UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

WENDSON MAGALHÃES DA SILVA

ANÁLISE DE TÉCNICAS DE EXPLICABILIDADE EM REDES NEURAIS CONVOLUCIONAIS PARA DIAGNÓSTICO DE GLAUCOMA, RETINOPATIA DIABÉTICA E CATARATA

CAMPINA GRANDE - PB

2024

WENDSON MAGALHÃES DA SILVA

ANÁLISE DE TÉCNICAS DE EXPLICABILIDADE EM REDES NEURAIS CONVOLUCIONAIS PARA DIAGNÓSTICO DE GLAUCOMA, RETINOPATIA DIABÉTICA E CATARATA

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Herman Martins Gomes

CAMPINA GRANDE - PB

2024

WENDSON MAGALHÃES DA SILVA

ANÁLISE DE TÉCNICAS DE EXPLICABILIDADE EM REDES NEURAIS CONVOLUCIONAIS PARA DIAGNÓSTICO DE GLAUCOMA, RETINOPATIA DIABÉTICA E CATARATA

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Herman Martins Gomes

Orientador – UASC/CEEI/UFCG

Leandro Balby Marinho

Examinador – UASC/CEEI/UFCG

Francisco Vilar Brasileiro

Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 15 de Maio de 2024.

CAMPINA GRANDE - PB

RESUMO

Doenças oftalmológicas, como catarata, glaucoma e retinopatia diabética, representam um desafio significativo para a saúde pública, com potencial de causar perda de visão. No entanto, a maioria desses casos poderia ser evitada ou tratada se diagnosticada precocemente. Neste contexto, a imagem de fundo de olho surge como uma ferramenta de diagnóstico eficaz, rápida e não invasiva.

A interpretação manual de imagens oftalmológicas é repetitiva e sujeita a erros. Assim, sistemas computacionais podem ser utilizados para auxiliar os profissionais na triagem automatizada, reduzindo tempo, erros e esforço na análise das doenças. Os sistemas de aprendizado profundo provaram ser eficazes nesse contexto, entretanto, sua falta de transparência tem sido um desafio para a adoção clínica, o que destaca a importância da explicabilidade nos modelos de aprendizado de máquina.

Este estudo contribui para o avanço da compreensão e interpretação de modelos de aprendizado profundo na área da saúde ocular, visando melhorar o diagnóstico e tratamento de condições oftalmológicas. Ele compara as técnicas LIME e Grad-CAM aplicadas a diferentes arquiteturas de redes neurais convolucionais (CNNs) treinadas para classificar condições oftalmológicas a partir de imagens de fundo de olho. Os resultados indicam que o modelo VGG16 se destaca, alcançando uma acurácia de 93,17% no treinamento e 87,16% na validação. Além disso, as técnicas de explicabilidade, embora distintas em abordagem, identificaram quase as mesmas regiões de interesse nas imagens oftalmológicas. Ainda assim, apesar de haver limitações, como a aleatoriedade do LIME e a necessidade de ajustes no Grad-CAM, o LIME destacou áreas críticas de forma mais sutil, enquanto o Grad-CAM forneceu representações visuais mais diretas e intuitivas.

ANALYSIS OF EXPLAINABILITY TECHNIQUES IN CONVOLUTIONAL NEURAL NETWORKS FOR THE DIAGNOSIS OF GLAUCOMA, DIABETIC RETINOPATHY, AND CATARACTS

ABSTRACT

Ophthalmic diseases, such as cataracts, glaucoma, and diabetic retinopathy, pose a significant challenge to public health, with the potential to cause vision loss. However, the majority of these cases could be prevented or treated if diagnosed early. In this context, fundus imaging emerges as an effective, fast and non-invasive diagnosis tool.

Manual interpretation of ophthalmic images is repetitive and prone to error. Thus, computational systems can be used to assist professionals in automated screening, reducing time, errors, and effort in disease analysis. Deep learning systems have proven effective in this context, however, their lack of transparency has been a challenge for clinical adoption, highlighting the importance of explainability in machine learning models.

This study contributes to advancing the understanding and interpretation of deep learning models in the field of ocular health, aiming to improve the diagnosis and treatment of ophthalmic conditions. It compares the LIME and Grad-CAM techniques applied to different architectures of convolutional neural networks (CNNs) trained to classify ophthalmic conditions from fundus images. The results indicate that the VGG16 model stands out, achieving an accuracy of 93.17% in training and 87.16% in validation. Additionally, explainability techniques, though distinct in approach, identified nearly the same regions of interest in ophthalmic images. Nevertheless, despite limitations such as the randomness of LIME and the need for adjustments in Grad-CAM, LIME highlighted critical areas more subtly, while Grad-CAM provided more direct and intuitive visual representations.

Análise de Técnicas de Explicabilidade em Redes Neurais Convolucionais para Diagnóstico de Glaucoma, Retinopatia Diabética e Catarata

Wendson Magalhães da Silva
Universidade Federal de Campina Grande
Campina Grande, Paraíba
E-mail: wendson.silva@ccc.ufcg.edu.br

Herman Martins Gomes
Universidade Federal de Campina Grande
Campina Grande, Paraíba
E-mail: hmg@dsc.ufcg.edu.br

ABSTRACT

Ophthalmic diseases, such as cataracts, glaucoma, and diabetic retinopathy, pose a significant challenge to public health, with the potential to cause vision loss. However, the majority of these cases could be prevented or treated if diagnosed early. In this context, fundus imaging emerges as an effective, fast and non-invasive diagnosis tool.

Manual interpretation of ophthalmic images is repetitive and prone to error. Thus, computational systems can be used to assist professionals in automated screening, reducing time, errors, and effort in disease analysis. Deep learning systems have proven effective in this context, however, their lack of transparency has been a challenge for clinical adoption, highlighting the importance of explainability in machine learning models.

This study contributes to advancing the understanding and interpretation of deep learning models in the field of ocular health, aiming to improve the diagnosis and treatment of ophthalmic conditions. It compares the LIME and Grad-CAM techniques applied to different architectures of convolutional neural networks (CNNs) trained to classify ophthalmic conditions from fundus images. The results indicate that the VGG16 model stands out, achieving an accuracy of 93.17% in training and 87.16% in validation. Additionally, explainability techniques, though distinct in approach, identified nearly the same regions of interest in ophthalmic images. Nevertheless, despite limitations such as the randomness of LIME and the need for adjustments in Grad-CAM, LIME highlighted critical areas more subtly, while Grad-CAM provided more direct and intuitive visual representations.

Keywords

Artificial intelligence, Convolutional neural networks, Explainability, Ocular health, Image diagnosis.

RESUMO

Doenças oftalmológicas, como catarata, glaucoma e retinopatia diabética, representam um desafio significativo para a saúde pública, com potencial de causar perda de visão. No entanto, a maioria desses casos poderia ser evitado ou tratado se diagnosticado precocemente. Neste contexto, a imagem de fundo de olho surge como uma ferramenta de diagnóstico eficaz, rápida e não invasiva.

A interpretação manual de imagens oftalmológicas é repetitiva e sujeita a erros. Assim, sistemas computacionais podem ser utilizados para auxiliar os profissionais na triagem automatizada, reduzindo tempo, erros e esforço na análise das doenças. Os sistemas de aprendizado profundo provaram ser eficazes nesse contexto, entretanto, sua falta de transparência tem sido um desafio para a adoção clínica, o que destaca a importância da explicabilidade nos modelos de aprendizado de máquina.

Este estudo contribui para o avanço da compreensão e interpretação de modelos de aprendizado profundo na área da saúde ocular, visando melhorar o diagnóstico e tratamento de condições oftalmológicas. Ele compara as técnicas LIME e Grad-CAM aplicadas a diferentes arquiteturas de redes neurais convolucionais (CNNs) treinadas para classificar condições oftalmológicas a partir de imagens de fundo de olho. Os resultados indicam que o modelo VGG16 se destaca, alcançando uma acurácia de 93,17% no treinamento e 87,16% na validação. Além disso, as técnicas de explicabilidade, embora distintas em abordagem, identificaram quase as mesmas regiões de interesse nas imagens oftalmológicas. Ainda assim, apesar de haver limitações, como a aleatoriedade do LIME e a necessidade de ajustes no Grad-CAM, o LIME destacou áreas críticas de forma mais sutil, enquanto o Grad-CAM forneceu representações visuais mais diretas e intuitivas.

Palavras-chave

Inteligência artificial, Redes neurais convolucionais, Explicabilidade, Saúde ocular, Diagnóstico por imagem.

1. INTRODUÇÃO

Doenças oculares como catarata, glaucoma, retinopatia diabética e erros de refração são comuns no Brasil e requerem atenção especial, assim como podem causar complicações graves, como perda parcial ou total da visão, se não forem diagnosticadas e tratadas precocemente. Segundo os dados da Organização Mundial de Saúde (OMS), cerca de 285 milhões de pessoas em todo o mundo têm deficiências visuais, e entre 60 e 80 por cento desses

casos poderiam ser evitados ou tratados se o diagnóstico fosse feito precocemente[1].

Devido à necessidade de análise minuciosa das estruturas oculares internas, o processo de diagnóstico dessas doenças costuma ser demorado e meticuloso. Uma alternativa eficaz é o exame de imagem de fundo de olho, que não é invasivo, de baixo custo e capaz de detectar diversas anormalidades oculares. Contudo, a interpretação manual dessas imagens demanda tempo e está sujeita a erros, além de ser uma tarefa repetitiva e monótona. [17][29]. Assim, o uso de sistemas computacionais para auxiliar os profissionais na triagem automatizada pode reduzir consideravelmente o tempo, os erros e o esforço exigidos para a análise dessas doenças.

Os sistemas de aprendizado profundo, uma subárea da aprendizagem de máquina que simula o processamento neural humano, demonstram eficácia em diversas tarefas de diagnóstico médico, incluindo doenças oculares [18]. Entretanto, a falta de transparência desses modelos tem sido um obstáculo para sua adoção clínica [45], ressaltando a importância da explicabilidade nos modelos de aprendizado de máquina.

A confiança nas previsões desses modelos muitas das vezes é comprometida pela sua complexidade e opacidade. Portanto, torna-se essencial o uso de técnicas de explicabilidade para compreender e interpretar as decisões desses modelos. Entre técnicas de explicabilidade existentes, LIME (*Local Interpretable Model-agnostic Explanations*) e Grad-CAM (*Gradient-weighted Class Activation Mapping*) se destacam ao buscar tornar o funcionamento interno desses modelos profundos compreensível para os humanos. Essa transparência é fundamental em contextos de saúde, nos quais as decisões humanas desempenham papel essencial no diagnóstico e tratamento de doenças.

Dito isso, o objetivo principal deste estudo é realizar uma análise comparativa das técnicas de explicabilidade LIME e Grad-CAM aplicadas a três modelos de redes neurais treinados (VGG16, ResNet50 e InceptionV3) no diagnóstico de catarata, glaucoma e retinopatia diabética. Os dados recebem uma análise qualitativa, com foco na interpretação das imagens de saída das técnicas LIME e Grad-CAM, em vistas a comparar as áreas destacadas nessas imagens com os padrões e características descritos na literatura científica para o diagnóstico das doenças oftalmológicas em questão.

2. FUNDAMENTAÇÃO TEÓRICA

Nas próximas seções, serão apresentados conceitos gerais sobre a estrutura do olho, doenças oculares e avanços científicos e tecnológicos relevantes. Na Seção 2.1, será discutida a estrutura do olho humano e seu funcionamento. A Seção 2.2 abordará algumas das doenças oculares mais comuns, incluindo glaucoma, retinopatia diabética e catarata. Em seguida, na Seção 2.3, será explorada a técnica da fundoscopia para examinar o fundo do olho. Posteriormente, na Seção 2.4, serão apresentados avanços significativos no campo científico e tecnológico, com foco em aprendizado de máquina e redes neurais. A Subseção 2.4.1 destacará a importância das redes neurais convolucionais na análise de imagens médicas. Já a Subseção 2.4.2 discorrerá sobre o aprendizado por transferência, incluindo algoritmos como ResNet50, VGG 16 e Inception V3, e sua aplicação no diagnóstico de doenças oculares.

Por fim, a Subseção 2.4.3 abordará o tema da interpretabilidade e explicabilidade, apresentando as técnicas LIME e Grad-Cam, que ajudam a compreender a tomada de decisão das redes neurais em diagnóstico por imagem.

2.1 Estrutura do Olho

O olho humano é um sistema sensorial complexo que desempenha um papel vital junto ao cérebro na captação e interpretação de imagens, assim, permitindo nossa interação natural com o mundo ao nosso redor.

O globo ocular (Figura 1) adota uma forma quase esférica, com um diâmetro de cerca de 25mm, composto por três camadas distintas: a camada externa, que compreende a córnea e a esclera; a camada intermediária, que abriga a íris, a coroide e o corpo ciliar; e a camada interna, constituída pela retina, onde as imagens são diretamente focadas [25].

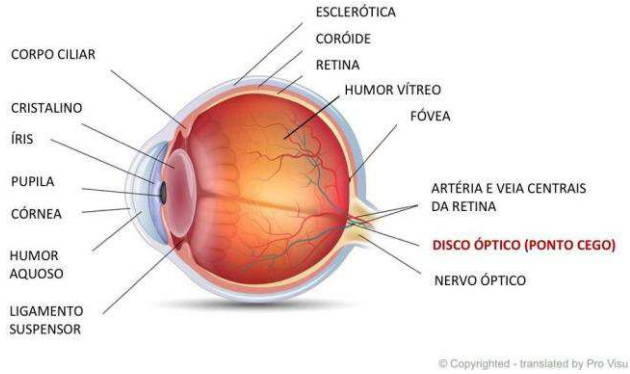

Figura 1: Esquema de um corte longitudinal do olho humano. Fonte: Singh et al [41].

De acordo com Courrol e Preto (2011) [10], a retina desempenha um papel essencial no processo da visão, recebendo, focalizando e transmitindo imagens para o cérebro. Na região da mácula, as imagens são especialmente nítidas, com a maior nitidez concentrada na área central conhecida como fóvea. O nervo óptico é a extensão das células nervosas da retina e tem a responsabilidade de transportar as imagens captadas pela retina para o cérebro, onde a visão será processada. Com isso, o processo de formação de imagens é rápido e complexo, pois, o olho capta a luz e a converte em impulsos nervosos, que são então interpretados pelo cérebro. Mais de cem milhões de células fotorreceptoras na retina convertem os sinais luminosos em impulsos eletroquímicos, que são processados pelo cérebro.

A retina (Figura 2) é a camada mais interna do olho, localizada na sua parede posterior e é fundamental na detecção de anomalias oculares e sistêmicas. Estruturas como o disco óptico, a rede de vasos sanguíneos, a mácula e a fóvea são cruciais na análise oftalmológica, fornecendo informações sobre possíveis alterações patológicas [28][32].

[1] https://www.gov.br/saude/pt-br/assuntos/noticias/2023/fevereiro/oms-alerta-que-285-milhoes-de-pessoas-no-mundo-tem-a-visao-prejudicada

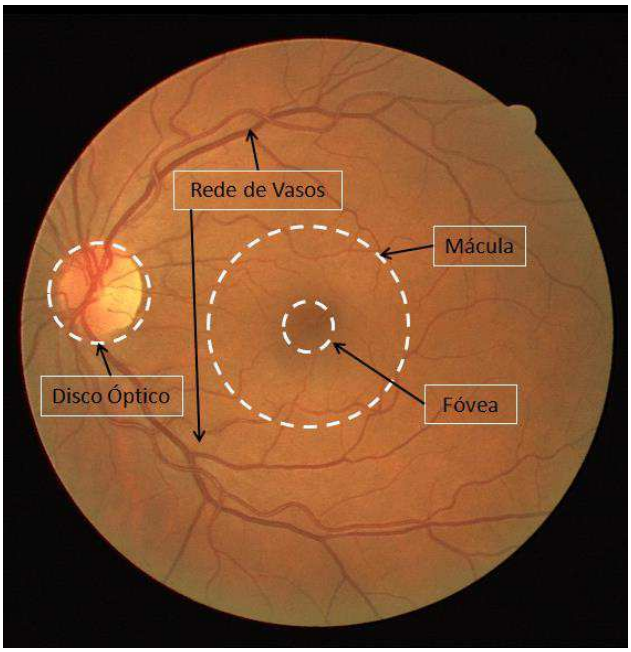

Figura 2: Principais estruturas de uma retina saudável. Fonte: Veras [50].

O disco óptico é a porção visível do nervo óptico, ele apresenta uma estrutura circular com bordas definidas. Apesar de corresponder ao ponto cego no campo de visão humano, geralmente não é perceptível devido à integração das informações nos córtices visuais.

A compreensão minuciosa da anatomia e do funcionamento do olho humano é essencial para a detecção precoce de doenças oculares e sistêmicas, ressaltando a importância da oftalmologia na preservação da saúde visual e geral.

2.2 Doenças Oculares

No contexto brasileiro, dados do último Censo Demográfico, conduzido pelo Instituto Brasileiro de Geografia e Estatística (IBGE) em 2010, revelam que mais de 35 milhões de pessoas apresentam algum grau de dificuldade visual [16].

O panorama oftalmológico no Brasil, conforme estatísticas do Ministério da Saúde, destaca uma série de doenças oculares prevalentes que representam a maioria dos atendimentos na área. Estas incluem catarata, glaucoma, conjuntivite, retinopatia diabética, degeneração macular relacionada à idade e erros de refração (miopia, hipermetropia, astigmatismo e presbiopia, também conhecida como vista cansada). É fundamental ressaltar que algumas destas condições, se não tratadas de forma adequada, podem resultar em perda total ou parcial da visão. Nas subseções seguintes, são apresentados maiores detalhes sobre três dessas doenças abordadas neste trabalho.

2.2.1 Glaucoma

O glaucoma é uma doença ocular neurodegenerativa[2] crônica e irreversível que afeta as células ganglionares da retina e seus axônios, resultando em danos ao nervo óptico (ONH) e perda progressiva da visão [4][51][21]. Essas células são responsáveis por transmitir os impulsos nervosos ao cérebro, sendo essenciais para a percepção visual [11]. Com cerca de 60 milhões de casos em 2010 e uma estimativa de 80 milhões em 2020, o glaucoma é uma das principais causas de cegueira em todo o mundo [36].

O alargamento anormal da escavação, conhecida como *Cup*, em comparação com o nervo óptico alargado, é um dos principais indicadores do glaucoma. Isso é medido pela relação do raio da escavação para o raio do nervo óptico, conhecido como *Cup-to-Disc Ratio* (CDR). Um CDR elevado sugere a presença da doença [4][51].

Na Figura 3, estão representadas as marcações do contorno do disco óptico (DO) e da escavação, realizadas por um oftalmologista. É visível a distinção entre as marcações do contorno do DO e da escavação em um olho saudável (Figura 3(a)) e em um olho com glaucoma (Figura 3(b)).

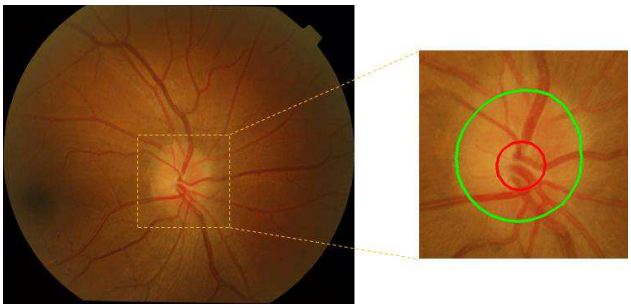
Figura 3a: Exemplo de retina saudável. Fonte: Claro et al [9].

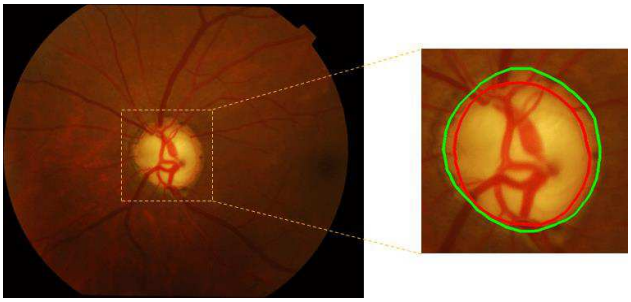
Figura 3b: Exemplo de retina com glaucoma. Fonte: Claro et al [9].

A detecção precoce dos danos glaucomatosos é desafiadora e não há um padrão definitivo para o diagnóstico. Os métodos oftalmológicos tradicionais incluem exames de acuidade visual, testes de campo visual, tonometria, exames de dilatação dos olhos e análise do nervo óptico. Sendo a análise do nervo óptico considerada a mais sensível para detectar glaucoma, porém é manual, subjetiva, demorada e cara [36][6][12][20]. Diante desses desafios, o uso de imagens digitais da retina tem sido cada vez mais explorado como uma ferramenta complementar ao diagnóstico clínico. Essa abordagem melhora a precisão diagnóstica, especialmente em casos complexos, e pode ser uma contribuição valiosa para a prática oftalmológica.

[2] As doenças neurodegenerativas são condições incuráveis e debilitantes que resultam em degeneração progressiva e/ou morte das células nervosas.

2.2.2 *Retinopatia Diabética*

A retinopatia diabética (RD) consiste em uma complicação ocular prevalente em pessoas com diabetes, caracterizada pelo dano aos vasos sanguíneos da retina devido aos altos níveis de glicose no sangue. Globalmente, a RD afeta milhões de pessoas, sendo uma das principais causas de deficiência visual, particularmente em países como o Brasil, onde cerca de 35 a 40% dos pacientes com diabetes são afetados [3].

A detecção precoce é primordial, especialmente considerando que entre 30% e 50% dos pacientes com diabetes não passam por triagem anual, o que aumenta o risco de complicações visuais graves. Os métodos diagnósticos tradicionais, como exames fundoscópicos e retinografias, têm uma sensibilidade considerável, mas métodos alternativos automatizados como Os modelos de análise automática de imagens da retina, que têm como objetivo principal diminuir a carga de trabalho dos médicos, oferecendo um método prático e eficiente com uma boa relação custo-benefício [30][46], estão sendo pesquisados para facilitar o processo de triagem.

A RD se desenvolve em estágios progressivos, cada um com implicações específicas para a visão. Desde os estágios iniciais com microaneurismas até os estágios avançados com proliferação de novos vasos sanguíneos e descolamento da retina, os sintomas podem variar de visão turva a perda súbita e permanente da visão.

A Figura 4 apresenta imagens ilustrativas dos diferentes estágios da RD, evidenciando a importância da identificação precoce para prevenir complicações visuais graves.

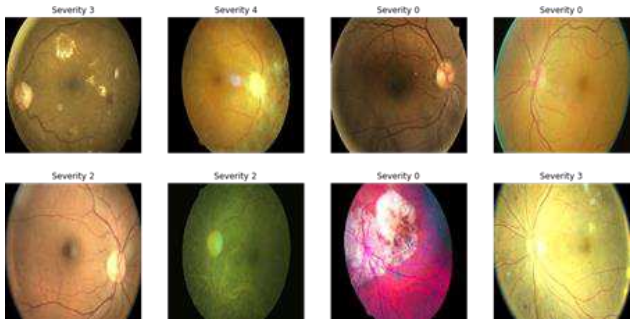

Figura 4: Imagens de fundo de retina com diferentes graus de gravidade. Fonte: Reis [38].

2.2.3 *Catarata*

A catarata é uma doença ocular em que o cristalino, localizado atrás da pupila, se torna opaco, prejudicando a visão. O cristalino atua como uma lente, concentrando a luz na retina, mas sua opacidade reduz essa capacidade ao longo do tempo [14]. A doença é definida como perda de transparência do que refrate a luz, acarretando um efeito adverso na visão, inicialmente, afeta a visão de longe e depois causa embaçamento gradual da visão. A mancha esbranquiçada na pupila é perceptível apenas em estágios avançados.

A catarata, considerada a principal causa de cegueira no mundo, não pode ser completamente evitada, pois pode ser causada por fatores genéticos e pela idade avançada. A detecção precoce é essencial para evitar complicações futuras [14][31].

A catarata pode ser caracteriza em: congênita, juvenil, senil e as relacionadas com doenças sistêmicas, intraoculares, traumas e substâncias tóxicas. A forma senil é a mais comum, concentrando-se nos estudos epidemiológicos e de prevalência.

Dentro do grupo senil, existem três subtipos: nuclear, cortical e subcapsular posterior (PSC), dependendo da localização da opacidade no cristalino. A Figura 5 mostra os tipos mais comuns de cataratas senis, incluindo a forma avançada ou madura [24].

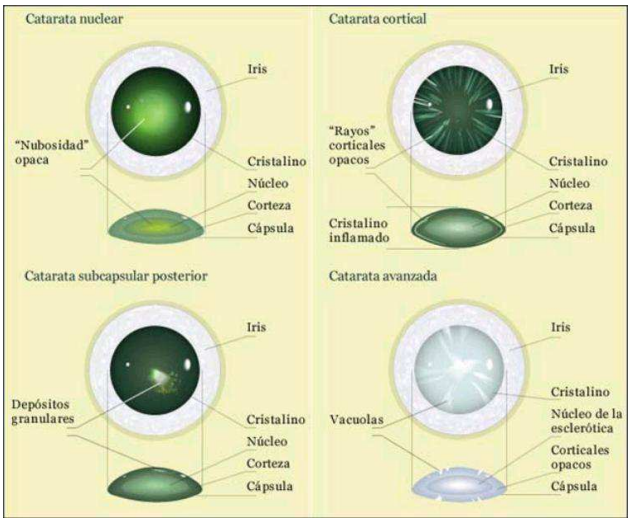

Figura 5: Principais tipos de cataratas senis. Fonte: Lopez et al [24].

O diagnóstico da catarata é realizado por oftalmologistas, que podem identificar lesões no cristalino através de uma análise com lâmpada de fenda. Sintomas como diminuição da visão, sensação de visão nublada, maior sensibilidade à luz e alterações na percepção de cores são indicativos. Existem diversos sistemas de classificação de opacidades, como o sistema LOCS III (Figura 6), que utiliza imagens de referência para graduar o desenvolvimento da doença [24].

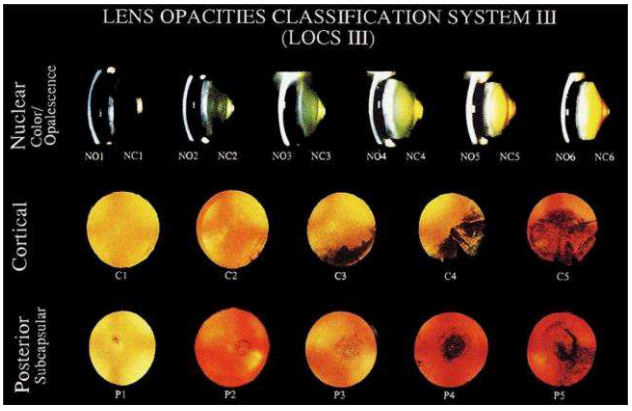

Figura 6: Sistema LOCS III de graduação de opacidade
Fonte: Lopez et al [24].

Avanços na cirurgia de catarata, incluindo técnicas de pequena incisão, o uso de viscoelásticos e o desenvolvimento de lentes intraoculares, tornaram o tratamento bastante eficaz, com rápida recuperação visual na maioria dos casos [5]. A cirurgia de catarata,

juntamente com o implante de uma lente intraocular, pode restaurar praticamente a visão normal.

2.3 Fundoscopia

A técnica de aquisição de imagens digitais de fundo de olho é uma modalidade essencial empregada no diagnóstico de doenças oculares, devido à sua natureza não invasiva [28]. Com apenas uma câmera, os oftalmologistas podem extrair características vitais das imagens do fundo de olho, como o disco óptico (OD) e os vasos sanguíneos, fornecendo informações cruciais para o diagnóstico da condição [28]. O equipamento utilizado para essa aquisição, conhecido como retinógrafo, consiste em um microscópio conectado a uma câmera e uma fonte de luz, projetado especificamente para capturar imagens da retina, o segmento posterior ocular [28].

A fundoscopia, realizada através do retinógrafo, é a técnica de triagem mais amplamente utilizada para a detecção de doenças oculares, sendo de maior preferência devido à sua simplicidade e custo mais baixo em comparação com outros métodos de análise de retina disponíveis [2].
O exame de fundo de olho, ou fundoscopia, é crucial para que oftalmologistas possam avaliar alterações oculares, incluindo aquelas em pacientes diabéticos, enfocando as estruturas do fundo do olho, como o nervo óptico, os vasos sanguíneos e a mácula. No entanto, sua interpretação depende muito da habilidade e experiência do especialista, o que ressalta a necessidade de sistemas de apoio ao diagnóstico como uma ferramenta valiosa no contexto clínico atual [1].

2.4 Avanços Científicos e Tecnológicos

Nos últimos anos, avanços na ciência e tecnologia têm proporcionado oportunidades significativas na área da oftalmologia, especialmente no combate a doenças oculares. A aplicação de tecnologias de inteligência artificial, como as Redes Neurais Convolucionais (CNNs), tem demonstrado eficácia em várias áreas médicas, incluindo radiologia e patologia, devido à sua capacidade de análise de imagens, a aplicação de tecnologias de inteligência artificial na oftalmologia tem o potencial de melhorar o diagnóstico e a graduação de doenças oculares, como glaucoma, catarata e retinopatia diabética. Sistemas de diagnóstico auxiliados por computador têm como objetivo melhorar a consistência e eficiência dos diagnósticos, colaborando com médicos especialistas.

2.4.1 Rede Neural Convolucional

A Rede Neural Convolucional é um modelo de aprendizado profundo adequado ao processamento de sinais e imagens. O pré-processamento exigido em uma CNN é muito menor em comparação com outros algoritmos de classificação. Enquanto nos métodos primitivos, os filtros são feitos à mão, com treinamento suficiente, as CNNs têm a capacidade de aprender esses filtros [13]. As CNNs surgiram como uma ferramenta poderosa no trabalho de classificação de imagens, especialmente médicas. Elas proporcionam precisão no diagnóstico da doença alvo, ao mesmo tempo em que reduzem o custo computacional.
De acordo com LeCun et al. (1998) [23], as CNNs são modelos biologicamente inspirados que aprendem características em múltiplos estágios. Sua arquitetura é composta por camadas organizadas em módulos ou blocos (Figura 7). Esses módulos são comumente empilhados para criar um modelo profundo. A imagem é introduzida diretamente na rede, passando por diversos estágios de convolução e *pooling*. Em seguida, as representações resultantes dessas operações são alimentadas em uma ou mais camadas totalmente conectadas. Por fim, a última camada totalmente conectada produz o rótulo indicativo da classe [37].

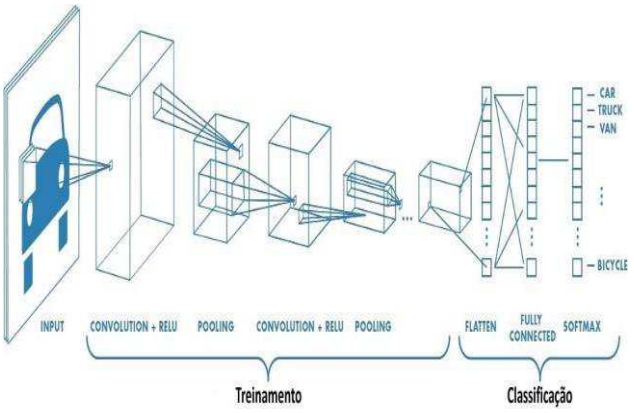

Figura 7: Exemplo de uma típica CNN em uma tarefa de classificação. Fonte: Rawat, W., Wang, Z [37].

2.4.2 Aprendizado por Transferência

O aprendizado por transferência é uma técnica que permite aplicar o conhecimento adquirido por uma rede neural em uma determinada tarefa para outra tarefa relacionada. Essa abordagem envolve o uso de um modelo pré-treinado em uma tarefa semelhante como base para resolver outro problema relacionado, transferindo o conhecimento adquirido para a nova tarefa. Isso resulta em economia de tempo e recursos computacionais.
Por exemplo, uma Rede Neural Convolucional treinada para reconhecimento de objetos em imagens pode ser ajustada para treinar e classificar objetos de interesse em uma imagem diferente. Nas redes convolucionais, os filtros aprendem informações relevantes sobre a imagem de entrada. As camadas iniciais identificam características básicas, como bordas e texturas, enquanto as camadas mais profundas capturam informações mais específicas. Utilizar modelos pré-treinados pode ser eficiente, já que esses modelos aprenderam pesos relevantes anteriormente em grandes conjuntos de dados.
Ao adotar a aprendizagem por transferência, pode-se aproveitar informações valiosas de modelos pré-existentes, como características e pesos. Isso é especialmente útil quando não se dispõe de um grande conjunto de dados próprios, pois reduz o custo computacional em comparação com o treinamento de uma CNN do zero.
Nesse processo, as arquiteturas de modelos pré-treinados, como ResNet50 [19], VGG16 [44] e InceptionV3 [48], são usadas como extratores de características. A saída de uma camada anterior à camada de saída é empregada como entrada para um novo classificador. Esses modelos pré-treinados são essenciais como inicialização de pesos, reduzindo a necessidade de treinamento a partir do zero.

2.4.2.1 ResNet50

As Redes Neurais Residuais (ResNet) foram desenvolvidas por pesquisadores da *Microsoft Research* [19] em 2015 e se destacam pelo uso de blocos residuais, que permitem atalhos entre camadas. Esses atalhos ajudam a mitigar o problema do gradiente desvanecente em redes mais profundas, permitindo o treinamento eficaz de modelos com um grande número de camadas.
A arquitetura da ResNet50 utiliza blocos residuais compostos por três camadas de convolução com diferentes tamanhos de filtro. Esses blocos são conectados por atalhos, permitindo que os gradientes fluam mais facilmente durante o treinamento. Isso ajuda a resolver o problema do desaparecimento do gradiente e permite a construção de redes mais profundas e eficazes [47].

2.4.2.2 VGG 16

A arquitetura VGGNet [44] é uma versão aprimorada da arquitetura AlexNet [22], composta por modelos VGG16 e VGG19. Ela consiste em uma série de blocos com diferentes camadas de convolução, subamostragem e conexões densas. A VGG16, por exemplo, possui 13 camadas convolucionais, 5 camadas de subamostragem e 3 camadas densas, totalizando 16 camadas de peso. Essa arquitetura venceu a competição ILSVR em 2014 e é reconhecida por sua eficácia em tarefas de visão computacional.
Na área de oftalmologia, as arquiteturas comumente utilizadas fazem parte da biblioteca KERAS[3]. Essas arquiteturas incluem pesos pré-treinados para facilitar a previsão e extração de recursos. Uma vantagem significativa de usar as arquiteturas disponíveis no KERAS é que elas foram testadas e avaliadas em grandes conjuntos de dados, demonstrando eficácia na classificação de imagens. Além disso, muitas dessas arquiteturas são pré-treinadas em conjuntos de dados abrangentes, o que permite que a rede neural seja utilizada como ponto de partida para tarefas específicas de classificação.

2.4.2.3 Inception V3

A arquitetura Inception V3, proposta por Szegedy et al. (2015) [48], destaca-se pela presença de módulos chamados inception. Esses módulos são extratores de características convolucionais que aprendem representações ricas de informação com poucos parâmetros. Tradicionalmente, uma camada convolucional aprende filtros em um espaço 3D por meio das dimensões altura, largura e dimensão de canal de cor. A ideia por trás do bloco Inception é simplificar o processo de mapeamento entre canais e correlações espaciais, fatorando explicitamente a série de operações que examinam de forma independente essas correlações, em contraste com o modelo tradicional em que um único kernel é responsável por essas tarefas.
O modelo Inception V3, baseado nos conceitos de *inception* do Googlenet e desenvolvido após o ILSVRC 2014, introduz melhorias significativas em relação aos modelos inception anteriores [48]. Inclui técnicas como regularização por suavização de rótulos (LSR) e um classificador auxiliar. O LSR é aplicado para reduzir o *overfitting*, tornando o modelo menos confiante e distribuindo melhor as probabilidades entre os rótulos. Aliado a um classificador auxiliar, permite que a rede seja treinada com grandes conjuntos de dados sem sofrer overfitting.
A estrutura da Inception V3 é composta por uma série de blocos contendo camadas convolucionais, *average pooling*, *max-pooling*, *dropout*, além de camadas totalmente conectadas e *softmax*, totalizando 42 camadas e quase 30 milhões de parâmetros.

2.4.3 Interpretabilidade e Explicabilidade

Apesar de passarem por vertiginosos avanços, as técnicas de aprendizado de máquina ainda representam um desafio significativo para a adoção generalizada em áreas críticas, como medicina, mercado financeiro e sistema de justiça criminal, onde as decisões baseadas em inferências podem ter impactos substanciais. Torna-se crucial compreender as decisões tomadas por essas máquinas e entender como essas previsões são geradas pode contribuir para aprimorar o modelo, aumentando sua confiabilidade.
Nesse contexto, surge a necessidade de compreender, racionalizar e justificar as decisões tomadas pelo modelo, o que implicitamente promove confiança nas previsões feitas [32].
A interpretabilidade surge como uma tendência em ascensão, permitindo apresentar informações úteis para explicar predições individuais ou expor seu funcionamento interno em termos compreensíveis para os humanos.
As explicações geradas pelas técnicas de interpretabilidade devem ser capazes de identificar as principais causas do comportamento do modelo, sem necessariamente exigir o entendimento exato de cada decisão [33]. Um modelo confiável e preciso em suas previsões torna-se também uma fonte de conhecimento, além dos próprios dados, e a interpretabilidade permite extrair esse conhecimento.
Segundo Molnar (2022) [7], a interpretabilidade no aprendizado de máquina busca fornecer explicações para o comportamento do modelo na tomada de decisões.
O termo interpretabilidade não possui um significado amplamente aceito na comunidade especializada, sendo muitas vezes utilizado de forma descuidada. No entanto, é essencial compreender o objetivo subjacente em cada contexto de uso. Termos como transparência, confiança, explicabilidade e auditabilidade têm significados mais precisos no contexto da interpretabilidade e são categorizados em três objetivos abstratos:

- Ética e Regulação: Refere-se ao uso da interpretabilidade para garantir conformidade com regulamentações e padrões éticos. Isso inclui adequação às leis que exigem explicabilidade em decisões automatizadas, como a GDPR da União Europeia.
- Apoio à Decisão: Envolve o uso de interpretabilidade durante a utilização prática de modelos de aprendizado de máquina para fornecer informações que auxiliem na compreensão das decisões tomadas pelo modelo. Isso é particularmente importante em cenários críticos, como diagnósticos médicos.
- Confiança e Entendimento: Diz respeito à exposição do funcionamento interno do modelo, o que estabelece confiança na sua capacidade preditiva e ajuda na compreensão do problema e da solução. Isso é útil durante a fase de modelagem e validação do modelo.

Os termos frequentemente associados à interpretabilidade, como transparência, explicabilidade e auditabilidade, referem-se às características que um modelo pode possuir. A presença dessas características torna viáveis os objetivos anteriormente apresentados. Nos tópicos a seguir, esses termos são melhor contextualizados:

- Auditabilidade: Modelos auditáveis são essenciais para atender às exigências dos órgãos reguladores e serem utilizados em aplicações críticas, como medicina, direito

[3] https://keras.io/

e aeronáutica. A auditabilidade está vinculada aos objetivos categorizados por ética e regulação.

- Explicabilidade: Modelos explicáveis têm a capacidade de fornecer explicações aos usuários e estão associados aos objetivos categorizados por apoio à decisão.
- Transparência: Modelos transparentes podem ser apresentados em termos compreensíveis para os seres humanos e estão relacionados aos objetivos categorizados por confiança e compreensão.

As técnicas de interpretabilidade podem ser categorizadas de acordo com seu funcionamento e escopo:

- Intrínsecas ou Pós Treinamento: Técnicas de interpretabilidade podem ser classificadas como intrínsecas ou pós treinamento (*Intrinsic* ou *Post-hoc*). As técnicas intrínsecas, aproveitam a própria estrutura do modelo para fornecer interpretabilidade. As pós treinamento são aplicadas após o treinamento do modelo.
- Local ou Global: O escopo das técnicas de interpretabilidade pode ser local ou global. As técnicas globais visam entender como o modelo faz previsões em um contexto geral, descrevendo seu comportamento esperado. Já as técnicas locais buscam explicar como o modelo faz previsões em casos ou grupos específicos, detalhando as previsões para conjuntos de interesse.
- Agnóstica a Modelo ou Específica a Modelo: A abrangência da técnica pode ser classificada como agnóstica a modelo ou específica a modelo (*Model Agnostic*) ou *Model Specific*). Técnicas especificas são limitadas a determinadas classes de modelos, enquanto as técnicas agnósticas à arquitetura do modelo avaliam o modelo após o treinamento, lidando apenas com as previsões retornadas pelo modelo, e não com sua estrutura interna. Assim, essas técnicas podem ser aplicadas a qualquer tipo de modelo.

2.4.3.1 *LIME*

O LIME (*Local Interpretable Model-agnostic Explanations*) é uma técnica introduzida por (Ribeiro;Singh;Guestrin, 2016) [39]. Projetada para tornar os modelos de aprendizado de máquina mais transparentes e interpretáveis, especialmente em nível local. Em vez de fornecer uma explicação abrangente para o todo o modelo, o LIME se concentra em explicar as previsões para instâncias de dados individuais, permitindo que os usuários compreendam como o modelo chegou a uma decisão específica para um caso particular. Os modelos, como as redes neurais profundas, são frequentemente complexos e difíceis de interpretar. O LIME aborda essa complexidade gerando modelos locais (surrogates), que são muito mais simples do que o modelo de aprendizado de máquina original, mas ainda são explicativos. Esses modelos locais são criados para imitar o comportamento do modelo original em torno da instância de dados de interesse.

O processo do LIME se dá ao examinar o efeito das perturbações nos dados de entrada do modelo sobre a previsão. Isso resulta na geração de um novo conjunto de dados contendo a imagem original com perturbações simples e as respectivas previsões fornecidas pelo modelo. Em seguida, a técnica treina um novo modelo interpretável utilizando esse conjunto de dados, o qual se baseia na proximidade entre as amostras geradas e a amostra real. Esse modelo interpretável é projetado para oferecer uma boa aproximação local, embora possa não ser tão precisa globalmente, esse tipo de acurácia é denominada fidelidade local pelos autores. (Molnar, 2019; Ribeiro; Singh; Guestrin, 2016) [39][26].

Para imagens faz sentido perturbar grupos de pixels, visto que mais de um pixel influencia as características de uma classe. A Figura 8 ilustra como a segmentação de pixels é empregada para explicar o conjunto de pixels que mais contribui para essa explicação.

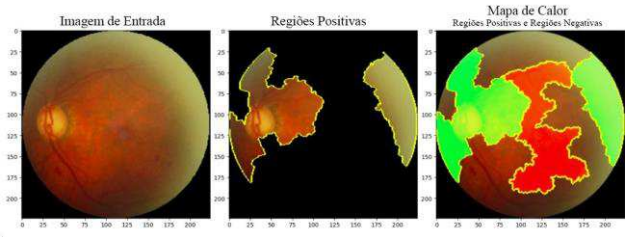

Figura 8: Previsão do LIME. Fonte: Elaborada pelo Autor (2024).

2.4.3.2 *Grad-CAM*

O Grad-CAM(*Gradient-weighted Class Activation Mapping*), ou Mapeamento de Ativação de Classe Ponderada por Gradiente [42], é uma generalização do CAM (*Class Activation Maps*), introduzido por Zhou et al. (2015). É uma técnica de interpretação utilizada em modelos de aprendizado profundo, que auxilia na visualização das partes mais influentes de uma imagem na tomada de decisão do modelo durante a classificação [43].

O Grad-CAM é aplicado principalmente em modelos de CNN, os quais são amplamente utilizados em tarefas de visão computacional. Esses modelos consistem em várias camadas convolucionais que são seguidas por camadas totalmente conectadas, as quais realizam a decisão final sobre a classe à qual a imagem pertence. Tais camadas produzem mapas de ativação, os quais destacam as regiões da imagem onde determinadas características foram detectadas. O Grad-CAM concentra-se em compreender como essas camadas totalmente conectadas ponderam as características extraídas pelas camadas convolucionais.

O processo pode ser descrito da seguinte maneira: inicialmente, uma imagem é fornecida ao modelo, que produz uma previsão. A classe alvo é determinada com base nessa previsão. Em seguida, o Grad-CAM calcula os gradientes da classe alvo em relação às ativações na camada convolucional mais profunda. Isso indica quais ativações foram mais influentes na decisão da classe. O Grad-CAM, então, combina esses gradientes com as ativações da camada convolucional usando uma média ponderada [8]. Essa média ponderada é utilizada para obter os pesos de ativação de cada canal na camada convolucional. Por fim, esses pesos de ativação são empregados para criar um mapa de ativação que destaca as regiões cruciais da imagem que mais contribuíram para a decisão da classe alvo.

Na Figura 9, é possível observar o resultado da aplicação do Grad-CAM para visualizar as zonas de ativações do modelo na tarefa de classificação de glaucoma. As áreas destacadas em vermelho concentram-se nas regiões que o modelo considera mais significativo na classificação.

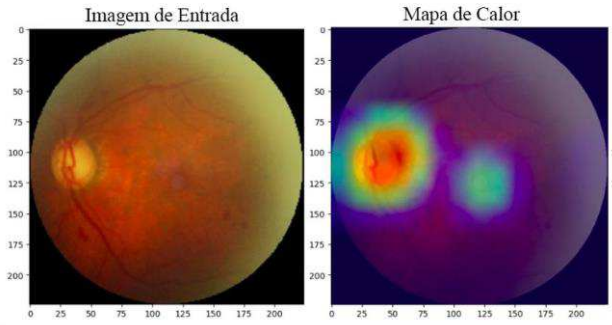

Figura 9: Previsão do Grad-CAM. Fonte: Elaborada pelo Autor (2024).

3. MATERIAIS E MÉTODOS

A metodologia utilizada nesse estudo compreende a implementação e comparação das técnicas de explicabilidade LIME e Grad-CAM, aplicadas aos modelos de CNNs (VGG16, ResNet50 e InceptionV3) treinados para o diagnóstico de três doenças oculares: catarata, glaucoma e retinopatia diabética. A linguagem de programação Python foi utilizada nessa implementação com os auxílios das bibliotecas *Keras* e *TensorFlow*[4]

Após o treinamento dos modelos, uma avaliação comparativa é realizada, analisando quantitativa e qualitativamente os resultados apresentados pelo treinamento dos modelos e a capacidade de localização de regiões relevantes nas imagens pelas técnicas de explicabilidade. A análise quantitativa engloba a medição e comparação de métricas específicas, como: Acurácia, Precisão, Revocação e pontuação-F1. Além disso, é realizada uma análise qualitativa, examinando-se visualmente os mapas de calor e áreas destacadas produzidas por cada técnica, a fim de compreender como as decisões dos modelos foram baseadas nas características oculares.

3.1 Conjuntos de Dados

O conjunto de dados utilizado foi obtido da plataforma do *Kaggle*[5]. A base de dados de nome *Eye Diseases Classification*[6], consiste em uma base de dados público, contendo imagens no formato JPEG, no sistema RGB (vermelho, verde e azul, do inglês *red*, *green* e *blue*), com imagens de fundo do olho, que foram coletadas de várias fontes como IDRiD [35], Oculur recognition, HRF etc.

A base de dados possui imagens das classes normal, glaucoma, retinopatia diabética e catarata contendo ao todo 4217 imagens, das quais 1074 pertencem à classe normal, 1007 à classe glaucoma, 1098 à classe retinopatia diabética e 1038 à classe catarata.

O uso da base de dados se deu pelos seguintes motivos: a quantidade elevada de imagens, a diversidade das estruturas internas presentes nas imagens e a qualidade da resolução das imagens.

As imagens foram adquiridas com diferentes tamanhos de largura e altura em pixels, foi feito o redimensionamento de todas as imagens para o tamanho 224x224 pixels, pois todos os modelos utilizados esperam imagens com esse dimensionamento e são normalizadas com base na média e no desvio padrão das imagens no conjunto de treinamento *ImageNet*[7] [40].

Os conjuntos de dados gerados a partir da aplicação das estratégias de pré-processamento citadas acima foram submetidos ao procedimento de aumento de dados com base nas rotações incrementais das imagens com valores aleatórios entre -20º e 20º. Utilizando a classe *ImageDataGenerator* do *Keras*, a base de dados foi dividida aleatoriamente da seguinte forma de um total de 4217 imagens, dados de treinamento 3586 imagens e dados de validação 631, ou seja, 85%/15%. Em relação às classes, a divisão entre dados de treinamento e dados de validação pode ser vista na Tabela 01.

	TREINAMENTO	VALIDAÇÃO
catarata	883	155
retinopatia diabética	934	164
glaucoma	856	151
normal	913	161

Tabela 1: Distribuição das imagens da base de dados. Fonte: Elaborada pelo Autor (2024).

3.2 Arquiteturas

Neste trabalho, foram exploradas as arquiteturas VGG16, ResNet50 e InceptionV3, todas com pesos pré-treinados na base de dados *ImageNet*. Utilizou-se a técnica de ajuste fino da transferência de aprendizagem para adaptar essas arquiteturas à classificação de uma nova base de dados. Essa abordagem permite refinar as camadas das redes neurais convolucionais, aproveitando-se o conhecimento prévio adquirido durante o treinamento na base de dados ImageNet.

O processo de treinamento se dá com a inicialização dos pesos, utilizando-se os pesos pré-treinados de modelos treinados na *ImageNet* como ponto de partida. Para adaptar as redes ao novo problema de classificação com apenas 4 classes, remove-se a camada de saída original e substitui-se por uma camada de *Global Average Pooling*, seguida por duas camadas densas personalizadas. Essas camadas densas possuem ativação ReLU e são regularizadas com técnica de regularização L2 para evitar *overfitting*. Todas as camadas são mantidas treináveis, permitindo que os pesos se ajustem ao novo conjunto de dados durante o treinamento.

Para o treinamento de cada modelo foi utilizado a técnica de regularização *early stopping* por ser um mecanismo que interrompe o treinamento se a métrica monitorada *(val_loss)* não melhorar após 3 épocas. Devido à natureza do problema, que envolve a classificação de múltiplas classes, foi selecionada a Softmax como função da camada de saída. Para a função de custo foi utilizada a *Categorical Crossentropy*, uma das principais funções utilizadas na literatura para obter o valor de custo em problemas com múltiplas classes devido às suas propriedades matemáticas, capacidade de promover previsões precisas e facilitar o treinamento e a interpretação do modelo. Como otimizador foi escolhido o *Adam* configurado a uma taxa de aprendizado de 0,0001 devido à sua

[4] https://www.tensorflow.org/?hl=pt-br
[5] https://www.kaggle.com/
[6] https://www.kaggle.com/datasets/gunavenkatdoddi/eye-diseases-classification
[7] https://www.image-net.org/challenges/LSVRC/

eficiência computacional, velocidade de convergência e adaptabilidade a uma variedade de problemas de otimização em aprendizado profundo.

3.3 Matriz de Confusão e Métricas de Avaliação

A matriz de confusão, também conhecida como tabela de contingência, é uma ferramenta amplamente utilizada no campo do aprendizado de máquina. Ela fornece informações sobre as classificações reais e esperadas realizadas por um classificador [52], simplificando o processo de avaliar se o modelo está confundindo as classes. Esta matriz é construída com base no número de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos [15]. Em resumo, compara as previsões de um modelo com o padrão real, a diagonal principal representa as classes que foram corretamente previstas (*True Positives* - TP), enquanto os elementos fora dessa diagonal representam as classificações equivocadas (*False Positives* - FP). Existem várias métricas que podem ser calculadas usando a análise da matriz de confusão. Uma delas é a Acurácia, que mede a taxa de classificação correta em relação ao número total de classificações realizadas, mas que pode ser problemática em base de dados desbalanceados [49]. A Precisão, que indica a taxa de verdadeiros positivos em relação ao total de positivos classificados pelo modelo, calculada como a razão entre previsões corretas (TP) e o total de observações positivas previstas da classe. A Revocação indica a capacidade do modelo de classificar corretamente elementos da classe positiva, sendo calculada como a razão entre o número de exemplos classificados corretamente como pertencentes à classe real (TP) e o total de exemplos que pertencem a essa classe (TP+FN). A pontuação-F1, também conhecido como F-Score, é a média harmônica entre precisão e revocação, sendo particularmente vantajosa porque leva em consideração tanto a revocação quanto a precisão, fornecendo uma avaliação mais precisa do desempenho [34].

3.4 Técnicas de Explicabilidade

Na implementação das técnicas de explicabilidade foi utilizada a versão do Keras 2.15.0 assim como a versão 2.15.0 do *TensorFlow*. Dentro da estrutura experimental, utilizamos tanto o LIME quanto o Grad-CAM de maneira independente em cada imagem. Para o Grad-CAM, foi utilizada uma implementação baseada nos exemplos oferecidos pelo *Keras*[8]. Esse método gera um mapa de calor que mostra a intensidade da ativação em diferentes regiões da imagem. As regiões mais quentes indicam as partes da imagem que tiveram maior contribuição para a predição da classe de interesse, enquanto as regiões mais frias indicam áreas menos relevantes para a classificação.

Por outro lado, o LIME é utilizado com seu próprio pacote, fornecido pelos autores da técnica, com a versão 0.2.0.1 utilizada nos experimentos. O resultado do LIME é fornecido na forma de duas imagens. A primeira imagem mostra as regiões da entrada que foram consideradas importantes pelo modelo para a tomada da decisão de classificação. Essas regiões aumentaram a probabilidade ou a confiança da predição para a classe atribuída pelo modelo. A segunda imagem mostra as regiões positivas e negativas da imagem, onde as regiões verdes indicam uma contribuição positiva para a predição, ou seja, regiões que aumentaram a probabilidade da classe atribuída, enquanto as regiões vermelhas indicam uma contribuição negativa, ou seja, regiões que diminuíram a probabilidade da classe atribuída. Essa representação ajuda a entender não apenas quais partes da entrada foram importantes, mas também o impacto relativo de cada parte na decisão final do modelo.

3.5 Ambiente de Desenvolvimento

Como ambiente de desenvolvimento foi utilizado o *Google Colaboratory*, ou *Colab*[9], uma ferramenta do Google que permite aos usuários escrever e executar códigos *Python* através do navegador. O *Colab* é amplamente utilizado para aprendizado de máquina e análise de dados. Tecnicamente, o *Colab* é um serviço de *Jupyter Notebook* hospedado na nuvem, que dispensa instalações prévias e oferece acesso gratuito a recursos computacionais, incluindo *Graphics Processing Units* (GPUs).

As Máquinas Virtuais (VMs) disponibilizadas pelo *Colab* vêm pré-configuradas com uma configuração padrão para VMs com GPU, geralmente utilizando Nvidia K80 com 12GB de memória, além das bibliotecas essenciais para aprendizado de máquina e inteligência artificial, como *TensorFlow, Matplotlib e Keras*.

O uso do Colab se deu para eliminar a dependência de acesso físico a um computador para realizar experimentos, garantindo assim a reprodutibilidade dos resultados. Além disso, não é necessário um computador com hardware específico para obter bom desempenho nos experimentos. Outra vantagem, inclui a facilidade de execução sem a necessidade de configurar um ambiente de desenvolvimento local.

A desvantagem encontrada é que após um certo período de execução, a VM é desativada automaticamente, resultando na perda de todos os dados e configurações do usuário.

4. RESULTADOS E DISCUSSÕES

Nas próximas seções, serão apresentados os resultados obtidos. Na Seção 4.1, destacam-se os resultados da análise de desempenho das redes neurais convolucionais implementadas. Onde o modelo VGG16 obteve a melhor acurácia, seguido pelo ResNet50 e InceptionV3.

Já na Seção 4.2, foram discutidos os resultados da análise de desempenho das técnicas de explicabilidade, as técnicas de LIME e Grad-CAM foram aplicadas a 100 imagens de teste. As análises comparativas entre as regiões destacadas por cada técnica e as regiões críticas mencionadas anteriormente forneceram esclarecimentos valiosos sobre o comportamento dos modelos no diagnóstico das doenças oculares.

4.1 Resultados da Análise de Desempenho das Redes Neurais Convolucionais Implementadas

Todos os modelos foram treinados com *Early Stopping*, preservando assim as melhores métricas evitando overfitting, o modelo com melhores métricas foi salvo e aplicado na base de dados de validação, buscando avaliar quão bom o modelo está performando para amostras completamente novas para os modelos. Na Figura 10 e Tabela 2, é possível observar o comportamento da acurácia (*accuracy*) e perda (*loss*) na base de dados de treino e validação. A partir desses dados, é possível identificar o melhor resultado dos modelos com respeito à base de dados de validação.

[8] https://keras.io/examples/vision/grad_cam/

[9] https://colab.research.google.com/

Dentre eles o que obteve o melhor resultado foi o VGG16 com acurácia de 93,17% na base de treino e 87,16% na base de validação. O modelo ResNet50 obteve 98,91% na base de treino e 85,58% na base de validação, enquanto o modelo InceptionV3 obteve 99,41% na base de treino e 82,88% na base de validação.

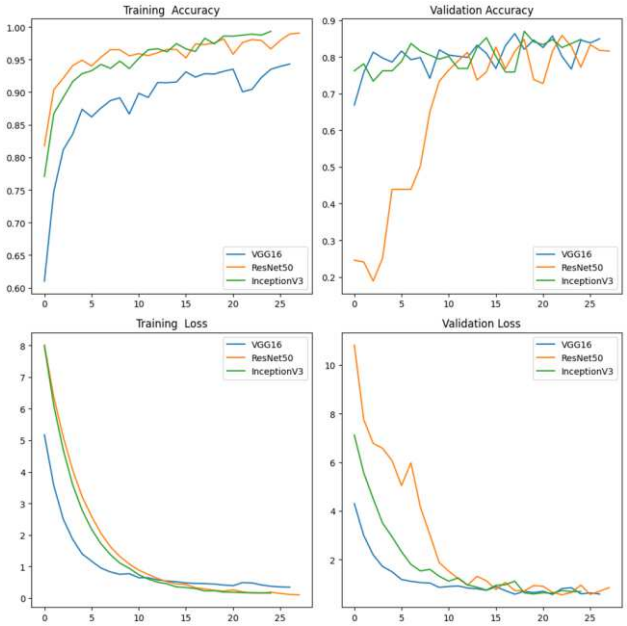

Figura 10: Curvas de aprendizagem da base de treino e validação. Fonte: Elaborada pelo Autor (2024).

Modelo	Acurácia		Perda	
	Treino	Validação	Treino	Validação
VGG16	0,9317	0,8716	0,3894	0,5648
ResNet50	0,9891	0,8558	0,1389	0,55
InceptionV3	0,9941	0,8288	0,1612	0,6174

Tabela 2: Métricas de acurácia e perda dos modelos treinados. Fonte: Elaborada pelo Autor (2024).

Na Figura 11 são apresentadas as matrizes de confusão, dos modelos VGG16, ResNet50 e InceptionV3. Nas linhas observam-se as classes preditas e, nas colunas, as classes esperadas (*ground truth*). A partir da matriz de confusão, é possível visualizar com mais facilidade algumas informações, como a quantidade de vezes que um modelo classifica uma classe nas demais classes. De modo geral, um ponto interessante de analisar é a dificuldade dos modelos predizer a algumas classes. Como pode ser visto a classe glaucoma é confundida com a classe normal por possuírem características semelhantes. A partir disso, a maior quantidade de erros da classe normal foram predições na classe glaucoma e a maior quantidade de erros da classe glaucoma foram predições na classe normal.

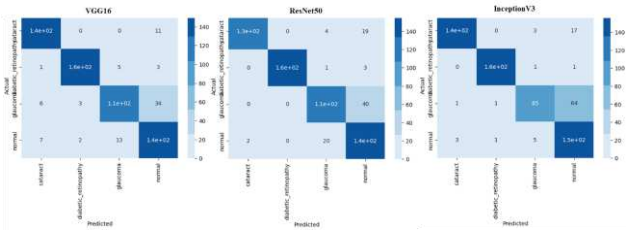

Figura 11: Matrizes de confusão dos modelos VGG16, ResNet50 e InceptionV3. Fonte: Elaborada pelo Autor (2024).

A partir das matrizes de confusão geradas, obtiveram-se as métricas de precisão, revocação e pontuação-F1, que são apresentadas nas Tabelas 3, 4 e 5, respectivamente. A partir dessas métricas é possível identificar as classes retinopatia diabética e catarata como as mais fáceis para a identificação com valores de pontuação-F1 em média de 96,50% e 91,67% e desvio padrão de 2,06% e 0,47% respectivamente.

Já as classes em que houve maior dificuldade para identificação foram normal e glaucoma, com valores de pontuação-F1 em média de 75,75% e 80,75% e desvio padrão de 2,66% e 7,30%, respectivamente. Uma possível justificativa para esse comportamento seria o fato das características das imagens serem bem semelhantes, diferindo apenas na escavação do disco óptico.

VGG16			
Classe	Precisão	Revocação	Pontuação-F1
catarata	0,91	0,93	0,92
retinopatia diabética	0,97	0,95	0,96
glaucoma	0,86	0,72	0,78
normal	0,74	0,86	0,8

Tabela 3: Métricas de precisão, revocação e pontuação-F1 do modelo VGG16. Fonte: Elaborada pelo Autor (2024).

ResNet50			
Classe	Precisão	Revocação	Pontuação-F1
catarata	0,99	0,85	0,91
retinopatia diabética	1,00	0,98	0,99
glaucoma	0,82	0,74	0,77
normal	0,69	0,86	0,77

Tabela 4: Métricas de precisão, revocação e pontuação-F1 do modelo Resnet50. Fonte: Elaborada pelo Autor (2024).

InceptionV3			
Classe	Precisão	Revocação	Pontuação-F1
catarata	0,97	0,87	0,92
retinopatia diabética	0,99	0,99	0,99
glaucoma	0,9	0,56	0,69
normal	0,65	0,94	0,77

Tabela 5: Métricas de precisão, revocação e pontuação-F1 do modelo InceptionV3. Fonte: Elaborada pelo Autor (2024).

4.2 Resultados da Análise de Desempenho das Técnicas de Explicabilidade Implementadas

Neste experimento foi utilizada a base de validação, onde as técnicas de explicabilidade foram aplicadas a 100 imagens, tendo sido gerado um número aleatório de imagens por classe. Nas Figuras 12, 13 e 14, são apresentados alguns dos resultados gerados de cada modelo de aprendizagem para cada classe predita, onde a primeira imagem é a da imagem que entra no modelo, a segunda é aplicação do LIME onde representa o superpixel que mais contribuiu para a predição, a terceira imagem é aplicação do LIME e representa os superpixels de maior (verde) e menor contribuição aplicada na imagem original. Por fim, a última imagem é a aplicação do Grad-CAM, representando o mapa de calor onde as cores quentes destacam as regiões que influenciaram de forma positiva para a predição.

A partir das informações apresentadas nas Seções 2.4.2.4.1 e 2.4.2.4.2 de como como são geradas as explicações apresentadas pelas técnicas do LIME e Grad-CAM, foram elaboradas algumas análises e explicações, com base na comparação entre as regiões destacadas por cada técnica de explicabilidade e as regiões críticas mencionadas nas Seções 2.2.1, 2.2.2 e 2.2.3, que desempenham um papel fundamental no diagnóstico das doenças oculares em questão.

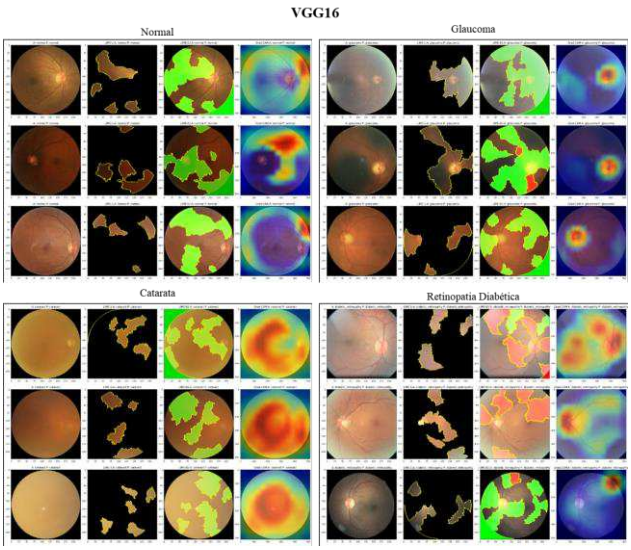

Figura 12: Aplicação do LIME e Grad-CAM na base de validação predica como correta no modelo VGG16. Fonte: Elaborada pelo Autor (2024).

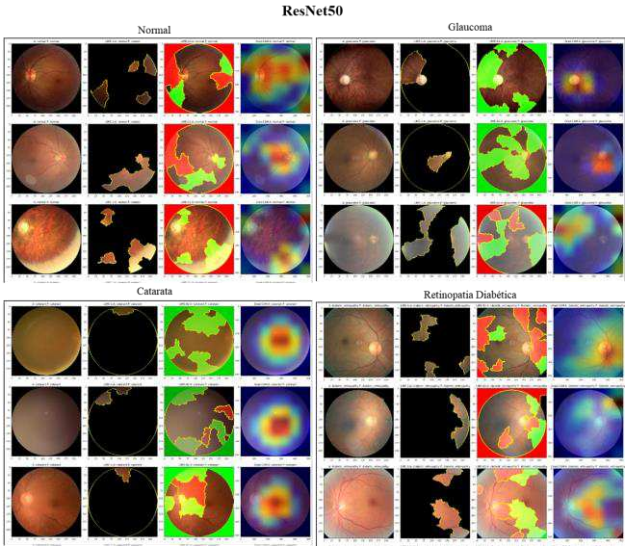

Figura 13: Aplicação do LIME e Grad-CAM na base de validação predica como correta no modelo ResNet50. Fonte: Elaborada pelo Autor (2024).

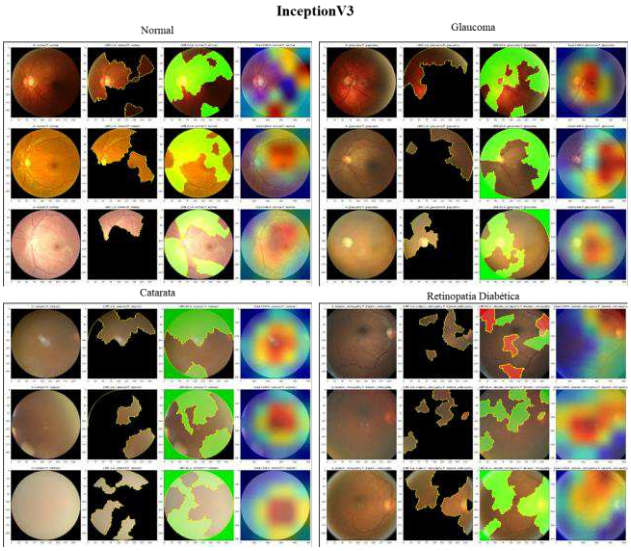

Figura 14: Aplicação do LIME e Grad-CAM na base de validação predica como correta no modelo InceptionV3. Fonte: Elaborada pelo Autor (2024).

Ao analisar as imagens geradas pelos modelos VGG16, ResNet50 e InceptionV3, conforme ilustrado nas Figuras 12, 13 e 14, podemos observar que tanto o LIME quanto o Grad-CAM abordam de maneira distinta a demarcação das regiões críticas para o diagnóstico de cada condição oftalmológica.

O LIME, ao analisar imagens preditas como olho normal, destaca várias regiões em busca de identificar anomalias, mas não demonstra um padrão. Nas imagens preditas como glaucoma, concentra suas marcações no disco óptico, área crucial para o diagnóstico dessa condição. Nas imagens diagnosticadas como catarata, demarca diversas regiões ao redor do cristalino, mas assim como as imagens de olho normal não se segue um padrão,

enquanto, nas imagens preditas como retinopatia diabética, foca nas regiões dos vasos sanguíneos, além de identificar o disco óptico como região negativa em algumas imagens.

Por outro lado, ao aplicar o Grad-CAM, observamos padrões distintos de destaque. Nas imagens preditas como normais, o mapa de calor abrange quase toda a imagem, sugerindo que toda a região foi relevante para a decisão do modelo. Nas imagens preditas como glaucoma, o mapa de calor concentra-se no disco óptico, região crítica para o diagnóstico dessa condição. Para imagens preditas como catarata, o mapa de calor focaliza-se no cristalino, enquanto, nas imagens preditas como retinopatia diabética, o mapa de calor está distribuído nos vasos sanguíneos.

Com base nas análises realizadas, podemos afirmar que tanto o LIME quanto o Grad-CAM demonstram padrões de comportamento que, em sua maioria, estão alinhados com as expectativas em relação às classificações das imagens oftalmológicas. Essas técnicas revelam regiões que são clinicamente relevantes para o diagnóstico de cada condição, de acordo com o conhecimento médico estabelecido.

Ao destacar áreas como o disco óptico para o glaucoma, o cristalino para a catarata e os vasos sanguíneos para a retinopatia diabética, essas técnicas validam a capacidade dos modelos de reconhecer características distintivas de cada condição oftalmológica. Esses *insights* são fundamentais não apenas para a interpretação das decisões dos modelos de classificação, mas também para fornecer uma visão mais transparente e explicável do processo de classificação das imagens oftalmológicas.

Portanto, é evidente que tanto o LIME quanto o Grad-CAM têm o potencial de fornecer informações valiosas que podem complementar o processo de diagnóstico médico, oferecendo uma visão mais transparente e explicável das predições dos modelos utilizados. Essas técnicas têm o poder de fortalecer a confiança nos resultados dos modelos de classificação, além de facilitar a colaboração entre médicos e algoritmos de inteligência artificial no contexto da oftalmologia.

5. CONCLUSÃO

Este trabalho realizou um estudo comparativo na aplicação das técnicas de explicabilidade LIME e Grad-CAM no desempenho de diferentes arquiteturas de CNNs treinadas na classificação de condições oftalmológicas, como glaucoma, retinopatia diabética e catarata.

Após a análise dos resultados na Seção 4.1, constatou-se que o modelo VGG16 obteve o melhor desempenho, com uma acurácia de 93,17% na base de treino e 87,16% na base de validação. Este resultado destaca a eficácia do modelo ResNet50 na generalização para amostras completamente novas.

Um ponto que vale destacar são as classes normal e glaucoma que, em geral, obtiveram os piores resultados. Um possível motivo da baixa pontuação-F1 da classe normal seria a presença de características visuais do glaucoma nas imagens das demais classes. A classe que teve mais predições erradas em relação à classe normal foi a classe glaucoma, que possui características semelhantes àquelas da classe normal. Como foi observado, classes com certo nível de similaridade apresentam maior dificuldade para a classificação de condições oftalmológicas, gerando certo grau de confusão para o modelo.

As 2 técnicas de interpretabilidade investigadas, embora sigam abordagens diferentes, identificaram as mesmas regiões de interesse nas imagens oftalmológicas. No entanto, há diferenças na forma como essas regiões são apresentadas. Enquanto o LIME destacou as áreas críticas de forma mais sutil, exigindo uma análise mais detalhada para interpretar suas visualizações, o Grad-CAM forneceu representações visuais mais diretas e de fácil compreensão, facilitando a interpretação das regiões influentes para as predições dos modelos. No entanto, é importante observar que tanto o Grad-CAM quanto o LIME possuem suas limitações. O LIME mostrou fragilidades devido à aleatoriedade das perturbações nas amostras de referência, enquanto o Grad-CAM produz mapas de calor mais precisos, mas ainda sujeitos a ajustes e melhorias.

6. REFERÊNCIAS

[1] Abbas, Q. 2017. Glaucoma-Deep: Detection of glaucoma eye disease on retinal fundus images using deep learning. International Journal Of Advanced Computer Science And Applications 8(6): 41–45.

[2] Abràmoff, M. D., Garvin, M. K., & Sonka, M. 2010. Retinal imaging and image analysis. IEEE reviews in biomedical engineering 3: 169–208.

[3] Alam U, Asghar O, Azmi S, Malik RA. 2014. General aspects of diabetes mellitus. Handb Clin Neurol. 126:211-222.

[4] Ali, Naila, Syed Ali Wajid, Nasir Saeed, and Muhammad Daud Khan. 2007. "The relative frequency and risk factors of primary open angle glaucoma and angle closure glaucoma." Pak J Ophthalmol 23(3): 117–121.

[5] Asbell, Penny A. et al. 2005. Age-related cataract. The Lancet 365(9459): 599-609.

[6] Boland, Michael V., Ann-Margret Ervin, David S. Friedman, Henry D. Jampel, Barbara S. Hawkins, Daniela Vollenweider, Yohalakshmi Chelladurai, Darcy Ward, Catalina Suarez Cuervo, and Karen A. Robinson. 2013. "Comparative effectiveness of treatments for open angle glaucoma: a systematic review for the US Preventive Services Task Force." Annals of Internal Medicine 158(4): 271–279.

[7] Chakraborty, S. et al. 2017. Interpretability of deep learning models: A survey of results. In: 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI).

[8] Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 839–847.

[9] Claro, M.L., Veras, R., Santos, L., Frazão, M., Carvalho Filho, A. e Leite, D. 2018. Métodos computacionais para segmentação do disco óptico em imagens de retina: uma revisão. Revista Brasileira de Computação Aplicada. 10, 2 (jul. 2018), 29-43. DOI:https://doi.org/10.5335/rbca.v10i2.7661.

[10] Courrol, Lilia Coronato, and Preto, André Oliveira. 2011. Óptica Geométrica. 1st ed. São Paulo.

[11] Curcio, Christine A., and Kimberly A. Allen. 1990. "Topography of ganglion cells in human retina." Journal of Comparative Neurology 300(1): 5–25.

[12] Divya, L., and Jaison Jacob. 2018. "Performance analysis of glaucoma detection approaches from fundus images." Procedia Computer Science 143: 544–551.

[13] Dlb. 2019. Deep Learning Book. Data Science Academy. Disponível em: http://deeplearningbook.com.br/.

[14] Domingues, Vinícius Oliveira et al. 2016. Catarata senil: uma revisão de literatura. Revista de Medicina e Saúde de Brasília 5: 135-134, 28 mar. 2016.

[15] Duda, R. O.; Hart, P. E. & Stork, D. G. 2000. Pattern Classification (2Nd Edition). Wiley-Interscience, New York, NY, USA.

[16] Frasão, G. 2019. Doenças oculares: quais são, tratamento, diagnóstico e prevenção. Ministério da Saúde. Disponível em: https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/d/doencas-oculares. Acesso em: 13 abr. 2024.

[17] Haleem, M. S., Han, L., Hemert, J. van, & Li, B. 2013. Automatic extraction of retinal features from colour retinal images for glaucoma diagnosis: A review. Computerized Medical Imaging and Graphics 37(7–8): 581 – 596. ISSN 0895-6111. Disponível em: http://www.sciencedirect.com/science/article/pii/S0895611113001468.

[18] Haykin, S. S. 2009. Neural networks and learning machines. Pearson Upper Saddle River, NJ, USA: v. 3.

[19] He, K. et al. 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[20] Heng, J., Liu, J., Xu, Y., Yin, F., Wong, D. W. K., Tan, N.-M., Tao, D., Cheng, C.-Y., Aung, T., Wong, T. Y. 2013. Superpixel classification based optic disc and optic cup segmentation for glaucoma screening. IEEE Transactions on Medical Imaging 32(6): 1019–1032.

[21] Int. 2016. Guidelines for Glaucoma Eye Care.

[22] Krizhevsky, A.; Sutskever, I. & Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In Pereira, F.; Burges, C. J. C.; Bottou, L. & Weinberger, K. Q., editores, Advances in Neural Information Processing Systems 25, pp. 1097--1105. Curran Associates, Inc.

[23] Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11): 2278–2324.

[24] Lopez, H. I. M., Garcia, J. C. S., & Mendez, J. A. D. 2016. Cataract Detection Techniques: A Review. IEEE Latin America Transactions 14. Doi: 10.1109/TLA.2016.7587604.

[25] Magalhães, L. 2019. Olhos. Toda Matéria. Disponível em: https://www.todamateria.com.br/olhos/. Acesso em: 13 abr. 2024.

[26] Molnar, C. 2019. Interpretable Machine Learning: A guide for making black box models explainable. Disponível em: https://christophm.github.io/interpretable-ml-book/. Acesso em: 16 abr. 2024.

[27] Molnar, C. 2022. Interpretable Machine Learning. 2 edition. Disponível em: https://christophm.github.io/interpretable-ml-book. Acesso em: 16 abr. 2024.

[28] Nayak, J., U., R. A., Bhat, P. S., Shetty, N., & Lim, T.-C. 2008. Automated diagnosis of glaucoma using digital fundus images. Journal of Medical Systems 33(5): 337. ISSN 1573-689X.

[29] Noronha, K. P., Acharya, U. R., Nayak, K. P., Martis, R. J., & Bhandary, S. V. 2014. Automated classification of glaucoma stages using higher order cumulant features. Biomedical Signal Processing and Control 10: 174 – 183. ISSN 1746-8094. Disponível em: http://www.sciencedirect.com/science/article/pii/S1746809413001699.

[30] Oliveira CM, Cristovão LM, Ribeiro ML, Abreu JR. 2011. Improved automated screening of diabetic retinopathy. Opthalmologica. 226(4):191-197.

[31] Pascolini, D., & Mariotti, S. P. 2012. Global estimates of visual impairment: 2010. British Journal of Ophthalmology 96: 614-618. Doi: 10.1136/bjophthalmol-2011-300539.

[32] Pinheiro, P. 2019. Exame de fundo de olho - o que é e como é feito. Md. Saúde. Disponível em: https://www.mdsaude.com/oftalmologia/exame-de-fundo-de-olho/. Acesso em: 16 abr. 2024.

[33] Pinto, João Victor dos Santos. 2022. As implicações da pandemia do Covid-19 para economia brasileira no período de 2020 a 2021. Uma abordagem minskyana. Orientação de Simone Silva de Deos. Avaliação de Ana Rosa Ribeiro de Mendonça. Campinas, SP: [s.n.], 2022. TCC. (1 recurso online (35 p.)), il., digital, arquivo PDF. Disponível em: https://hdl.handle.net/20.500.12733/5557. Acesso em: 19 abr. 2024.

[34] Powers, David M. W. 2019. What the f-measure doesn't measure: Features, flaws, fallacies and fixes.

[35] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, Fabrice Meriaudeau, April 24, 2018. Indian Diabetic Retinopathy Image Dataset (IDRiD). IEEE Dataport, doi: https://dx.doi.org/10.21227/H25W98.

[36] Quigley, Harry A., Earl M. Addicks, W. Richard Green, and AE Maumenee. 1981. "Optic nerve damage in human glaucoma: II. The site of injury and susceptibility to damage." Archives of Ophthalmology 99(4): 635–649.

[37] Rawat, W., & Wang, Z. 2017. Deep convolutional neural networks for image classification: A comprehensive review. Neural Computation 29(9): 2352–2449. PMID: 28599112.

[38] Reis, L. P. dos. 2023. Algoritmo de detecção de retinopatia diabética baseado em aprendizado de máquina. Trabalho de Conclusão de Curso (Graduação em Engenharia Elétrica) – Universidade Federal de São Carlos, São Carlos.

[39] Ribeiro, M. T., Singh, S., & Guestrin, C. 2016. Model-agnostic interpretability of machine learning. ArXiv, abs/1606.05386.

[40] Russakovsky, O. et al. 2014. Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115: 211–252.

[41] Segre, L. 2019. Eye anatomy: A closer look at the parts of the eye. All About Vision. Disponível em: https://www.allaboutvision.com/resources/anatomy.htm. Acesso em: 13 abr. 2024.

[42] Selvaraju, R. R., A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. 2016. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. arXiv preprint arXiv:1610.02391.

[43] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 618–626.

[44] Simonyan, K. & Zisserman, A. 2014. Very deep convolutional networks for largescale image recognition. CoRR, abs/1409.1556.

[45] Singh, A., Sengupta, S., & Lakshminarayanan, V. 2020. Explainable deep learning models in medical image analysis. Journal of Imaging 6(6): 52.

[46] Soto-Pedre E, Neve A, Millan S, et al. 2015. Evaluation of automated image analysis software for the detection of diabetic retinopathy to reduce the ophthalmologists' workload. Acta Ophthalmol. 93(1):e52-e56.

[47] Souza V, Araújo L, Silva L, Santos A. 2020. Análise comparativa de redes neurais convolucionais no reconhecimento de cenas, in: XI Computer on the Beach, Balneário Camburiú, SC.

[48] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition.

[49] Vakili, Meysam, Mohammad Ghamsari e Masoumeh Rezaei. 2020. Performance analysis and comparison of machine and deep learning algorithms for iot data classification.

[50] Veras, R. d. M. S. 2014. Detecção e segmentação de estruturas em imagens médicas de retina. Tese (Doutorado) — Programa de Pós-Graduação em Engenharia de Teleinformática, Universidade Federal do Ceará.Universidade Federal do Ceará.

[51] Weinreb, Robert N., and Khaw, Peng Tee. 2004. Primary open-angle glaucoma. Disponível em: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(04)16257-0. Acesso em: 13 abr. 2024.

[52] Xinyang Deng, Qi Liu, Yong Deng, and Sankaran Mahadevan. 2016. An improved method to construct basic probability assignment based on the confusion matrix for classification problem. Information Sciences 340-341:250–261. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2016.01.033.