L.; ZHOU, R.; SONG, Y.; SUN, Y.; CHENG, W.; YANG, C.; FAN, Y.; LI, Y.; ZHANG, H.; YUAN, Y.; XU, Y.; XIONG, Y.; JIN, L.; LV, A.; NIU, L.; LIU, Y.; LI, S.; ZHANG, J.; ZANGWILL, L. M.; FRANGI, A. F.; AUNG, T.; CHENG, C.-y.; QIAO, Y.; ZHANG, X.; TING, D. S. W. Development and clinical deployment of a smartphone-based visual field deep learning system for glaucoma detection. *npj Digital Medicine*, v. 3, n. 1, p. 123, Sep 2020. ISSN 2398-6352. Disponível em: <https://doi.org/10.1038/s41746-020-00329->. Citado na página 42.

OLDEN, J. D.; JOY, M. K.; DEATH, R. G. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, v. 178, n. 3, p. 389–397, 2004. ISSN 0304-3800. Disponível em: <https://www.sciencedirect.com/science/article/pii/S030438000400156>. Citado 2 vezes nas páginas 43 e 47.

TSANG, M.; CHENG, D.; LIU, Y. *Detecting Statistical Interactions from Neural Network Weights*. 2018. Citado 2 vezes nas páginas 43 e 47.

FONG, R. C.; VEDALDI, A. Interpretable explanations of black boxes by meaningful perturbation. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. [S.l.: s.n.], 2017. Citado 2 vezes nas páginas 44 e 47.

MENG, Q.; HASHIMOTO, Y.; SATOH, S. How to extract more information with less burden: Fundus image classification and retinal disease localization with ophthalmologist intervention. *IEEE Journal of Biomedical and Health Informatics*, v. 24, n. 12, p. 3351–3361, 2020. Citado 3 vezes nas páginas 44, 46 e 47.

AHMAD, M.; KASUKURTHI, N.; PANDE, H. Deep learning for weak supervision of diabetic retinopathy abnormalities. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. [S.l.: s.n.], 2019. p. 573–577. Citado 3 vezes nas páginas 44, 46 e 47.

JANG, Y.; SON, J.; PARK, K. H.; PARK, S. J.; JUNG, K.-H. Laterality classification of fundus images using interpretable deep neural network. *Journal of Digital*

Imaging, v. 31, n. 6, p. 923–928, Dec 2018. ISSN 1618-727X. Disponível em: <https://doi.org/10.1007/s10278-018-0099->. Citado 2 vezes nas páginas 44 e 47.

COSTA, P.; ARAúJO, T.; ARESTA, G.; GALDRAN, A.; MENDONçA, A. M.; SMAILAGIC, A.; CAMPILHO, A. Eyewes: Weakly supervised pre-trained convolutional neural networks for diabetic retinopathy detection. In: *2019 16th International Conference on Machine Vision Applications (MVA)*. [S.l.: s.n.], 2019. p. 1–6. Citado 2 vezes nas páginas 45 e 47.

THAKOOR, K. A.; LI, X.; TSAMIS, E.; SAJDA, P.; HOOD, D. C. Enhancing the accuracy of glaucoma detection from oct probability maps using convolutional neural networks. In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. [S.l.: s.n.], 2019. p. 2036–2040. Citado 2 vezes nas páginas 45 e 47.

ZHOU, K.; GAO, S.; CHENG, J.; GU, Z.; FU, H.; TU, Z.; YANG, J.; ZHAO, Y.; LIU, J. Sparse-gan: Sparsity-constrained generative adversarial network for anomaly detection in retinal oct image. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. [S.l.: s.n.], 2020. p. 1227–1231. Citado 2 vezes nas páginas 46 e 47.

LI, L.; XU, M.; LIU, H.; LI, Y.; WANG, X.; JIANG, L.; WANG, Z.; FAN, X.; WANG, N. A large-scale database and a cnn model for attention-based glaucoma detection. *IEEE Transactions on Medical Imaging*, v. 39, n. 2, p. 413–424, 2020. Citado 2 vezes nas páginas 46 e 47.

BAJWA, M. N.; SINGH, G. A. P.; NEUMEIER, W.; MALIK, M. I.; DENGEL, A.; AHMED, S. *G1020: A Benchmark Retinal Fundus Image Dataset for Computer-Aided Glaucoma Detection*. 2020. Citado 2 vezes nas páginas 50 e 51.

ISLAM, M. T.; MASHFU, S. T.; FAISAL, A.; SIAM, S. C.; NAHEEN, I. T.; KHAN, R. Deep learning-based glaucoma detection with cropped optic cup and disc and blood vessel segmentation. *IEEE Access*, v. 10, p. 2828–2841, 2022. Citado na página 51.

SIVASWAMY, J.; KRISHNADAS, S. R.; JOSHI, G. D.; JAIN, M.; TABISH, A. U. S. Drishti-gs: Retinal image dataset for optic nerve head(onh) segmentation. In: *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*. [S.l.: s.n.], 2014. p. 53–56. Citado na página 51.

BUDAI, A.; BOCK, R.; MAIER, A.; HORNEGGER, J.; MICHELSON, G. Robust vessel segmentation in fundus images. *International Journal of Biomedical Imaging*, Hindawi Publishing Corporation, v. 2013, p. 154860, Dec 2013. ISSN 1687-4188. Disponível em: <https://doi.org/10.1155/2013/15486>. Citado na página 52.

KIM, U. Machine learn for glaucoma. Harvard Dataverse, 2018. Disponível em: <https://doi.org/10.7910/DVN/1YRRA>. Citado na página 52.

LI, L.; XU, M.; WANG, X.; JIANG, L.; LIU, H. Attention based glaucoma detection: A large-scale database and cnn model. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2019. Citado na página 52.

ZHANG, Z.; YIN, F.; LIU, J.; WONG, W.; TAN, N.; LEE, B.; CHENG, J.; WONG, T. Origa-light : An online retinal fundus image database for glaucoma analysis and research. 2010. Citado na página 52.

ORLANDO, J. I.; FU, H.; BREDA, J. B.; KEER, K. van; BATHULA, D. R.; DIAZ-PINTO, A.; FANG, R.; HENG, P.-A.; KIM, J.; LEE, J.; LEE, J.; LI, X.; LIU, P.; LU, S.; MURUGESAN, B.; NARANJO, V.; PHAYE, S. S. R.; SHANKARANARAYANA, S. M.; SIKKA, A.; SON, J.; HENGEL, A. van den; WANG, S.; WU, J.; WU, Z.; XU, G.; XU, Y.; YIN, P.; LI, F.; ZHANG, X.; XU, Y.; BOGUNOVIĆ, H. REFUGE challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Med Image Anal*, Netherlands, v. 59, p. 101570, out. 2019. Citado na página 52.

BATISTA, F. J. F.; DIAZ-ALEMAN, T.; SIGUT, J.; ALAYON, S.; ARNAY, R.; ANGEL-PEREIRA, D. Rim-one dl: A unified retinal image database for assessing glaucoma using deep learning. *Image Analysis & Stereology*, v. 39, n. 3, p. 161–167, 2020. ISSN 1854-5165. Disponível em: <https://www.ias-iss.org/ojs/IAS/article/view/234>. Citado 2 vezes nas páginas 52 e 53.

CLARO, M.; VOGADO, L. H.; SANTOS, J.; VERAS, R. Utilização de técnicas de data augmentation em imagens: Teoria e prática. In: _____. [S.l.: s.n.], 2020. p. 47–71. ISBN 9786587003115. Citado na página 53.

TAYLOR, L.; NITSCHKE, G. Improving deep learning using generic data augmentation. *CoRR*, abs/1708.06020, 2017. Disponível em: <https://arxiv.org/abs/1708.0602>. Citado 2 vezes nas páginas 53 e 55.

PEREZ, L.; WANG, J. The effectiveness of data augmentation in image classification using deep learning. *CoRR*, abs/1712.04621, 2017. Disponível em: <https://arxiv.org-/abs/1712.0462>. Citado na página 53.

CUBUK, E. D.; ZOPH, B.; MANÉ, D.; VASUDEVAN, V.; LE, Q. V. Autoaugment: Learning augmentation policies from data. *CoRR*, abs/1805.09501, 2018. Disponível em: <https://arxiv.org/abs/1805.0950>. Citado na página 53.

ORLANDO, J. I.; PROKOFYEVA, E.; FRESNO, M. del; BLASCHKO, M. B. Convolutional neural network transfer for automated glaucoma identification. In: ROMERO, E.; LEPORE, N.; BRIEVA, J.; BRIEVA, J.; AND, I. L. (Ed.). *12th International Symposium on Medical Information Processing and Analysis*. SPIE, 2017. v. 10160, p. 101600U. Disponível em: <https://doi.org/10.1117/12.225574>. Citado na página 54.

DENG, J.; DONG, W.; SOCHER, R.; LI, L.-J.; LI, K.; FEI-FEI, L. 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 na página 56.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009. (Springer series in statistics). ISBN 9780387848846. Disponível em: <https://books.google.com.br-/books?id=eBSgoAEACAA>. Citado na página 56.

MILANI, A.; SILVA, F. da; GUEDES, E.; RIOS, R. A deep learning application for psoriasis detection. In: *Anais do XX Encontro Nacional de Inteligência Artificial e Computacional*. Porto Alegre, RS, Brasil: SBC, 2023. p. 315–329. ISSN 2763-9061. Disponível em: <https://sol.sbc.org.br/index.php/eniac/article/view/2571>. Citado na página 56.

LING, C. X.; HUANG, J.; ZHANG, H. Auc: A statistically consistent and more discriminating measure than accuracy. In: *Proceedings of the 18th International Joint Conference on Artificial Intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003. (IJCAI'03), p. 519–524. Citado na página 57.

Apêndices

APÊNDICE A – Questionário Aplicado ao Profissional de Oftalmologia para Avaliação da Interpretabilidade Visual Gerada pelas Técnicas de XAI nos Diferentes Modelos de CNNs

UFSJ

Universidade Federal de São João del-Rei

Avaliação da Interpretabilidade Visual de Técnicas de Inteligência Artificial Explicável (XAI) para Apoio no Diagnóstico do Glaucoma

Esse questionário faz parte do projeto de pesquisa intitulado "Inteligência Artificial Explicável (XAI) aplicada na classificação de imagens de retinografia para apoio no diagnóstico do Glaucoma".

Objetivo: Realizar a avaliação da interpretabilidade visual gerada por diferentes técnicas de explicabilidade aplicadas em diferentes modelos de redes neurais convolucionais (CNNs) para apoio no diagnóstico do glaucoma.

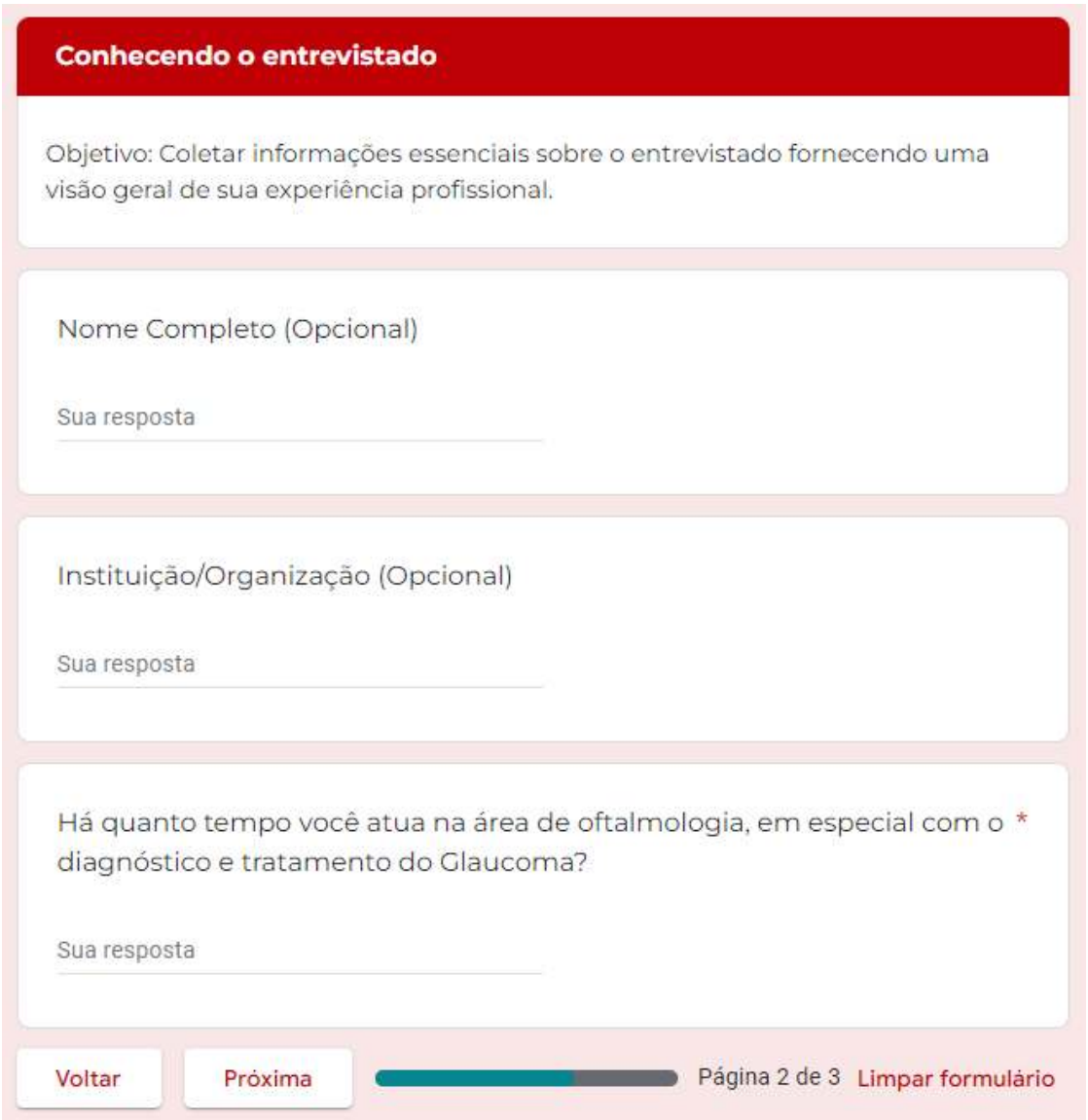
Público alvo: Profissionais de Oftalmologia.

Discente: Cleverson Marques Vieira
Orientador: Diego Roberto Colombo Dias

Programa de Pós Graduação em Ciência da Computação (PPGCC)
Universidade Federal de São João del-Rei (UFSJ)

Faça login no Google para salvar o que você já preencheu. Saiba mais

Próxima Página 1 de 3 Limpar formulário


Conhecendo o entrevistado
Objetivo: Coletar informações essenciais sobre o entrevistado fornecendo uma visão geral de sua experiência profissional.
Nome Completo (Opcional)
Sua resposta
Instituição/Organização (Opcional)
Sua resposta
Há quanto tempo você atua na área de oftalmologia, em especial com o diagnóstico e tratamento do Glaucoma? *
Sua resposta
Voltar
Próxima
Página 2 de 3
Limpar formulário

Avaliando a Interpretabiliade Visual gerada pelas Técnicas de Inteligência Artificial Explicável (XAI)

Objetivo:

Avaliação crítica da interpretabilidade visual produzida por técnicas de Inteligência Artificial Explicável (XAI) aplicadas a modelos baseados em redes neurais convolucionais (CNNs) para apoio ao diagnóstico do Glaucoma.

Buscamos compreender a eficácia dessas técnicas em traduzir as decisões complexas dos modelos em interpretações visuais de fácil compreensão. Ao explorar a qualidade e a utilidade das representações visuais geradas, pretendemos identificar os pontos fortes e desafios associados a essas técnicas.

Suas percepções e observações serão fundamentais para aprimorar a interpretabilidade visual em XAI, contribuindo para um entendimento mais profundo das decisões tomadas por modelos de CNNs em diferentes contextos.

As figuras abaixo representam imagens de fundo de retina rotuladas como "**Glaucoma Positivo**".

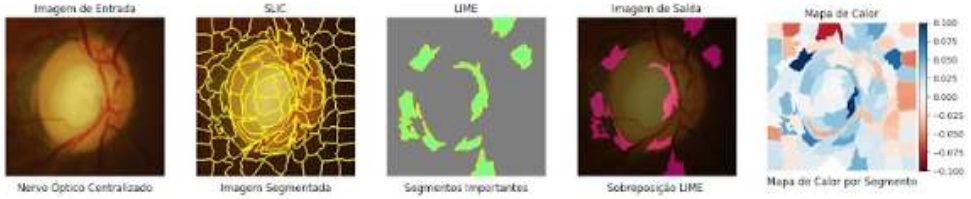
Diferentes técnicas de interpretabilidade visual foram aplicadas em diferentes modelos de redes neurais convolucionais.

Verificando essas imagens e analisando a interpretabilidade gerada, por favor, responda as perguntas a seguir.

PERGUNTA 1 - O quanto você concorda com a afirmação: "As imagens abaixo me facilitam a diagnosticar o Glaucoma". *

Obs.: Use a escala de (1) a (5), onde (1) indica discordo totalmente, (2) discordo, (3) indiferente (ou neutro), (4) concordo e (5) concordo totalmente.

Para ver a imagem ampliada, Clique aqui.



O SLIC representa as regiões segmentadas da Imagem de Entrada.

O LIME representa as regiões importantes que foram significativas para o diagnóstico.

A Imagem de Saída representa a sobreposição do LIME na Imagem de Entrada, destacando os segmentos da imagem mais significativos para o diagnóstico.

As cores no Mapa de Calor representam a intensidade / força de cada segmento da Imagem de Entrada para o diagóstico do Glaucoma.

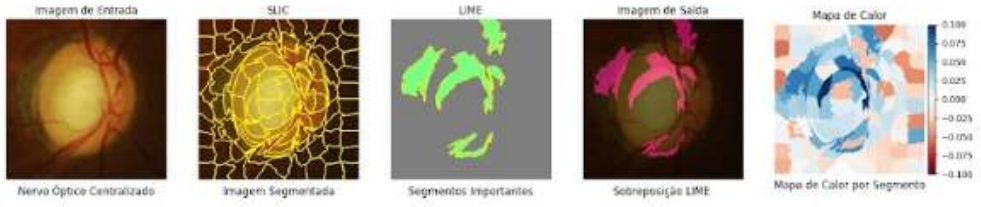
A cor AZUL no Mapa de Calor representa maior intensidade, enquanto a cor VERMELHA representa menor intensidade.

	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

PERGUNTA 2 - O quanto você concorda com a afirmação: "As imagens abaixo me facilitam a diagnosticar o Glaucoma". *

Obs.: Use a escala de (1) a (5), onde (1) indica discordo totalmente, (2) discordo, (3) indiferente (ou neutro), (4) concordo e (5) concordo totalmente.

Para ver a imagem ampliada, Clique aqui.



O SLIC representa as regiões segmentadas da Imagem de Entrada.

O LIME representa as regiões importantes que foram significativas para o diagnóstico.

A Imagem de Saída representa a sobreposição do LIME na Imagem de Entrada, destacando os segmentos da imagem mais significativos para o diagnóstico.

As cores no Mapa de Calor representam a intensidade / força de cada segmento da Imagem de Entrada para o diagóstico do Glaucoma.

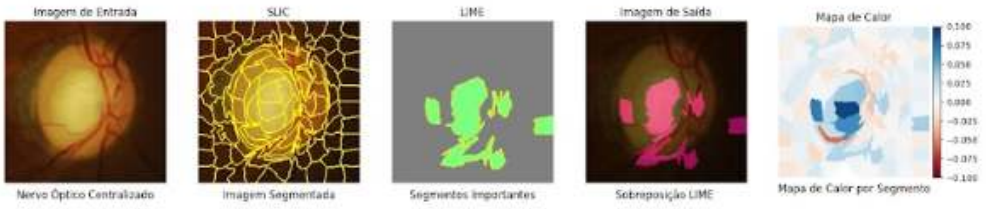
A cor AZUL no Mapa de Calor representa maior intensidade, enquanto a cor VERMELHA representa menor intensidade.

	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

PERGUNTA 3 - O quanto você concorda com a afirmação: "As imagens abaixo me facilitam a diagnosticar o Glaucoma". *

Obs.: Use a escala de (1) a (5), onde (1) indica discordo totalmente, (2) discordo, (3) indiferente (ou neutro), (4) concordo e (5) concordo totalmente.

Para ver a imagem ampliada, Clique aqui.



O SLIC representa as regiões segmentadas da imagem de Entrada.

O LIME representa as regiões importantes que foram significativas para o diagnóstico.

A Imagem de Saída representa a sobreposição do LIME na Imagem de Entrada, destacando os segmentos da imagem mais significativos para o diagnóstico.

As cores no Mapa de Calor representam a intensidade / força de cada segmento da Imagem de Entrada para o diagóstico do Glaucoma.

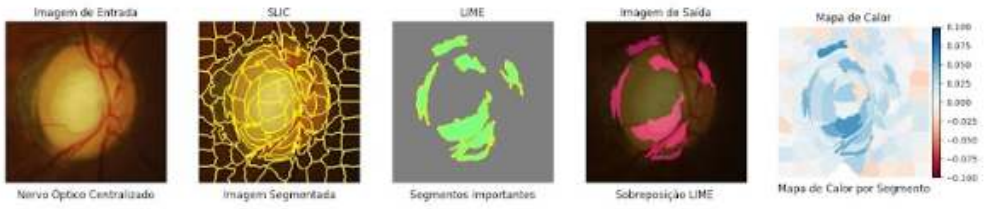
A cor AZUL no Mapa de Calor representa maior intensidade, enquanto a cor VERMELHA representa menor intensidade.

	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

PERGUNTA 4 - O quanto você concorda com a afirmação: "As imagens abaixo me facilitam a diagnosticar o Glaucoma". *

Obs.: Use a escala de (1) a (5), onde (1) indica discordo totalmente, (2) discordo, (3) indiferente (ou neutro), (4) concordo e (5) concordo totalmente.

Para ver a imagem ampliada, Clique aqui.



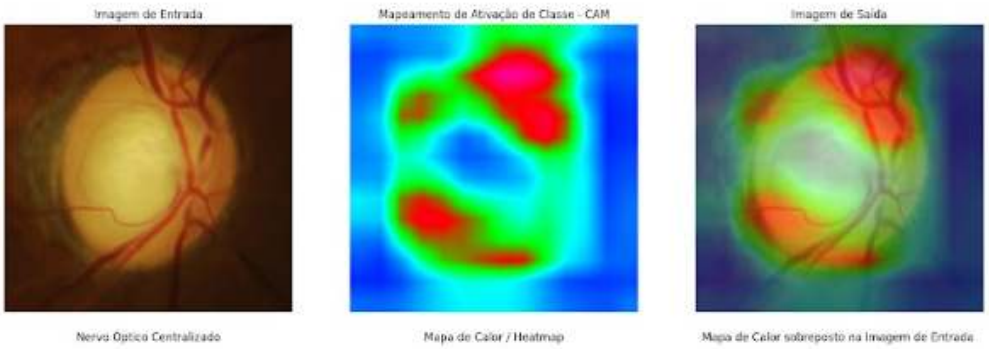
O SLIC representa as regiões segmentadas da Imagem de Entrada.
O LIME representa as regiões importantes que foram significativas para o diagnóstico.
A Imagem de Saída representa a sobreposição do LIME na Imagem de Entrada, destacando os segmentos da imagem mais significativos para o diagnóstico.
As cores no Mapa de Calor representam a intensidade / força de cada segmento da Imagem de Entrada para o diagóstico do Glaucoma.
A cor AZUL no Mapa de Calor representa maior intensidade, enquanto a cor VERMELHA representa menor intensidade.

	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

PERGUNTA 5 - O quanto você concorda com a afirmação: "As imagens abaixo me facilitam a diagnosticar o Glaucoma". *

Obs.: Use a escala de (1) a (5), onde (1) indica discordo totalmente, (2) discordo, (3) indiferente (ou neutro), (4) concordo e (5) concordo totalmente.

Para ver a imagem ampliada, Clique aqui.



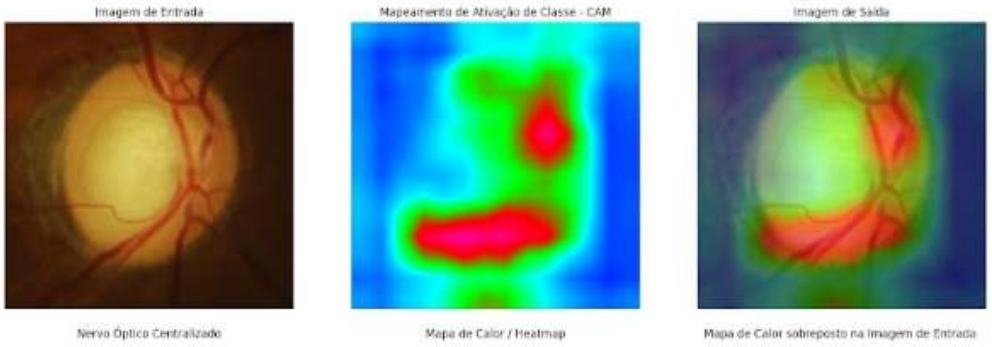
As cores no Mapa de Calor / Heatmap representam a intensidade de cada região da Imagem de Entrada para o diagóstico do Glaucoma.
O VERMELHO é mais intenso / forte que o VERDE e o VERDE é mais intenso / forte que o AZUL

	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

PERGUNTA 6 - O quanto você concorda com a afirmação: "As imagens abaixo me facilitam a diagnosticar o Glaucoma". *

Obs.: Use a escala de (1) a (5), onde (1) indica discordo totalmente, (2) discordo, (3) indiferente (ou neutro), (4) concordo e (5) concordo totalmente.

Para ver a imagem ampliada, Clique aqui.



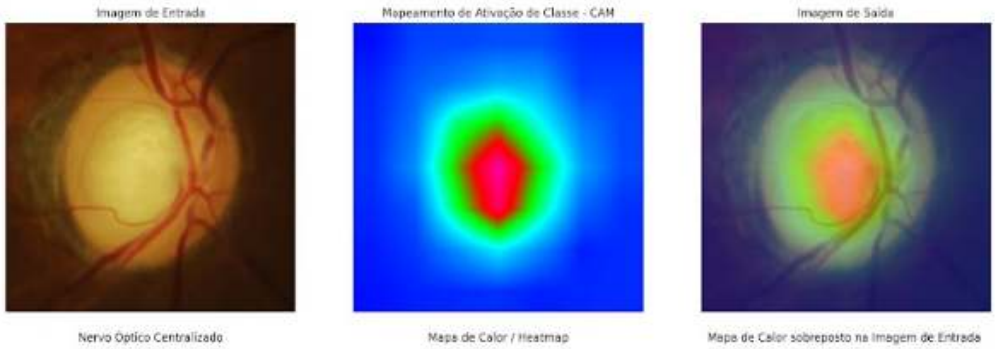
As cores no Mapa de Calor / Heatmap representam a intensidade de cada região da Imagem de Entrada para o diagóstico do Glaucoma.
O VERMELHO é mais intenso / forte que o VERDE e o VERDE é mais intenso / forte que o AZUL

	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

PERGUNTA 7 - O quanto você concorda com a afirmação: "As imagens abaixo me facilitam a diagnosticar o Glaucoma". *

Obs.: Use a escala de (1) a (5), onde (1) indica discordo totalmente, (2) discordo, (3) indiferente (ou neutro), (4) concordo e (5) concordo totalmente.

Para ver a imagem ampliada, Clique aqui.



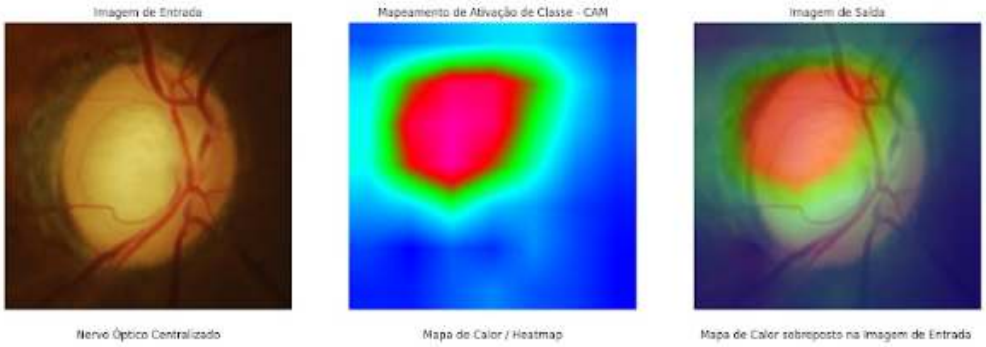
As cores no Mapa de Calor / Heatmap representam a intensidade de cada região da Imagem de Entrada para o diagóstico do Glaucoma.
O VERMELHO é mais intenso / forte que o VERDE e o VERDE é mais intenso / forte que o AZUL.

	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

PERGUNTA 8 - O quanto você concorda com a afirmação: "As imagens abaixo me facilitam a diagnosticar o Glaucoma". *

Obs.: Use a escala de (1) a (5), onde (1) indica discordo totalmente, (2) discordo, (3) indiferente (ou neutro), (4) concordo e (5) concordo totalmente.

Para ver a imagem ampliada, Clique aqui.



As cores no Mapa de Calor / Heatmap representam a intensidade de cada região da Imagem de Entrada para o diagóstico do Glaucoma.
O VERMELHO é mais intenso / forte que o VERDE e o VERDE é mais intenso / forte que o AZUL.

	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

PERGUNTA 9 - O quanto você concorda com a afirmação: "As imagens abaixo me facilitam a diagnosticar o Glaucoma". *

Obs.: Use a escala de (1) a (5), onde (1) indica discordo totalmente, (2) discordo, (3) indiferente (ou neutro), (4) concordo e (5) concordo totalmente.

Para ver a imagem ampliada, Clique aqui.

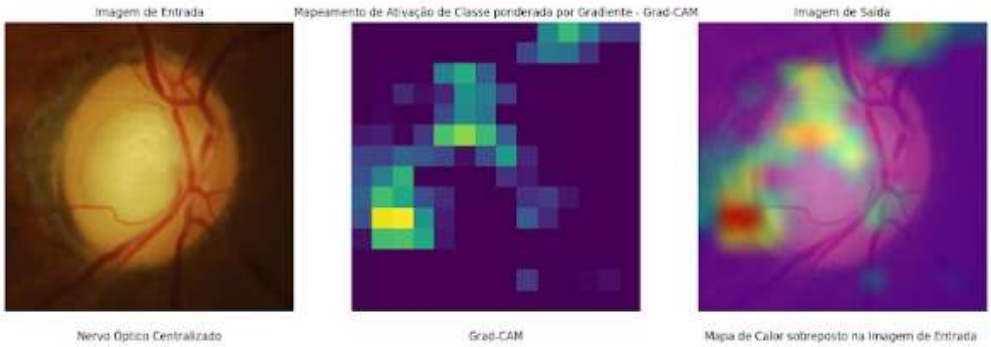
A cor AMARELA no Grad-CAM representa a região da Imagem de Entrada que mais contribuiu para o diagóstico do Glaucoma.
A Imagem de Saída representa a sobreposição do Grad-CAM na Imagem de Entrada, destacando a região mais significativa para o diagnóstico.

	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

PERGUNTA 10 - O quanto você concorda com a afirmação: "As imagens abaixo me facilitam a diagnosticar o Glaucoma". *

Obs.: Use a escala de (1) a (5), onde (1) indica discordo totalmente, (2) discordo, (3) indiferente (ou neutro), (4) concordo e (5) concordo totalmente.

Para ver a imagem ampliada, Clique aqui.



A cor AMARELA no Grad-CAM representa a região da Imagem de Entrada que mais contribuiu para o diagóstico de Glaucoma.
A Imagem de Saída representa a sobreposição do Grad-CAM na Imagem de Entrada, destacando a região mais significativa para o diagnóstico.

	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

PERGUNTA 11 - O quanto você concorda com a afirmação: "As imagens abaixo me facilitam a diagnosticar o Glaucoma". *

Obs.: Use a escala de (1) a (5), onde (1) indica discordo totalmente, (2) discordo, (3) indiferente (ou neutro), (4) concordo e (5) concordo totalmente.

Para ver a imagem ampliada, Clique aqui.

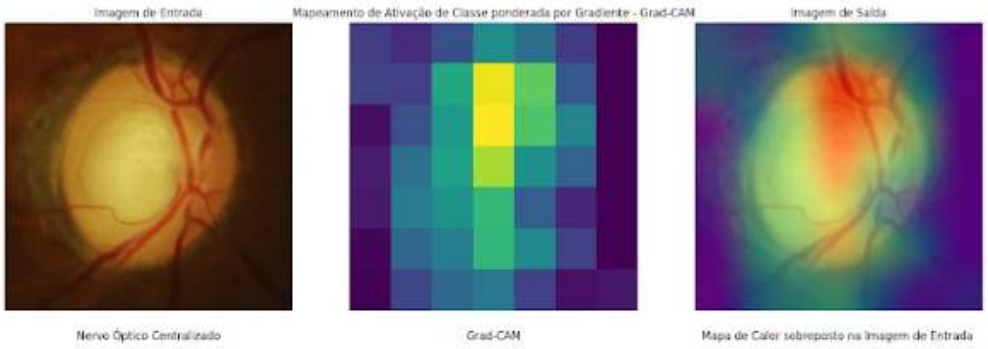
A cor AMARELA no Grad-CAM representa a região da Imagem de Entrada que mais contribuiu para o diagóstico do Glaucoma.
A Imagem de Saída representa a sobreposição do Grad-CAM na Imagem de Entrada, destacando a região mais significativa para o diagnóstico.

	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

PERGUNTA 12 - O quanto você concorda com a afirmação: "As imagens abaixo me facilitam a diagnosticar o Glaucoma". *

Obs.: Use a escala de (1) a (5), onde (1) indica discordo totalmente, (2) discordo, (3) indiferente (ou neutro), (4) concordo e (5) concordo totalmente.

Para ver a imagem ampliada, Clique aqui.



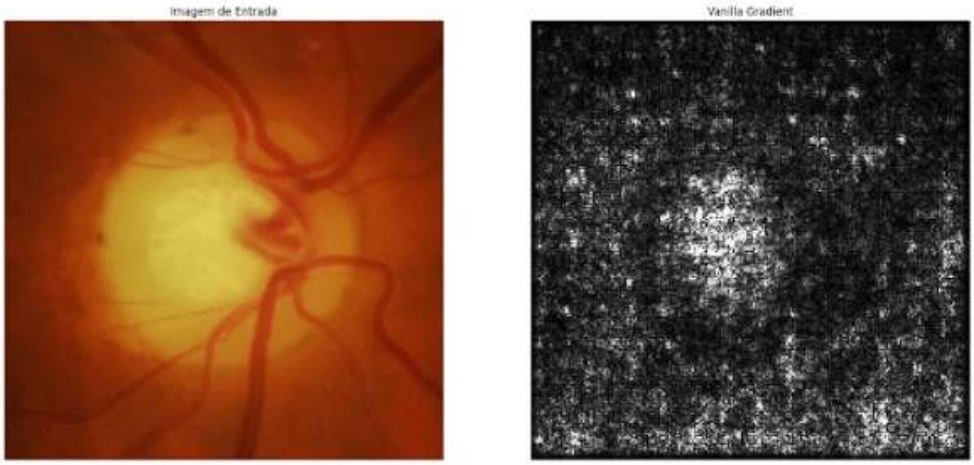
A cor AMARELA no Grad-CAM representa a região da Imagem de Entrada que mais contribuiu para o diagóstico do Glaucoma.
A Imagem de Saída representa a sobreposição do Grad-CAM na Imagem de Entrada, destacando a região mais significativa para o diagnóstico.

	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

PERGUNTA 13 - O quanto você concorda com a afirmação: "As imagens abaixo me facilitam a diagnosticar o Glaucoma". *

Obs.: Use a escala de (1) a (5), onde (1) indica discordo totalmente, (2) discordo, (3) indiferente (ou neutro), (4) concordo e (5) concordo totalmente.

Para ver a imagem ampliada, Clique aqui.



	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

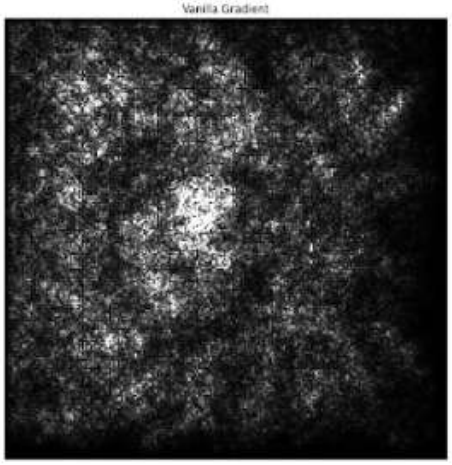
PERGUNTA 14 - O quanto você concorda com a afirmação: "As imagens abaixo me facilitam a diagnosticar o Glaucoma". *

Obs.: Use a escala de (1) a (5), onde (1) indica discordo totalmente, (2) discordo, (3) indiferente (ou neutro), (4) concordo e (5) concordo totalmente.

Para ver a imagem ampliada, Clique aqui.

Imagem de Entrada

Vanilla Gradient



Nervo Óptico Centralizado

Vanilla Gradient

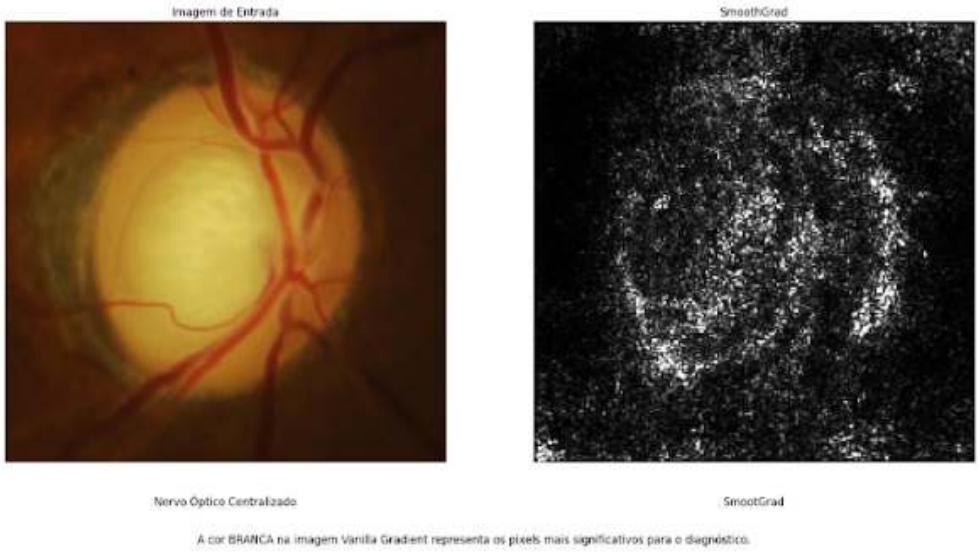
A cor BRANCA na imagem Vanilla Gradient representa os pixels mais significativos para o diagnóstico.

	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

PERGUNTA 15 - O quanto você concorda com a afirmação: "As imagens abaixo me facilitam a diagnosticar o Glaucoma". *

Obs.: Use a escala de (1) a (5), onde (1) indica discordo totalmente, (2) discordo, (3) indiferente (ou neutro), (4) concordo e (5) concordo totalmente.

Para ver a imagem ampliada, Clique aqui.

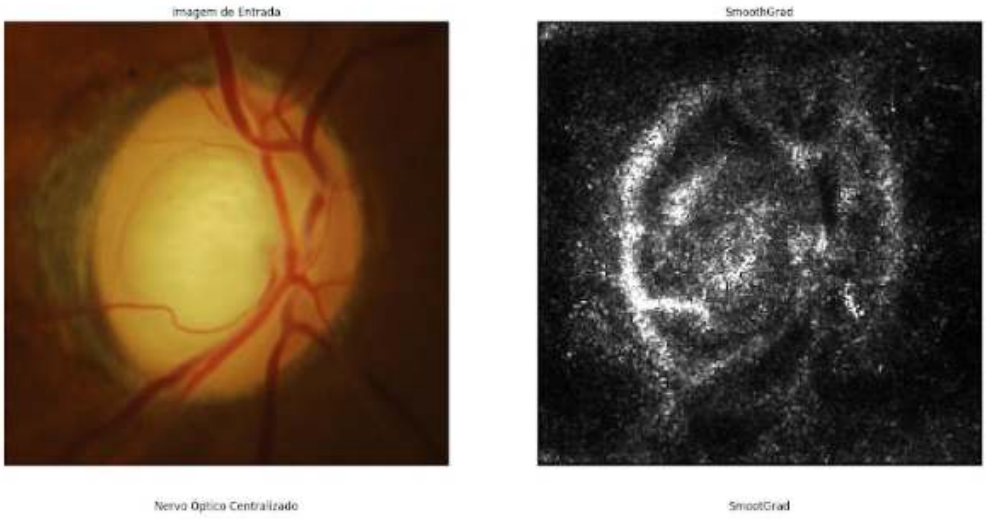


	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

PERGUNTA 16 - O quanto você concorda com a afirmação: "As imagens abaixo me facilitam a diagnosticar o Glaucoma". *

Obs.: Use a escala de (1) a (5), onde (1) indica discordo totalmente, (2) discordo, (3) indiferente (ou neutro), (4) concordo e (5) concordo totalmente.

Para ver a imagem ampliada, Clique aqui.



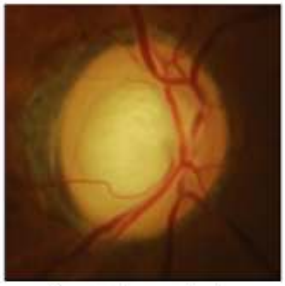
	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

PERGUNTA 17 - O quanto você concorda com a afirmação: "As imagens abaixo me facilitam a diagnosticar o Glaucoma". *

Obs.: Use a escala de (1) a (5), onde (1) indica discordo totalmente, (2) discordo, (3) indiferente (ou neutro), (4) concordo e (5) concordo totalmente.

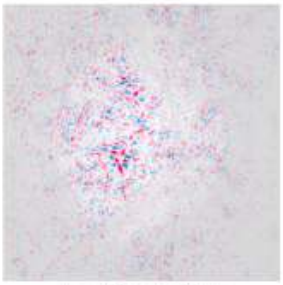
Para ver a imagem ampliada, Clique aqui.

Imagem de Entrada



Nervo óptico centralizado

Imagem de Saída



Importância dos Pixels

A cor ROSA na Imagem de Saída representa os pixels positivos que mais contribuíram para o diagnóstico.
A cor AZUL na Imagem de Saída representa os pixels negativos com uma menor contribuição.

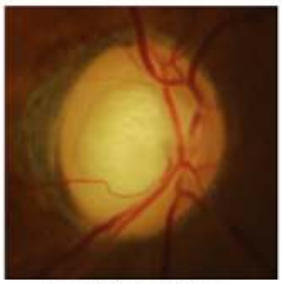
	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

PERGUNTA 18 - O quanto você concorda com a afirmação: "As imagens abaixo me facilitam a diagnosticar o Glaucoma". *

Obs.: Use a escala de (1) a (5), onde (1) indica discordo totalmente, (2) discordo, (3) indiferente (ou neutro), (4) concordo e (5) concordo totalmente.

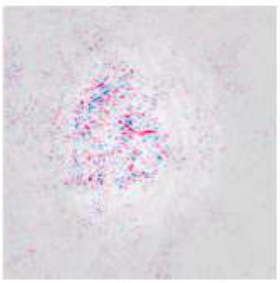
Para ver a imagem ampliada, Clique aqui.

Imagem de Entrada



Nervo óptico centralizado

Imagem de Saída



Importância dos Pixels

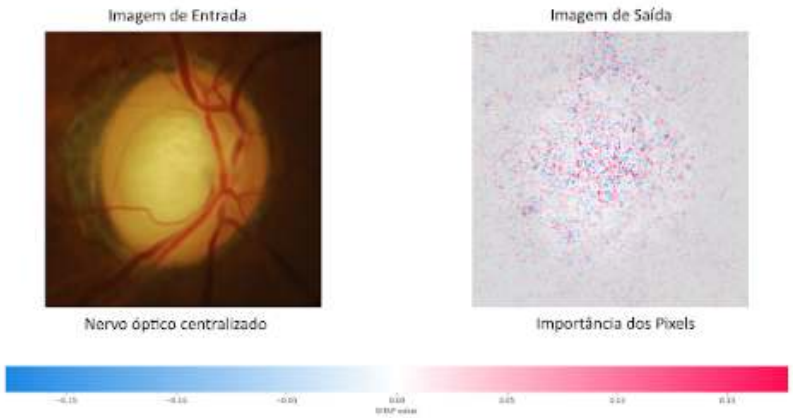
A cor ROSA na Imagem de Saída representa os pixels positivos que mais contribuíram para o diagnóstico.
A cor AZUL na Imagem de Saída representa os pixels negativos com uma menor contribuição.

	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

PERGUNTA 19 - O quanto você concorda com a afirmação: "As imagens abaixo me facilitam a diagnosticar o Glaucoma". *

Obs.: Use a escala de (1) a (5), onde (1) indica discordo totalmente, (2) discordo, (3) indiferente (ou neutro), (4) concordo e (5) concordo totalmente.

Para ver a imagem ampliada, Clique aqui.

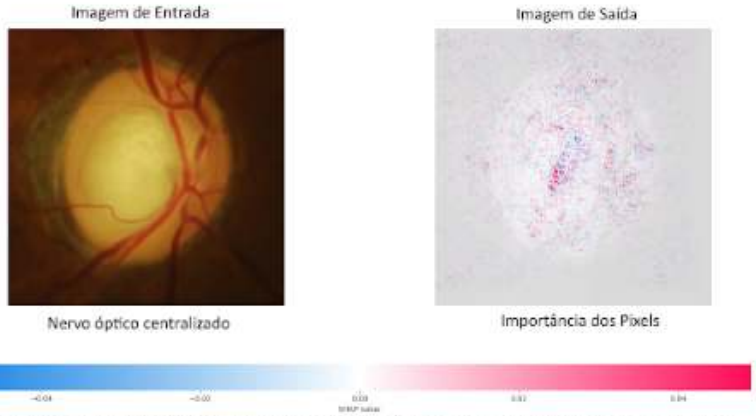


	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

PERGUNTA 20 - O quanto você concorda com a afirmação: "As imagens abaixo me facilitam a diagnosticar o Glaucoma". *

Obs.: Use a escala de (1) a (5), onde (1) indica discordo totalmente, (2) discordo, (3) indiferente (ou neutro), (4) concordo e (5) concordo totalmente.

Para ver a imagem ampliada, Clique aqui.



	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

PERGUNTA 21 - O quanto você concorda com a afirmação: "As imagens abaixo me facilitam a diagnosticar o Glaucoma". *

Obs.: Use a escala de (1) a (5), onde (1) indica discordo totalmente, (2) discordo, (3) indiferente (ou neutro), (4) concordo e (5) concordo totalmente.

Para ver a imagem ampliada, Clique aqui.

Nervo óptico centralizado

Importância dos Pixels

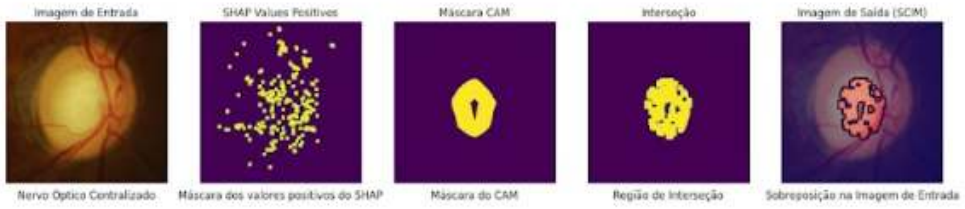
A cor ROSA na Imagem de Saída representa os pixels positivos que mais contribuíram para o dignóstico.
A cor AZUL na Imagem de Saída representa os pixels negativos com uma menor contribuição.

	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

PERGUNTA 22 - O quanto você concorda com a afirmação: "As imagens abaixo me facilitam a diagnosticar o Glaucoma". *

Obs.: Use a escala de (1) a (5), onde (1) indica discordo totalmente, (2) discordo, (3) indiferente (ou neutro), (4) concordo e (5) concordo totalmente.

Para ver a imagem ampliada, Clique aqui.



A área destacada na Imagem de Saída (SCIM) representa a região de confluência (região mais significativa para o diagnóstico) entre os valores positivos de SHAP e a região do CAM.

	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente

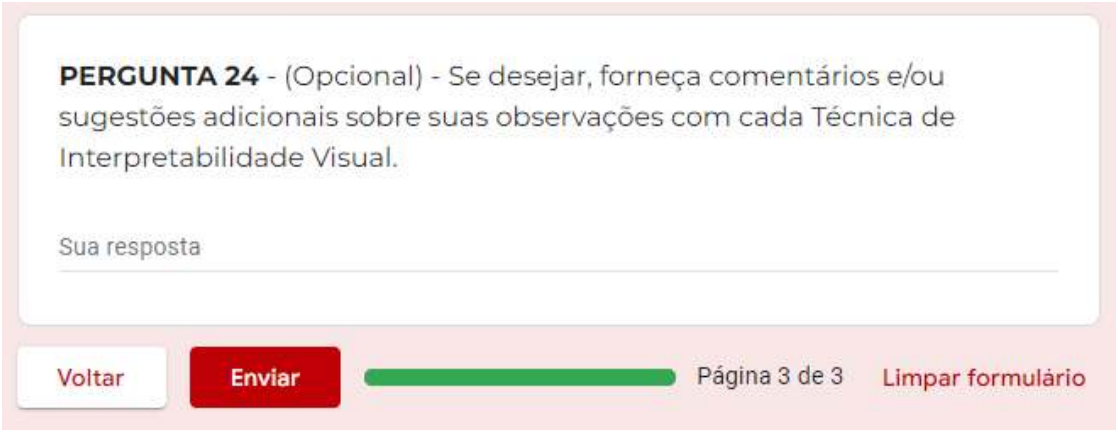
PERGUNTA 23 - O quanto você concorda com a afirmação: "As imagens abaixo me facilitam a diagnosticar o Glaucoma". *

Obs.: Use a escala de (1) a (5), onde (1) indica discordo totalmente, (2) discordo, (3) indiferente (ou neutro), (4) concordo e (5) concordo totalmente.

Para ver a imagem ampliada, Clique aqui.

A área destacada na Imagem de Saída (SCIM) representa a região de confluência (região mais significativa para o diagnóstico) entre os valores positivos de SHAP e a região do CAM.

	1	2	3	4	5	
Discordo totalmente	○	○	○	○	○	Concordo totalmente


PERGUNTA 24 - (Opcional) - Se desejar, forneça comentários e/ou sugestões adicionais sobre suas observações com cada Técnica de Interpretabilidade Visual.
Sua resposta
Voltar
Enviar
Página 3 de 3
Limpar formulário