



UNIVERSIDADE ESTADUAL PAULISTA - UNESP
FACULDADE DE CIÊNCIAS - CAMPUS DE BAURU

Miguel José das Neves

eXAI: Uma Técnica Inovadora de Ensemble com Inteligência Artificial Explicável Aplicada à Visão Computacional em Contextos Críticos

Miguel José das Neves

**eXAI: Uma Técnica Inovadora de Ensemble com
Inteligência Artificial Explicável Aplicada à Visão
Computacional em Contextos Críticos**

Tese apresentada à Universidade Estadual Paulista (UNESP), Faculdade de Ciências, Bauru, para obtenção do título de Doutor em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação desta universidade

Área de Concentração: Computação Aplicada

Orientador: Prof. Dr. Kelton Augusto Pontara da Costa

Coorientador: Prof. Dr. Felipe Rodrigues Perche Mahlow.

N518e

Neves, Miguel Jose das

eXAI : Uma Técnica Inovadora de Ensemble com Inteligência Artificial Explicável Aplicada à Visão Computacional em Contextos Críticos / Miguel Jose das Neves. -- Bauru, 2025
63 f.

Tese (doutorado) - Universidade Estadual Paulista (UNESP),
Faculdade de Ciências, Bauru

Orientador: Kelton Augusto Pontara da Costa

Coorientador: Felipe Rodrigues Perche Mahlow

1. Aprendizagem profunda (Aprendizado do computador). 2.
Inteligência artificial. 3. Redes neurais (Computação). I. Título.

ATA DA DEFESA PÚBLICA DA TESE DE DOUTORADO DE MIGUEL JOSÉ DAS NEVES, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO, DA FACULDADE DE CIÊNCIAS - CÂMPUS DE BAURU.

Aos 24 dias do mês de setembro do ano de 2025, às 9h, por meio de Videoconferência, realizou-se a defesa de TESE DE DOUTORADO de MIGUEL JOSÉ DAS NEVES, intitulada **eXAI: Uma Técnica Inovadora de Ensemble com Inteligência Artificial Explicável Aplicada à Visão Computacional em Contextos Críticos**. A Comissão Examinadora foi constituída pelos seguintes membros: Prof. Dr. KELTON AUGUSTO PONTARA DA COSTA (Orientador(a) - Participação Virtual) do(a) FC / UNESP Bauru SP, Profa. Dra. ROBERTA SPOLON (Participação Virtual) do(a) Departamento de Computação / Faculdade de Ciências/Unesp Bauru, Profa. Dra. SIMONE DAS GRAÇAS DOMINGUES PRADO (Participação Virtual) do(a) Departamento de Computação / Faculdade de Ciências de Bauru, Prof. Dr. ALEX MARINO GONÇALVES DE ALMEIDA (Participação Virtual) do(a) Computação / Faculdade de Tecnologia de Ourinhos, Prof. Dr. THIAGO JOSÉ LUCAS (Participação Virtual) do(a) Computação / Professor de Ensino Superior na Faculdade de Tecnologia de Ourinhos. Após a exposição pelo doutorando e arguição pelos membros da Comissão Examinadora que participaram do ato, de forma presencial e/ou virtual, o discente recebeu o conceito final APROVADO. Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelo(a) Presidente(a) da Comissão Examinadora.

Documento assinado digitalmente
 **KELTON AUGUSTO PONTARA DA COSTA**
Data: 24/09/2025 09:59:05-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. KELTON AUGUSTO PONTARA DA COSTA

À minha esposa, Kenia Maria C Neves, por seu amor incondicional, paciência infinita e por ser meu porto seguro em todos os momentos. E à minha filha, Juliana Neves, que com seu sorriso iluminou os dias mais desafiadores desta jornada. Vocês são minha maior inspiração.

Aos meus pais, Jacira Batista das Neves e Miguel Sebastião das Neves (in memoriam), por terem me ensinado o valor do esforço e por serem a base de toda a minha jornada. Onde quer que estejam, sei que se orgulham deste momento.

À minha sogra, Angela Maria P Labs, por me acolher como um filho, e ao meu sogro, Pedro Labs (in memoriam), cuja memória guardo com carinho. Agradeço também a todos os meus familiares pelo apoio fundamental.

Ao Prof. Dr. Kelton Augusto Pontara da Costa, que foi muito mais que um orientador. Agradeço por sua mentoria, por sua paciência inestimável e por me guiar com maestria pelos desafios da pesquisa científica. Sou imensamente grato pela oportunidade de aprender com um profissional e ser humano tão exemplar.

Resumo

A crescente complexidade dos modelos de Aprendizado Profundo (*deep learning*), especialmente as Redes Neurais Convolucionais (*CNNs*), gerou um desafio fundamental em aplicações críticas: a falta de transparência e interpretabilidade de suas decisões. Esta tese aborda este problema ao propor um framework metodológico, o *eXAI* (*Ensemble Explainable Artificial Intelligence*), que visa desenvolver sistemas de análise de imagem que aliam alta performance preditiva à interpretabilidade. A metodologia proposta integra de forma sinérgica o *deep learning* com técnicas de Aprendizado de Conjunto (*Ensemble Learning*) e um *ensemble* de métodos de Inteligência Artificial Explicável (*XAI*), combinando o *SHAP* para atribuição em nível de pixel com o *Grad-CAM* para localização de regiões de interesse. A flexibilidade e a eficácia do framework *eXAI* são validadas através de dois estudos de caso em domínios distintos. O primeiro estudo de caso, focado no diagnóstico de Retinopatia Diabética, implementa uma arquitetura de *stacking* que combina as previsões de uma *CNN* com dados clínicos estruturados, sendo o resultado final classificado por um meta-modelo LightGBM (LGBM). Esta abordagem multimodal demonstrou altíssima acurácia, e a análise *XAI* validou a relevância clínica dos fatores de decisão do modelo. O segundo estudo de caso, no domínio da forense digital, aplica uma *CNN* de classificação direta para detectar manipulações de imagem pela técnica de *Seam Carving*. Neste cenário, o *XAI* foi crucial para revelar que o modelo aprendeu a identificar artefatos de distorção estrutural, em vez de se basear no conteúdo semântico da imagem. Os resultados conjuntos demonstram que o framework *eXAI* é uma abordagem robusta e adaptável, capaz de gerar modelos de *IA* confiáveis para contextos críticos. Ao fornecer explicações compreensíveis e de múltiplas perspectivas, a metodologia contribui para aumentar a confiança e a utilidade prática da inteligência artificial tanto na área da saúde quanto na segurança da informação.

Palavras-chave: Inteligência Artificial Explicável *XAI*. Ensemble Learning. Retinopatia Diabética. Seam Carving. *eXAI*.

Abstract

The growing complexity of Deep Learning models, particularly Convolutional Neural Networks (CNNs), presents a fundamental challenge in critical applications: the opacity of their decision-making processes. This thesis addresses this interpretability gap by proposing and validating a methodological framework termed *eXAI* (*Ensemble Explainable Artificial Intelligence*), designed to create image analysis systems that merge high predictive performance with transparency. Conceptually inspired by Ensemble Learning, the framework's novelty lies not in combining predictive models, but in synergistically integrating an ensemble of explainability methods—specifically combining *SHAP* for pixel-level attribution with *Grad-CAM* for region-level localization—to provide a multi-faceted interpretation of a single model's decisions. The framework's flexibility and efficacy are demonstrated through two case studies in distinct domains. The first, focused on Diabetic Retinopathy diagnosis, implements a stacking architecture that integrates deep visual features from a CNN with structured clinical data, classified by a LightGBM (LGBM) meta-model. This multimodal approach achieved outstanding accuracy, while the XAI analysis validated the clinical relevance of its decision factors. The second case study, in digital forensics, applies a direct classification CNN to detect *Seam Carving* image manipulations. Here, XAI was crucial in revealing that the model learned to identify subtle structural distortion artifacts rather than relying on semantic content. The collective results demonstrate that the *eXAI* framework is a robust and adaptable approach for building trustworthy AI models for critical contexts. By delivering comprehensive and understandable explanations, this methodology enhances the confidence in, and practical utility of, artificial intelligence in both healthcare and information security.

Keywords: Explainable Artificial Intelligence (XAI), Ensemble Learning, Diabetic Retinopathy, Seam Carving, eXAI.

Lista de ilustrações

Figura 1	– Fluxo de trabalho geral do framework <i>eXAI</i> proposto, desde a entrada de dados até a análise de explicabilidade. O framework pode ser instanciado com um modelo <i>CNN</i> de classificação direta (Abordagem 1) ou um modelo <i>Ensemble</i> de <i>Stacking</i> (Abordagem 2).	29
Figura 2	– Matrizes de confusão para os três modelos no conjunto de teste de RD: (a) Modelo Híbrido LGBM, (b) Modelo Pure CNN, e (c) Modelo Structured LGBM.	36
Figura 3	– Curvas ROC para os três modelos avaliados no conjunto de teste de RD.	37
Figura 4	– Gráfico de resumo SHAP para o modelo híbrido de RD, mostrando o impacto das <i>características</i> da CNN (e.g., <i>cnn_feat_270</i>) e das <i>características</i> clínicas.	38
Figura 5	– Visualização da Maximização de Ativação para camadas internas da <i>CNN</i> , revelando os padrões visuais aprendidos associados à Retinopatia Diabética.	39
Figura 6	– Visualização SHAP para a CNN, destacando regiões retinianas importantes para um caso negativo (topo) e um caso positivo (base).	40
Figura 7	– Arquitetura da Rede Neural Convolucional customizada para a detecção de <i>Seam Carving</i>	44
Figura 8	– Curvas de acurácia e perda (<i>loss</i>) do modelo <i>CNN</i> durante o treinamento para o estudo de caso de <i>Seam Carving</i>	46
Figura 9	– Matrizes de confusão para o modelo CNN no conjunto de teste. A matriz à esquerda exibe as contagens brutas das predições. A matriz à direita apresenta os resultados normalizados, onde cada linha é normalizada pelo total da classe verdadeira; os valores na diagonal representam, portanto, a revocação (taxa de verdadeiros positivos) para cada classe.	47
Figura 10	– Curva ROC (Receiver Operating Characteristic) para o modelo CNN no conjunto de teste. O alto valor de AUC (0.99) indica um excelente desempenho discriminativo.	47
Figura 11	– Visualizações complementares do framework <i>eXAI</i> . (a) Imagem de entrada forjada. (b) Mapa de calor Grad-CAM, indicando regiões de interesse. (c) Mapa de atribuição SHAP, onde pixels vermelhos contribuíram para a classificação de fraude.	49

Lista de tabelas

Tabela 1 – Compilação da pesquisa bibliográfica.	22
Tabela 2 – Resumo de trabalhos correlatos selecionados ilustrando o estado da arte. . .	27
Tabela 3 – Hiperparâmetros utilizados no estudo de caso de Retinopatia Diabética. . .	35
Tabela 4 – Métricas de desempenho para os modelos avaliados no conjunto de teste de RD. C1 refere-se à classe RD-positivo (minoritária).	35
Tabela 5 – Desempenho do Modelo Híbrido sob Compressão JPEG.	37
Tabela 6 – Hiperparâmetros utilizados no estudo de caso de <i>Seam Carving</i>	45
Tabela 7 – Métricas de desempenho detalhadas para o modelo <i>CNN</i> no conjunto de teste de <i>Seam Carving</i>	46
Tabela 8 – Acurácia do modelo vs. Qualidade da Compressão JPEG.	48
Tabela 9 – Comparação de performance com métodos SOTA para detecção de <i>Seam Carving</i>	48
Tabela 10 – Comparativo das abordagens nos dois estudos de caso.	51

Lista de abreviaturas e siglas

AUC	Area Under the Curve (Área Sob a Curva)
BRSET	Brazilian Multilabel Ophthalmological Dataset
CNN	Rede Neural Convolucional (<i>Convolutional Neural Network</i>)
eXAI	Ensemble Explainable Artificial Intelligence
HbA1c	Hemoglobina Glicada
IA	Inteligência Artificial
LGBM	Light Gradient Boosting Machine
RD	Retinopatia Diabética
RF	Random Forest (Floresta Aleatória)
ROC	Receiver Operating Characteristic
SHAP	Shapley Additive Explanations
VGG	Visual Geometry Group
XAI	Inteligência Artificial Explicável (<i>Explainable Artificial Intelligence</i>)

Sumário

1	INTRODUÇÃO	13
1.1	Problema de Pesquisa e Justificativa	13
1.1.1	Hipótese	14
1.2	Objetivos	15
1.2.1	Objetivo Geral	15
1.2.2	Objetivos Específicos	15
1.3	Estrutura da Tese	16
2	FUNDAMENTAÇÃO TEÓRICA	17
2.1	Aprendizado Profundo para Visão Computacional	17
2.1.1	Redes Neurais Convolucionais (<i>CNNs</i>)	17
2.1.2	Aprendizado por Transferência (<i>Transfer Learning</i>)	18
2.2	Aprendizado de Conjunto (<i>Ensemble Learning</i>)	18
2.2.1	Modelos Baseados em Árvores: <i>Random Forest</i> e <i>Gradient Boosting</i>	18
2.2.2	Empilhamento (<i>Stacking</i>)	19
2.3	Inteligência Artificial Explicável (<i>XAI</i>)	19
2.3.1	SHAP (<i>Shapley Additive Explanations</i>)	19
2.3.2	Grad-CAM (<i>Gradient-weighted Class Activation Mapping</i>)	20
2.3.3	Maximização de Ativação (<i>Activation Maximization</i>)	20
2.4	Domínios de Aplicação	20
2.4.1	Forense Digital e Detecção de <i>Seam Carving</i>	20
2.4.2	Diagnóstico Médico e Retinopatia Diabética	20
3	REVISÃO DE LITERATURA E TRABALHOS CORRELATOS	21
3.1	Metodologia da Revisão Sistemática	21
3.1.1	Bases de Dados e Critérios de Busca	21
3.2	Análise Temática do Estado da Arte	23
3.2.1	Detecção de Adulteração de Imagens com Foco em <i>Seam Carving</i>	23
3.2.2	Diagnóstico Assistido por Computador de Retinopatia Diabética	23
3.2.3	Inteligência Artificial Explicável (<i>XAI</i>) em Visão Computacional	24
3.3	Discussão Detalhada dos Trabalhos Selecionados	24
3.3.1	Ano base 2017	24
3.3.2	Ano base 2018	24
3.3.3	Ano base 2019	25
3.3.4	Ano base 2020	26
3.3.5	Ano base 2021	26

3.4	Síntese da Literatura e Lacuna de Pesquisa	26
4	METODOLOGIA GERAL PROPOSTA	28
4.1	Visão Geral do Framework eXAI	28
4.2	Construção e Preparação do Conjunto de Dados	29
4.3	Arquitetura do Modelo Preditivo	30
4.3.1	Abordagem 1: Modelo de Classificação Direta (CNN)	30
4.3.2	Abordagem 2: Modelo <i>Ensemble</i> de <i>Stacking</i> (CNN + Modelo Baseado em Árvores)	31
4.4	Framework de Explicabilidade eXAI	31
4.4.1	Atribuição de Importância em Nível de Pixel com SHAP	31
4.4.2	Localização de Região de Interesse com Grad-CAM	32
4.4.3	Visualização de Padrões Gerais com Maximização de Ativação	32
4.5	Métricas de Avaliação de Desempenho	32
5	ESTUDO DE CASO 1: DIAGNÓSTICO ASSISTIDO POR COMPUTADOR DE RETINOPATIA DIABÉTICA	33
5.1	Introdução ao Estudo de Caso	33
5.2	Materiais e Métodos Específicos	33
5.2.1	conjunto de dados e Preparação (BRSET)	33
5.2.2	Pré-processamento de Dados de Imagem e Clínicos	34
5.2.3	Arquitetura do Modelo Híbrido (CNN-LGBM Stacking)	34
5.3	Resultados e Análises	35
5.3.1	Avaliação Comparativa de Desempenho	35
5.3.2	Análise de Robustez contra Compressão JPEG	37
5.4	Análise de Explicabilidade (Resultados do XAI)	38
5.4.1	Análise de Importância de características com SHAP	38
5.4.2	Padrões Visuais Aprendidos pela CNN	39
5.4.3	Custo Computacional e Reprodutibilidade	40
5.5	Discussão e Conclusão do Capítulo	40
6	ESTUDO DE CASO 2: DETECÇÃO DE ADULTERAÇÃO DE IMAGENS POR <i>SEAM CARVING</i>	42
6.1	Introdução ao Estudo de Caso	42
6.2	Materiais e Métodos Específicos	42
6.2.1	Construção do conjunto de dados e Preparação	42
6.2.2	Arquitetura e Treinamento do Modelo CNN	43
6.3	Resultados e Análises	45
6.3.1	Análise do Treinamento e Desempenho no Teste	45
6.3.2	Análise de Robustez e Comparação com o Estado da Arte	48

6.3.3	Análise de Explicabilidade (Resultados do eXAI)	48
6.3.3.1	Análise Qualitativa com Grad-CAM e SHAP	48
6.3.3.2	Análise Quantitativa dos Mapas de Explicabilidade	49
6.3.4	Custo Computacional e Reprodutibilidade	50
6.4	Discussão e Conclusão do Capítulo	50
7	DISCUSSÃO GERAL E COMPARATIVA	51
7.1	Análise Comparativa dos Resultados e Abordagens	51
7.2	O Papel Contextual do XAI: Insights Distintos para Problemas Distintos	52
7.3	Implicações e Contribuições Gerais para a Pesquisa	53
7.4	Limitações da Pesquisa e Trabalhos Futuros	53
8	CONCLUSÃO	55
8.1	Limitações e Trabalhos Futuros	55
	REFERÊNCIAS	57

1 Introdução

A análise de imagens através de Inteligência Artificial (IA), especialmente com o uso de técnicas de Aprendizado Profundo (*Deep Learning*), revolucionou inúmeros campos. Modelos baseados em Redes Neurais Convolucionais (*CNNs*) alcançaram desempenho em nível humano, ou até sobre-humano, em tarefas de classificação, detecção e segmentação de imagens. No entanto, o sucesso desses modelos trouxe um desafio intrínseco: a sua complexidade. A natureza de caixa-preta das *CNNs*, onde o processo de tomada de decisão interna é opaco, representa uma barreira significativa para a sua adoção em aplicações críticas, onde a confiança, a segurança e a interpretabilidade são requisitos fundamentais (DOSHI-VELEZ; KIM, 2017a).

Essa lacuna entre desempenho e transparência é particularmente evidente em domínios de alto risco. No diagnóstico médico assistido por computador, por exemplo, um clínico necessita não apenas de uma predição precisa, mas também de uma justificativa que possa validar e na qual possa confiar antes de tomar uma decisão sobre a saúde de um paciente. Similarmente, no campo da segurança da informação e forense digital, um analista precisa entender como um modelo detecta uma imagem adulterada para validar a detecção como evidência e para desenvolver defesas contra novas formas de manipulação.

Para endereçar essa necessidade, o campo da Inteligência Artificial Explicável (*XAI*) emerge com o objetivo de desenvolver métodos e ferramentas para tornar os sistemas de IA mais transparentes, interpretáveis e confiáveis. Paralelamente, técnicas de Aprendizado de Conjunto (*Ensemble Learning*) têm se mostrado eficazes em aumentar a robustez e a acurácia dos modelos, combinando as forças de múltiplos classificadores (DIETTERICH, 2000).

Esta tese se posiciona na interseção dessas áreas, propondo um framework metodológico flexível, denominado *eXAI* (*Ensemble Explainable Artificial Intelligence*), que integra o poder preditivo do *Deep Learning* com a robustez conceitual do *Ensemble Learning* e a transparência do *XAI*. O objetivo central é criar sistemas de análise de imagem que sejam não apenas precisos, mas cujas decisões possam ser compreendidas e validadas por especialistas humanos em seus respectivos domínios.

1.1 Problema de Pesquisa e Justificativa

O problema central de pesquisa abordado nesta tese pode ser formulado da seguinte maneira:

Como desenvolver um framework de inteligência artificial flexível que possa ser aplicado a diferentes problemas de classificação de imagens em domínios críticos,

garantindo simultaneamente alta acurácia preditiva e um alto grau de interpretabilidade para fomentar a confiança e a utilidade prática por parte de especialistas?

A justificativa para esta pesquisa reside na crescente demanda por sistemas de IA responsáveis e confiáveis. Modelos que apenas fornecem uma saída binária (e.g., doença presente ou imagem adulterada) sem qualquer explicação são de utilidade limitada em contextos onde as consequências de um erro são graves. A capacidade de um modelo explicar por que ele chegou a uma conclusão é o que pode viabilizar sua transição de uma ferramenta de pesquisa para uma ferramenta de trabalho no mundo real, aumentando a colaboração humano-máquina (ARRIETA et al., 2020).

1.1.1 Hipótese

Com base no problema de pesquisa delineado, a hipótese central que norteia este trabalho é a seguinte:

A integração de modelos de *Deep Learning* (CNNs) com classificadores de *Ensemble Learning*, como o *LightGBM* ou o *Random Forest*, em arquiteturas flexíveis, a exemplo do *stacking* para dados multimodais e da classificação direta para dados unimodais, combinada com a aplicação dual de técnicas de *XAI*, permite a criação de sistemas de análise de imagem que não são apenas precisos, mas também transparentes e robustos. Valida-se que tal abordagem é suficientemente geral para ser eficaz em domínios distintos, como o diagnóstico médico e a forense digital, aumentando a confiança do usuário especialista e a utilidade prática da solução ao fornecer *insights* contextualmente relevantes para cada domínio.

Para validar esta hipótese, o framework proposto é aplicado e avaliado em dois estudos de caso representativos:

1. **Diagnóstico de Retinopatia Diabética:** Um problema de diagnóstico médico multimodal, onde imagens de retina são combinadas com dados clínicos do paciente para uma avaliação de risco. Neste cenário, a explicabilidade visa validar o raciocínio do modelo em relação ao conhecimento clínico, com o objetivo de auxiliar e dar confiança ao profissional de saúde.
2. **Deteção de Adulteração de Imagens por *Seam Carving*:** Um problema de forense digital unimodal, focado em detectar manipulações sutis em imagens (AVIDAN; SHAMIR, 2007). Para este caso, a explicabilidade se propõe a revelar a estratégia interna do detector e a identificar os artefatos da manipulação, garantindo que o modelo não utilize atalhos espúrios em sua decisão.

A escolha desses dois domínios permite testar e demonstrar como o framework *eXAI* se adapta a cenários com diferentes características de dados (multimodal vs. unimodal) e diferentes necessidades de explicação (validação clínica vs. revelação de estratégia), oferecendo evidências para a generalidade e eficácia da hipótese.

1.2 Objetivos

Com base no problema de pesquisa delineado, os objetivos desta tese são definidos da seguinte forma:

1.2.1 Objetivo Geral

Desenvolver e validar um framework metodológico de Inteligência Artificial Explicável (*eXAI*) que integre técnicas de *Deep Learning* e *Ensemble Learning* para a classificação de imagens, com foco em fornecer alta acurácia e interpretabilidade em aplicações de domínios críticos.

1.2.2 Objetivos Específicos

- Investigar e modelar uma arquitetura de *ensemble* de *stacking*, utilizando *CNN* e classificadores baseados em árvores, para problemas de classificação de imagens que se beneficiem da integração de dados multimodais.
- Investigar e modelar uma arquitetura de classificação direta com *CNN* para problemas de classificação baseados unicamente em *características* visuais.
- Integrar sistematicamente técnicas de *XAI*, como SHAP e Grad-CAM, e complementarmente a Maximização de Ativação, em ambas as arquiteturas para analisar o processo de decisão dos modelos.
- Aplicar e validar o framework proposto em um estudo de caso de diagnóstico médico de Retinopatia Diabética, avaliando seu desempenho e a relevância clínica de suas explicações.
- Aplicar e validar o framework proposto em um estudo de caso de segurança da informação para detecção de *Seam Carving*, avaliando sua precisão e os *insights* forenses fornecidos pelas explicações.
- Realizar uma análise comparativa dos resultados para discutir a generalidade, flexibilidade e as implicações do framework *eXAI*.

1.3 Estrutura da Tese

Este documento está organizado da seguinte forma para apresentar a pesquisa de maneira clara e lógica:

O **Capítulo 2** apresenta a fundamentação teórica, revisando os conceitos essenciais de *Deep Learning*, *Ensemble Learning*, Inteligência Artificial Explicável, e os domínios de aplicação (Retinopatia Diabética e *Seam Carving*).

O **Capítulo 3** analisa o estado da arte nas áreas de aplicação, contextualizando a pesquisa e identificando a lacuna na literatura que este trabalho se propõe a preencher.

O **Capítulo 4** detalha a metodologia geral do framework *eXAI* proposto, descrevendo seus componentes, arquiteturas de modelo e protocolo de avaliação e explicação.

O **Capítulo 5** apresenta o primeiro estudo de caso, detalhando a aplicação do framework para o diagnóstico de Retinopatia Diabética, incluindo a metodologia específica, os resultados e a discussão.

O **Capítulo 6** apresenta o segundo estudo de caso, focado na detecção de adulteração de imagens por *Seam Carving*, seguindo a mesma estrutura de apresentação.

O **Capítulo 7** realiza uma discussão geral e comparativa dos achados de ambos os estudos de caso, sintetizando as contribuições e as implicações do trabalho.

Finalmente, o **Capítulo 8** resume as conclusões da tese, reforça as contribuições e aponta direções para trabalhos futuros.

2 Fundamentação Teórica

Este capítulo apresenta os fundamentos teóricos que sustentam a pesquisa desenvolvida nesta tese. O objetivo é fornecer ao leitor o embasamento necessário para a compreensão da metodologia proposta e dos estudos de caso subsequentes. São abordados os conceitos essenciais de Aprendizado Profundo para Visão Computacional, com foco em Redes Neurais Convolucionais; as estratégias de Aprendizado de Conjunto (*Ensemble Learning*), incluindo *Random Forest* e *Stacking*; e os princípios e técnicas de XAI. Adicionalmente, serão contextualizados os domínios de aplicação explorados, estabelecendo o alicerce conceitual sobre o qual a contribuição prática desta pesquisa foi construída.

2.1 Aprendizado Profundo para Visão Computacional

O Aprendizado Profundo (*Deep Learning*) é um subcampo do aprendizado de máquina que utiliza redes neurais artificiais com múltiplas camadas (profundas) para modelar e resolver problemas complexos (RUSSELL; NORVIG, 2004). Essas redes são capazes de aprender representações hierárquicas de dados, o que lhes permite extrair e reconhecer padrões e relacionamentos complexos, de forma análoga ao cérebro humano. Nos últimos anos, o *Deep Learning* tornou-se proeminente devido à crescente disponibilidade de grandes volumes de dados e ao avanço do poder computacional, alcançando desempenho de ponta em tarefas como classificação de imagens, detecção de objetos e tradução de idiomas.

2.1.1 Redes Neurais Convolucionais (CNNs)

Uma *CNN* é uma classe de rede neural profunda, inspirada no córtex visual humano, e representa a arquitetura predominante para tarefas de visão computacional (LECUN et al., 1998). Sua eficácia reside na capacidade de aprender automaticamente um conjunto de filtros (*kernels*) que detectam hierarquias de características visuais a partir dos dados de entrada.

A arquitetura de uma *CNN* é tipicamente composta por uma sequência de camadas, sendo as principais:

- **Camada de Convolução:** Aplica um conjunto de filtros sobre a imagem de entrada para criar mapas de características (*feature maps*). Cada filtro é especializado em detectar um padrão específico, como bordas, texturas ou formas.
- **Função de Ativação:** Após a convolução, uma função de ativação não-linear é aplicada a cada elemento do mapa de características. A mais comum é a Unidade Linear Retificada

(ReLU - *Rectified Linear Unit*), que introduz não-linearidade no modelo, permitindo o aprendizado de padrões mais complexos (NAIR; HINTON, 2010).

- **Camada de Agrupamento (*Pooling*):** Reduz a dimensionalidade espacial dos mapas de características, tornando a representação mais compacta e invariante a pequenas translações. A técnica mais comum é o *Max Pooling*, que seleciona o valor máximo de uma vizinhança.
- **Camada Totalmente Conectada (*Fully Connected*):** Após várias camadas de convolução e *pooling*, os mapas de características são achatados (*flattened*) em um vetor e passados para uma ou mais camadas densas, que realizam a classificação final com base nas características extraídas.

2.1.2 Aprendizado por Transferência (*Transfer Learning*)

O treinamento de uma CNN profunda do zero requer um volume massivo de dados rotulados e alto custo computacional. Uma estratégia eficaz para contornar essa limitação é o aprendizado por transferência (*transfer learning*). Esta técnica consiste em utilizar uma rede neural já treinada em um grande conjunto de dados de referência, como a ImageNet, e adaptá-la para uma nova tarefa específica. A intuição é que as camadas iniciais da rede pré-treinada aprenderam a detectar características visuais genéricas (bordas, texturas, cores) que são úteis para uma vasta gama de problemas de visão computacional.

Nesta abordagem, a base convolucional da rede pré-treinada é geralmente congelada (seus pesos não são atualizados), e apenas as camadas de classificação no topo são substituídas e treinadas com os dados da nova tarefa. Modelos como a *VGGNet* (SIMONYAN; ZISSERMAN, 2015) são comumente utilizados como extratores de *características* de base neste processo, servindo como um ponto de partida poderoso para o desenvolvimento de modelos em domínios com dados mais limitados, como a área médica.

2.2 Aprendizado de Conjunto (*Ensemble Learning*)

O *Ensemble Learning* é uma técnica de aprendizado de máquina onde múltiplos modelos individuais são combinados para produzir uma predição final. A premissa fundamental é que, ao combinar as opiniões de vários modelos, o sistema resultante pode alcançar maior acurácia, robustez e capacidade de generalização do que qualquer modelo único conseguiria isoladamente (DIETTERICH, 2000).

2.2.1 Modelos Baseados em Árvores: *Random Forest* e *Gradient Boosting*

Uma árvore de decisão é um modelo preditivo que mapeia observações sobre um item para conclusões sobre seu valor alvo. Para evitar o sobreajuste de uma única árvore, métodos

de *ensemble* são frequentemente utilizados. O *Random Forest*, proposto por Breiman (2001), constrói uma infinidade de árvores de decisão em subconjuntos de dados e de *características*, e a predição final é dada pela votação majoritária. Outra abordagem poderosa é o *Gradient Boosting*, onde modelos são adicionados sequencialmente para corrigir os erros dos modelos anteriores. O **LightGBM (LGBM)**, utilizado nesta tese, é uma implementação altamente eficiente de *gradient boosting* (KE et al., 2017).

2.2.2 Empilhamento (*Stacking*)

O empilhamento, ou *Stacking*, é uma técnica de *ensemble* que visa aprender a melhor forma de combinar as predições de múltiplos modelos. A arquitetura consiste em dois níveis:

- **Nível 0 (Modelos Base):** Vários modelos diferentes são treinados no conjunto de dados de treinamento. As predições geradas por eles são coletadas.
- **Nível 1 (Meta-Classificador):** Um novo modelo (o meta-classificador) é treinado usando as predições dos modelos base como suas *características* de entrada, aprendendo a melhor forma de combiná-las.

Na abordagem híbrida desta tese (Estudo de Caso 1), a CNN atua como um modelo base de Nível 0, e um classificador baseado em árvores de decisão, como o **LightGBM**, atua como o meta-classificador de Nível 1.

2.3 Inteligência Artificial Explicável (XAI)

XAI é uma área de pesquisa focada no desenvolvimento de sistemas de IA que sejam mais transparentes, interpretáveis e confiáveis. A natureza de "caixa-preta" de modelos complexos como os de *Deep Learning* é uma barreira para sua adoção em aplicações de alto risco, como saúde e forense, onde as consequências de uma decisão errada são graves (DOSHI-VELEZ; KIM, 2017b). O objetivo do XAI é desenvolver métodos para que os sistemas de IA possam explicar suas decisões de forma compreensível.

2.3.1 SHAP (*Shapley Additive Explanations*)

O *SHAP* é uma abordagem unificada para explicar a saída de qualquer modelo de aprendizado de máquina, introduzida por Lundberg e Lee (2017). Baseado nos valores de *Shapley* da teoria dos jogos, o *SHAP* atribui a cada *feature* um valor de importância para uma predição particular, quantificando seu impacto. O método garante propriedades desejáveis como precisão local e consistência. Existem implementações otimizadas como o *TreeExplainer* para modelos baseados em árvores e o *GradientExplainer* para redes neurais.

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

O Grad-CAM é uma técnica de visualização que produz "mapas de calor" para identificar as regiões de uma imagem de entrada que foram mais importantes para uma determinada predição. Ele utiliza os gradientes que fluem para a última camada convolucional da *CNN* para entender a importância de cada *feature map*. O resultado é um mapa de ativação que destaca visualmente "onde" o modelo está olhando para tomar sua decisão, sendo uma ferramenta poderosa para a localização de evidências visuais.

2.3.3 Maximização de Ativação (*Activation Maximization*)

A Maximização de Ativação é uma técnica de visualização que ajuda a entender o que os filtros de uma *CNN* aprenderam a detectar de forma geral. A técnica gera uma imagem sintética que maximiza a ativação de um neurônio ou filtro específico, partindo de uma imagem de ruído (ERHAN et al., 2009). A imagem resultante representa o padrão visual "ideal" para aquele filtro, fornecendo uma visão direta dos conceitos, como texturas e formas, que a rede aprendeu durante o treinamento.

2.4 Domínios de Aplicação

Para demonstrar a validade e a versatilidade do *framework* proposto, ele foi implementado e avaliado em dois domínios de aplicação distintos. A seleção desses cenários permite uma análise comparativa rica sobre como a metodologia se adapta a diferentes tipos de dados e necessidades de interpretabilidade.

2.4.1 Forense Digital e Detecção de *Seam Carving*

O *Seam Carving* é uma técnica de redimensionamento de imagens ciente do conteúdo, introduzida por Avidan e Shamir (2007). O algoritmo remove ou insere caminhos de pixels de baixa energia (*seams*), permitindo redimensionar a imagem enquanto preserva o conteúdo importante. No entanto, ele introduz artefatos estruturais sutis, cuja detecção é um problema relevante em forense digital.

2.4.2 Diagnóstico Médico e Retinopatia Diabética

A Retinopatia Diabética (RD) é uma complicação microvascular do *diabetes mellitus* que afeta os vasos sanguíneos da retina, sendo uma das principais causas de cegueira em adultos (YAU et al., 2012). O diagnóstico precoce é realizado através da análise de retinografias em busca de lesões características. Fatores de risco clínicos, como a idade do paciente e o tempo de diagnóstico do diabetes, são cruciais para a avaliação da progressão da doença.

3 Revisão de Literatura e Trabalhos Correlatos

Este capítulo apresenta uma revisão da literatura científica que fundamenta esta tese, analisando o estado da arte nas áreas de detecção de adulteração de imagens, diagnóstico de Retinopatia Diabética e Inteligência Artificial Explicável (XAI). A análise dos trabalhos existentes é crucial para contextualizar a pesquisa, identificar as tendências atuais e, mais importante, apontar a lacuna na literatura que este trabalho se propõe a preencher.

Para garantir o rigor e a reprodutibilidade, foi adotada uma metodologia de Revisão Sistemática, conforme detalhado na seção a seguir. As seções subsequentes apresentam a análise temática dos trabalhos encontrados, uma discussão detalhada das contribuições mais relevantes e a síntese final que define o escopo desta tese.

3.1 Metodologia da Revisão Sistemática

Este processo foi desenvolvido por meio do estudo de obras correlatas a fim de possibilitar embasamento para esta pesquisa. Em seu trabalho, Kitchenham e Charters (2007) enfatizam a necessidade de definir critérios para realizar uma busca de forma confiável, permitindo que outros pesquisadores possam reproduzi-la. Este procedimento visa justificar a pesquisa, fornecendo uma base de dados explorada de forma confiável através de protocolos definidos (BIOLCHINI et al., 2005).

3.1.1 Bases de Dados e Critérios de Busca

A pesquisa foi conduzida nas principais bases de publicações científicas da área da computação: *IEEE Xplore*, *ACM Digital Library*, *Science Direct*, *SpringerLink*, *MDPI* e *Taylor & Francis*.

A busca foi estruturada para abranger os três conceitos centrais da tese: *Seam Carving*, *Deep Learning* e *XAI*. A Tabela 1 detalha os parâmetros e os resultados quantitativos da busca inicial, que retornou 145 trabalhos. A baixa sobreposição entre os três conceitos evidenciou o caráter inédito da pesquisa.

Tabela 1 – Compilação da pesquisa bibliográfica.

Base de Dados	Parâmetros de Busca	Semântica	Período	Trabalhos
IEEE e ACM	<i>Seam Carving</i> OR <i>detection seam carving</i> AND <i>machine learning</i>	Palavra-chave e Título	2016–2021	58
Science Direct	(Mesmos acima)	Palavra-chave e Título	2016–2021	6
SpringerLink	(Mesmos acima)	Palavra-chave e Título	2016–2021	26
MDPI	(Mesmos acima)	Palavra-chave e Título	2016–2021	13
Taylor & Francis	(Mesmos acima)	Palavra-chave e Título	2016–2021	2
IEEE e ACM	<i>XAI</i> AND <i>Deep Learning</i>	Título e Palavra-chave	2016–2021	9
Science Direct	<i>XAI</i> AND <i>Deep Learning</i>	Título e Palavra-chave	2016–2021	20
SpringerLink	<i>XAI</i> AND <i>Deep Learning</i>	Título e Palavra-chave	2016–2021	8
MDPI	<i>XAI</i> AND <i>Deep Learning</i>	Título e Palavra-chave	2016–2021	3
Taylor & Francis	<i>XAI</i> AND <i>Deep Learning</i>	Título e Palavra-chave	2016–2021	0

Fonte: Elaborado pelo autor (2025).

Após a busca, foi realizada a leitura e a filtragem dos artigos com base nos seguintes critérios de exclusão:

- O trabalho não se referia a pelo menos um dos conceitos de interesse ou não estabelecia nenhuma relação com os outros.
- O trabalho não discutia ou mostrava alguma aplicação de *Deep Learning*, base algorítmica desta tese.

Ao final do processo, 40 artigos foram selecionados por sua relevância e contribuição.

3.2 Análise Temática do Estado da Arte

3.2.1 Detecção de Adulteração de Imagens com Foco em *Seam Carving*

A detecção de manipulações do tipo *seam carving* é um problema desafiador em forense digital, pois os artefatos introduzidos são frequentemente sutis e estruturais. A literatura inicial sobre o tema focava em métodos de processamento de imagens e aprendizado de máquina clássico. Trabalhos como o de Chen e Grauman (2018) exploraram o uso de classificadores como *SVM* (*Support Vector Machine*) e *ELM* (*Extreme Learning Machine*), alimentados por *características* extraídas manualmente a partir de *patches* da imagem, como as baseadas no operador Sobel ou em matrizes de probabilidade. Embora tenham alcançado algum sucesso, esses métodos dependem fortemente da qualidade da engenharia de *características* e podem não generalizar bem para diferentes tipos de imagem.

Com o advento do *Deep Learning*, a abordagem predominante migrou para o uso de Redes Neurais Convolucionais (*CNNs*). A principal vantagem das *CNNs* é sua capacidade de aprender automaticamente as *características* relevantes diretamente dos dados. Trabalhos pioneiros, como o de Cieslak, Costa e Papa (2018), demonstraram a viabilidade de usar *CNNs* para esta tarefa. Pesquisas subsequentes, como a de Nam et al. (2019a), focaram em projetar arquiteturas de *CNN* específicas para capturar os artefatos de alta frequência e as distorções estatísticas deixadas pela remoção ou inserção de *seams*, frequentemente superando os métodos clássicos em precisão e robustez.

3.2.2 Diagnóstico Assistido por Computador de Retinopatia Diabética

O uso de *IA* para o diagnóstico de Retinopatia Diabética (RD) é uma das áreas mais maduras da oftalmologia computacional. A vasta maioria dos trabalhos no estado da arte utiliza *CNNs* para classificar a presença e a severidade da RD a partir de imagens de retinografia (SENAPATI et al., 2024). A evolução nesse campo seguiu a tendência geral do *Deep Learning*, passando de arquiteturas mais simples para modelos mais profundos e complexos, como *ResNet*, e mais recentemente, explorando arquiteturas baseadas em atenção e *transformers* (NAZIH et al., 2023).

Uma vertente particularmente relevante para esta tese é a de modelos multimodais. Reconhecendo que o diagnóstico médico raramente se baseia em uma única fonte de informação, pesquisadores começaram a desenvolver modelos híbridos que combinam dados de imagem com dados clínicos estruturados. O trabalho de Tan, Lim e Ting (2023), por exemplo, demonstrou que a integração de variáveis como idade, duração do diabetes e níveis de *HbA1c* com as predições de uma *CNN* melhora significativamente o desempenho do classificador de RD. Essa abordagem sinérgica é um pilar central do primeiro estudo de caso desta tese.

3.2.3 Inteligência Artificial Explicável (XAI) em Visão Computacional

À medida que os modelos de *Deep Learning* se tornaram mais precisos, a preocupação com sua natureza opaca cresceu, impulsionando o campo do XAI. O objetivo do XAI é fornecer transparência e interpretabilidade às decisões dos modelos de IA. Na visão computacional, os métodos de XAI podem ser amplamente categorizados.

Trabalhos pioneiros focaram em métodos baseados em gradientes, como os mapas de saliência, que destacam os pixels de uma imagem de entrada que mais influenciam a saída. Técnicas mais avançadas, como o *Grad-CAM*, melhoraram essa abordagem ao produzir mapas de calor mais localizados e semanticamente significativos.

Outra classe de métodos, na qual esta tese se baseia, foca na atribuição de importância de *características*. O SHAP (*Shapley Additive Explanations*) (LUNDBERG; LEE, 2017) se destaca por sua base teórica sólida na teoria dos jogos e por sua flexibilidade. Trabalhos como os de Quellec et al. (2021) e Niu et al. (2022) demonstram a aplicação de diferentes técnicas de XAI, incluindo sistemas baseados em casos (CBR), para fornecer explicações em contextos de imagens médicas, ressaltando a importância da interação humano-computador para a confiança no sistema.

3.3 Discussão Detalhada dos Trabalhos Seleccionados

A leitura dos 40 artigos seleccionados trouxe diversas considerações que embasam a pesquisa. A seguir, uma análise cronológica das principais contribuições.

3.3.1 Ano base 2017

O trabalho (NGUYEN; KANKANHALLI, 2017) desenvolve um paralelo de metodologias de comparação de imagens baseadas no mapa de saliência da imagem. A motivação baseia-se no princípio de que se as imagens têm semelhanças, seus mapas de saliência também têm semelhanças. Ainda o trabalho (SASAKI et al., 2017) apresenta uma técnica CNN para extrair uma camada para fornecer parâmetros a um algoritmo SVM para classificar as estruturas de corte de costura.

3.3.2 Ano base 2018

Em (ADADI; BERRADA, 2018) foi apresentado um estudo sobre XAI com 381 artigos, discutindo a importância do tema, mas registrando que a ideia de XAI naquela época ainda era inexplorada e indefinida. O artigo em (WEBER et al., 2018) discute a ideia de uso de XAI com base no conceito de Raciocínio Baseado em Casos (CBR) para categorizar a interpretação de texto. Em (LI et al., 2018) tem-se uma proposta de *Seam Carving* baseada em um algoritmo genético, com desempenho superior aos métodos tradicionais. O objetivo do artigo (CAI;

KONG; WANG, 2018) envolve a aplicação de dois fluxos de rede para otimizar a detecção de cliques de ação em vídeos. O estudo (FURUTA; TSUBAKI; YAMASAKI, 2018) propõe aplicar o algoritmo multipass como alternativa ao *Seam Carving*, focando em ganho computacional. Um trabalho de *Seam Carving* (WEI; CHOU; SU, 2018) propõe um esquema multioperador para determinar as áreas de corte baseado em mapa de saliência. A pesquisa (WEN; SU, 2018) desenvolve uma melhoria no processo de análise estatística de dados de imagem para aprimorar o dimensionamento de objetos. O trabalho (ERTORK; ERTORK, 2018) propõe a utilização do algoritmo KRX para redução de imagens com preservação de anomalias. O algoritmo OCAIR foi apresentado em (ROY; DASGUPTA; PRADHAN, 2018) combinando cortes gráficos iterativos (Grabcut) (ROTHER; KOLMOGOROV; BLAKE, 2004) para obter melhores resultados em velocidade. A pesquisa em Sun et al. (2018) teve como objetivo criar um método de identificação de adulteração de imagem com CNN, superando métodos anteriores. A pesquisa (CHENG et al., 2018) aborda uma aplicação em ELM para melhorar os métodos de detecção de *Seam Carving*, superando o SVM em precisão e velocidade. O trabalho (CHOI; EUN; KIM, 2018) propõe o uso de CNN para identificação de cárie em imagens de radiografias. Um trabalho pioneiro de Cieslak, Costa e Papa (2018) demonstrou a detecção de *Seam Carving* por meio de uma CNN, com resultados satisfatórios. O estudo (CHEN; GRAUMAN, 2018) explora o reconhecimento de imagens para identificar objetos sob interação de pessoas na cena. O problema explorado no estudo (WALIA; KUMAR, 2018) é identificar a adulteração de imagens trabalhadas por *Seam Carving* vertical usando CNN e ResNet v2.

3.3.3 Ano base 2019

O método proposto em (SCHLEGEL et al., 2019) apresenta uma forma de avaliar alguns métodos *XAI* existentes. O trabalho (CHITTAJALLU et al., 2019) apresenta uma proposta de um aplicativo *XAI* que utiliza CNN ResNet para ranquear imagens de similaridade de procedimentos cirúrgicos. Os autores (LAMY et al., 2019) demonstram a construção de um CBR para uso em medicina, como no diagnóstico de câncer de mama. Este trabalho (KEANE; KENNY, 2019) foca em uma revisão de literatura envolvendo RNA e CBR, notando que as aplicações de *XAI* ainda não seguiam um padrão. O artigo (SONG; LEE; LEE, 2019) propõe a utilização da CNN para montar o mapa de energia das imagens para realizar redimensionamento e recorte. O trabalho (IQBAL et al., 2019) considera uma CNN baseada em TensorFlow para permitir que o processamento de *Seam Carving* seja usado em dispositivos portáteis. A pesquisa (JAGTAP; TRIPATHI, 2019) propõe uma abordagem de RNA para redimensionamento de imagens com preservação do campo semântico. Este artigo (NAM et al., 2019b) apresenta uma proposta para identificar o *Seam Carving* usando uma CNN, com resultados superiores a métodos anteriores. O método proposto em (WEI; CHENG; CHANG, 2019) utiliza uma rede Hopfield para identificar se as imagens foram esculpidas com costuras e detectar essas costuras. Uma abordagem para detectar *Seam Carving* usando CNN é proposta em (YE et al., 2019), com resultados melhores que abordagens anteriores baseadas em SVM.

3.3.4 Ano base 2020

Este trabalho (DAVIS et al., 2020) propõe que a confiança em sistemas de *IA* deve ser medida pela utilidade da explicação do modelo. O artigo (FUTIA; VETRO, 2020) apresenta a proposta de utilizar *Knowledge Graphs* como uma ferramenta *XAI* para tornar o entendimento de algoritmos mais compreensível. O trabalho (WANG; NIU; WANG, 2021) baseia-se no uso da CNN para identificar adulterações de imagens causadas por métodos de pintura. A pesquisa (MYERS-DEAN; WEHRWEIN, 2020) explora o uso de CNN baseado em distâncias de pixel para combinar *Seam Carving* semântico com escala de mapa RGB. A CNN é explorada em (NAM et al., 2020) para identificar o *Seam Carving* de três maneiras (original, inserido e removido) através de uma rede customizada, a ILFNet. É proposto um trabalho com a CNN em (NAZARI; AKGÜN, 2020) que fornece quatro classes de resultados em entalhe por retalho de costura (0%, 10%, 20% e 40%).

3.3.5 Ano base 2021

Um estudo *XAI* (KENNY et al., 2021) considera os usuários e suas reações ao entender as explicações, propondo um método cognitivo baseado em modelo mental. O aplicativo (YEOM et al., 2019) demonstra uma técnica de remoção em camadas de CNN usando o método de visualização LRP. A pesquisa (BAKALO; GOLDBERGER; BEN-ARI, 2021) propõe um método de duas redes profundas para a tarefa de classificação do câncer de mama. A revisão em (JAM et al., 2021) destaca o uso de métodos de *Deep Learning* para detectar pinturas em imagens, mas aponta a falta de informações para reprodutibilidade nos trabalhos. Este trabalho (JOUIS et al., 2021) realizou testes em modelos *XAI* de caixa branca e caixa preta (CNN), destacando a importância da avaliação do usuário. Este estudo (BIACH et al., 2021) demonstra um método codificador/decodificador (Fals-Unet) para identificar alterações em imagens, baseado na arquitetura Resnet50. O trabalho (ZHAO, 2021) é desenvolvido para juntar várias imagens em uma foto panorâmica no celular, empregando uma técnica de costura semelhante ao *Seam Carving*.

3.4 Síntese da Literatura e Lacuna de Pesquisa

A análise da literatura revela que, embora haja um corpo de trabalho substancial em cada uma das áreas supracitadas, a intersecção entre elas é limitada. Existem modelos para detecção de *seam carving*, modelos para diagnóstico de *RD* e uma variedade de técnicas de *XAI*. No entanto, a revisão sistemática indicou que poucos trabalhos realizam uma aplicação e, principalmente, uma análise comparativa de um mesmo framework metodológico de *XAI* em domínios de aplicação tão distintos.

A Tabela 2 resume alguns dos trabalhos selecionados que formam o contexto desta pesquisa.

Tabela 2 – Resumo de trabalhos correlatos selecionados ilustrando o estado da arte.

Referência	Domínio/Foco	Técnica Principal e Contribuição
Cieslak, Costa e Papa (2018)	Detecção de <i>Seam Carving</i>	Demonstra o uso de <i>CNNs</i> para a tarefa.
Cheng et al. (2018)	Detecção de <i>Seam Carving</i>	Utiliza <i>ELM</i> e <i>SVM</i> com <i>características</i> extraídas manualmente.
Tan, Lim e Ting (2023)	Retinopatia Diabética	Propõe modelo multimodal (<i>CNN</i> + Dados Clínicos) para predição de risco.
Quellec et al. (2021)	<i>XAI</i> em Imagem Médica	Desenvolve modelo <i>XAI</i> customizado (<i>ExplAIIn</i>) para explicar classificação de severidade.

Fonte: Elaborado pelo autor (2025).

A lacuna de pesquisa que esta tese visa preencher é, portanto, a seguinte: falta um estudo que não apenas proponha um *framework* flexível (o *eXAI*) capaz de se adaptar a problemas unimodais e multimodais, mas que também compare os tipos de *insights* que as mesmas técnicas de *XAI* (*SHAP* e Maximização de Ativação) podem gerar em contextos diferentes. Ou seja, como a explicação de um modelo de segurança da informação difere da explicação de um modelo de diagnóstico médico? A resposta a essa pergunta constitui a principal contribuição original desta tese ao campo do *XAI*.

4 Metodologia Geral Proposta

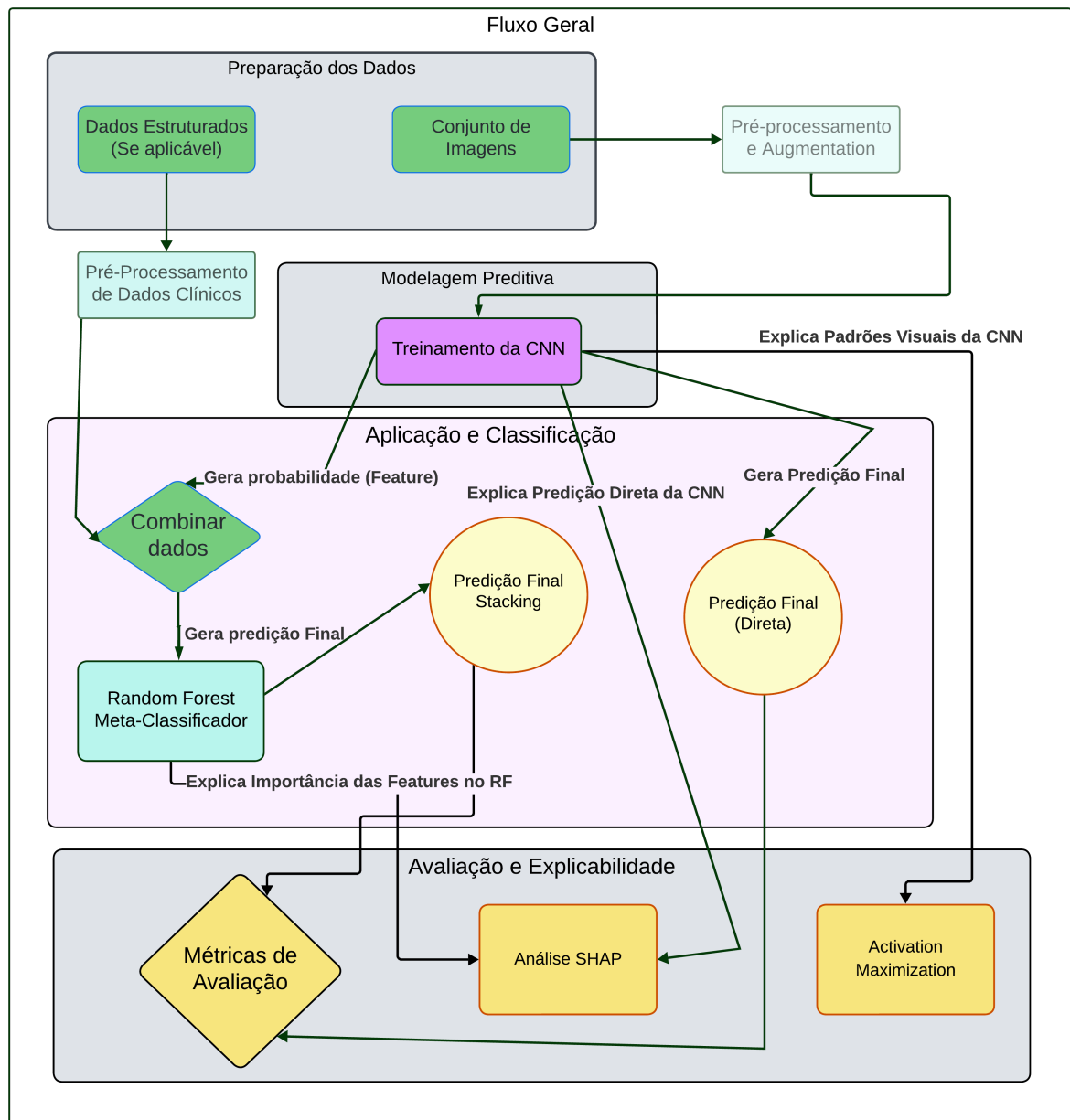
Este capítulo detalha a metodologia geral proposta nesta tese para o desenvolvimento e avaliação de sistemas de Inteligência Artificial Explicável (*XAI*) aplicados à classificação de imagens em contextos críticos. O framework aqui descrito, denominado *eXAI*, é projetado para ser flexível e aplicável a diferentes domínios, conforme demonstrado nos estudos de caso subsequentes sobre o diagnóstico médico de Retinopatia Diabética (Capítulo 5) e a detecção de adulteração de imagens por *Seam Carving* (Capítulo 6).

A abordagem central combina o poder de representação das *CNNs* com a robustez de técnicas de *Ensemble Learning* e a transparência de métodos de *XAI*. O objetivo é criar modelos que não apenas alcancem alta acurácia, mas que também forneçam *insights* compreensíveis sobre seu processo de tomada de decisão, aumentando a confiança e a utilidade em aplicações do mundo real.

4.1 Visão Geral do Framework *eXAI*

O fluxo de trabalho do framework *eXAI* é modular e compreende quatro etapas principais, sendo elas: (1) Construção e Preparação do Conjunto de Dados, (2) Arquitetura do Modelo Preditivo, (3) Protocolo de Avaliação de Desempenho, e (4) Análise de Explicabilidade. A Figura 1 ilustra a inter-relação dessas etapas e as duas principais instâncias do framework exploradas nesta tese.

Figura 1 – Fluxo de trabalho geral do framework eXAI/ proposto, desde a entrada de dados até a análise de explicabilidade. O framework pode ser instanciado com um modelo *CNN* de classificação direta (Abordagem 1) ou um modelo *Ensemble* de *Stacking* (Abordagem 2).



Fonte: Elaborado pelo autor (2025).

As seções a seguir detalham cada um desses componentes metodológicos.

4.2 Construção e Preparação do Conjunto de Dados

A base de qualquer modelo de aprendizado de máquina é um conjunto de dados bem preparado e representativo do problema. Os passos gerais para esta etapa, aplicados em ambos

os estudos de caso, incluem os itens a seguir.

- **Coleta e Definição dos Dados:** Consiste na obtenção das imagens e dos metadados associados, quando existentes. A tarefa de classificação é definida, sendo que nesta tese ambas as tarefas são binárias.
- **Pré-processamento de Imagens:** Inclui o redimensionamento das imagens para uma dimensão uniforme, como por exemplo 224×224 ou 256×256 pixels, e a normalização dos valores dos pixels para o intervalo $[0, 1]$. O pré-processamento de dados clínicos, quando aplicável, inclui etapas de limpeza, padronização (*z-score*) e imputação de valores ausentes.
- **Aumento de Dados (*Data Augmentation*):** Para o conjunto de treinamento, técnicas de aumento de dados, tais como rotações, zooms e espelhamentos aleatórios, são aplicadas para aumentar a diversidade dos dados e reduzir o risco de *overfitting* (SHORTEN; KHOSHGOFTAAR, 2019).
- **Tratamento de Desbalanceamento de Classes:** Em cenários com classes desbalanceadas, como no estudo de caso de RD, técnicas de reamostragem são aplicadas exclusivamente ao conjunto de treinamento para mitigar o viés do modelo. Estas podem incluir o sobre-amostragem (*oversampling*) da classe minoritária e/ou a sub-amostragem (*undersampling*) da classe majoritária (BUDA; MAKI; MAZUROWSKI, 2018).
- **Divisão dos Dados:** O conjunto de dados é dividido em subconjuntos de treinamento, validação e teste. É crucial que o conjunto de teste não sofra reamostragem para que a avaliação do modelo reflita seu desempenho em um cenário com a distribuição de classes original.

4.3 Arquitetura do Modelo Preditivo

O framework *eXAI* explora duas arquiteturas principais de modelo preditivo, demonstrando sua flexibilidade para problemas unimodais e multimodais.

4.3.1 Abordagem 1: Modelo de Classificação Direta (CNN)

Nesta abordagem, uma *CNN* é treinada para realizar a tarefa de classificação de ponta a ponta. A arquitetura, inspirada em modelos como a *VGGNet* (SIMONYAN; ZISSERMAN, 2015), tipicamente consiste em blocos convolucionais para extração de *características* e uma cabeça de classificação com camadas densas, que culmina em uma camada de saída Sigmóide para predição de probabilidade. Esta abordagem é ideal para problemas unimodais, onde a decisão se baseia unicamente em informações visuais, e é aplicada no estudo de caso de detecção de *Seam Carving* (Capítulo 6).

4.3.2 Abordagem 2: Modelo *Ensemble* de *Stacking* (CNN + Modelo Baseado em Árvores)

Esta abordagem híbrida combina as forças do *Deep Learning* e do aprendizado de máquina clássico em uma arquitetura de *stacking* de dois níveis (DIETTERICH, 2000), sendo particularmente adequada para problemas multimodais. A estrutura consiste em dois níveis:

1. **Nível 0 - Extrator de *características* de Imagem:** Uma *CNN* pré-treinada, como a VGG16, é utilizada como uma extratora de *características* de alto nível. Utiliza-se a saída de uma de suas últimas camadas, a exemplo da camada de *Global Average Pooling*, como um vetor de *características* (ou *embedding*) que resume o conteúdo visual da imagem.
2. **Nível 1 - Meta-Classificador:** Um classificador baseado em árvores de decisão, como o *LightGBM* (LGBM) (KE et al., 2017) ou o *Random Forest* (RF) (BREIMAN, 2001), é treinado usando um conjunto de *características* combinado. Este conjunto inclui o vetor de *características* da imagem extraído pela *CNN* e os dados clínicos ou outros metadados estruturados relevantes para o problema.

O meta-classificador aprende a ponderar as *características* visuais com o contexto fornecido pelos dados estruturados para realizar a classificação final. Esta abordagem é aplicada no estudo de caso de diagnóstico de Retinopatia Diabética (Capítulo 5).

4.4 Framework de Explicabilidade eXAI

Um pilar central desta tese é a garantia de que os modelos não operem como caixas-pretas. Para isso, o framework eXAI incorpora uma abordagem sinérgica de explicabilidade, combinando técnicas de XAI que oferecem perspectivas complementares sobre a decisão do modelo. A principal combinação utilizada é a de SHAP com Grad-CAM.

4.4.1 Atribuição de Importância em Nível de Pixel com SHAP

Para quantificar a contribuição de cada *feature* para uma predição, utiliza-se a técnica *SHAP* (*Shapley Additive Explanations*) (LUNDBERG; LEE, 2017). A metodologia SHAP é flexível e pode ser aplicada a ambas as arquiteturas. No modelo de classificação direta (*CNN*), o *GradientExplainer* é utilizado para aproximar os valores SHAP para os pixels da imagem, respondendo à pergunta: "*quais pixels específicos influenciaram a decisão?*". Para o modelo de *stacking*, o *TreeExplainer* calcula os valores SHAP exatos para o meta-classificador, revelando a importância de cada *feature* de entrada, seja ela visual (do vetor da *CNN*) ou clínica.

4.4.2 Localização de Região de Interesse com Grad-CAM

Enquanto o SHAP oferece uma visão granular, o *Grad-CAM* (*Gradient-weighted Class Activation Mapping*) fornece uma perspectiva de mais alto nível, respondendo à pergunta: "em qual região da imagem o modelo focou?". A técnica gera mapas de calor que destacam as áreas da imagem mais influentes para uma predição, utilizando os gradientes da última camada convolucional. Essa abordagem é ideal para entender o foco espacial do modelo, identificando as regiões que ele considerou mais relevantes para sua análise. A combinação de Grad-CAM e SHAP permite uma interpretação mais completa: o Grad-CAM mostra "onde" olhar, e o SHAP detalha "o que" ver naquela região.

4.4.3 Visualização de Padrões Gerais com Maximização de Ativação

Como técnica complementar, emprega-se a Maximização de Ativação (*Activation Maximization*) (ERHAN et al., 2009). Diferente de explicar uma predição específica, seu objetivo é entender o que um filtro da CNN aprendeu de forma geral. Ela gera imagens sintéticas que maximizam a ativação de filtros específicos, oferecendo uma visualização direta dos conceitos visuais (texturas, formas) que a rede foi treinada para detectar.

4.5 Métricas de Avaliação de Desempenho

Para avaliar rigorosamente o desempenho dos modelos, um protocolo padronizado é adotado, focado no conjunto de teste que mantém a distribuição de classes original. As seguintes métricas são utilizadas:

- **Matriz de Confusão:** Detalhando Verdadeiros Positivos (TP), Falsos Positivos (FP), Verdadeiros Negativos (TN) e Falsos Negativos (FN).
- **Acurácia (*acurácia*):** $\frac{TP+TN}{TP+TN+FP+FN}$.
- **Precisão (*precisão*):** $\frac{TP}{TP+FP}$.
- **Revocação (*Recall* ou *Sensibilidade*):** $\frac{TP}{TP+FN}$.
- **F1-Score:** $2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}}$.
- **Curva ROC (*Receiver Operating Characteristic*) e AUC (*Area Under the Curve*):** Mede a capacidade geral do modelo de distinguir entre as classes.

Adicionalmente, em cenários com classes desbalanceadas, como o diagnóstico de Retinopatia Diabética, o limiar de decisão padrão de 0.5 pode não ser o ideal para equilibrar as métricas de precisão e Recall. Nesses casos, o limiar pode ser otimizado no conjunto de validação para maximizar uma métrica de interesse, como o F1-score, garantindo que o modelo final seja mais bem ajustado às necessidades específicas do problema.

5 Estudo de Caso 1: Diagnóstico Assistido por Computador de Retinopatia Diabética

5.1 Introdução ao Estudo de Caso

A Retinopatia Diabética (RD) é uma das principais complicações do *diabetes mellitus* e uma das maiores causas de cegueira evitável em adultos em todo o mundo (YAU et al., 2012). O diagnóstico precoce, realizado através da análise de imagens de fundo de olho (retinografias), é crucial para mitigar a progressão da doença. No entanto, a análise manual dessas imagens é um processo demorado e subjetivo, representando um desafio em cenários com grande volume de pacientes para triagem (SENAPATI et al., 2024).

Este capítulo apresenta a primeira aplicação prática do framework *eXAI*, detalhado no Capítulo 4, a um problema crítico no domínio do diagnóstico médico. O objetivo é desenvolver e avaliar um sistema de inteligência artificial preciso e, fundamentalmente, interpretável para a classificação binária de RD (presença vs. ausência da doença), utilizando dados de imagem e informações clínicas.

Devido à natureza multimodal do problema, onde o contexto do paciente agrega grande valor diagnóstico, este estudo de caso implementa a Abordagem 2 da metodologia: um Modelo *Ensemble* de *Stacking*. A abordagem visa integrar sinergicamente *características* visuais extraídas por uma *CNN* com dados clínicos estruturados, utilizando um meta-classificador LightGBM para aumentar a acurácia, a robustez e a confiança no diagnóstico.

5.2 Materiais e Métodos Específicos

A implementação seguiu o protocolo geral do framework *eXAI*. As especificidades desta aplicação, incluindo uma etapa crítica de curadoria de dados, são detalhadas a seguir.

5.2.1 conjunto de dados e Preparação (BRSET)

O conjunto de dados utilizado foi o *Brazilian Multilabel Ophthalmological conjunto de dados* (BRSET) (NAKAYAMA et al., 2024), que compreende aproximadamente 16.266 imagens e dados clínicos associados.

Um passo metodológico fundamental foi a curadoria e filtragem por qualidade. O conjunto de dados original foi filtrado para reter apenas as imagens onde todos os parâmetros de qualidade (foco, iluminação, campo da imagem e artefatos) foram classificados como satisfatórios (classe 1). Este processo, embora tenha reduzido o tamanho do conjunto de dados,

foi crucial para minimizar o ruído e garantir que o modelo fosse treinado em padrões visuais clinicamente relevantes e claros.

O conjunto de dados curado foi então dividido em conjuntos de treinamento (70%), validação (15%) e teste (15%) por amostragem estratificada, preservando a distribuição natural de classes em cada subconjunto.

5.2.2 Pré-processamento de Dados de Imagem e Clínicos

As imagens foram redimensionadas para 224×224 pixels e pré-processadas utilizando a função específica da arquitetura VGG16.

Para os dados clínicos estruturados, um conjunto de *características* candidatas foi selecionado. Um modelo preliminar de *Random Forest* com a técnica `SelectFromModel` foi utilizado para identificar as variáveis mais impactantes. As *características* selecionadas, como `patient_age` e `diabetes_time_y`, foram então limpas, padronizadas (usando `StandardScaler` ajustado apenas nos dados de treino) e tiveram valores ausentes imputados com a mediana do conjunto de treinamento.

5.2.3 Arquitetura do Modelo Híbrido (CNN-LGBM Stacking)

Conforme a Abordagem 2 (Seção 4.3.2), a arquitetura de *stacking* foi implementada em dois níveis:

- **Nível 0 - Extrator de características (VGG16):** Uma *CNN* com arquitetura VGG16, pré-treinada na ImageNet, foi utilizada estritamente como extratora de *características*. A base convolucional foi mantida "congelada", e a saída foi extraída da camada `GlobalAveragePooling2D`, gerando um vetor de 512 dimensões para cada imagem.
- **Nível 1 - Meta-Classificador (LightGBM):** Um classificador LightGBM (LGBM) foi treinado para realizar a predição final. A entrada para este modelo consistiu na concatenação do vetor de 512 *características* da *CNN* com as *características* clínicas selecionadas e pré-processadas.

Para fins de comparação, outros dois modelos foram treinados: um modelo "Pure CNN"(utilizando apenas as imagens) e um modelo "Structured LGBM"(utilizando apenas os dados clínicos). A seleção dos hiperparâmetros para todos os modelos seguiu uma abordagem metodológica, partindo de valores consolidados na literatura e realizando ajustes empíricos. Parâmetros como o otimizador Adam para a CNN e a função de perda binária foram escolhidos por serem padrões em suas respectivas áreas. Outros valores, como a taxa de aprendizado e o número de estimadores do LGBM, foram ajustados com base no monitoramento do desempenho no conjunto de validação, visando o melhor equilíbrio entre performance e prevenção do

sobreajuste. Os principais hiperparâmetros utilizados nos modelos estão detalhados na Tabela 3.

Tabela 3 – Hiperparâmetros utilizados no estudo de caso de Retinopatia Diabética.

Parâmetro	Valor
<i>CNN (VGG16 Feature Extractor)</i>	
Modelo Base	VGG16 (pré-treinada)
Camadas Treináveis	Nenhuma (base congelada)
Camada de Saída	GlobalAveragePooling2D (512 <i>características</i>)
<i>LightGBM (Meta-Classificador)</i>	
Número de Estimadores	300
Taxa de Aprendizagem	0.05
Objetivo	Binary
Peso das Classes	Balanced

Fonte: Elaborado pelo autor (2025).

5.3 Resultados e Análises

Esta seção apresenta os resultados quantitativos da avaliação dos modelos. Inicialmente, é realizada uma análise comparativa de desempenho no conjunto de teste para demonstrar a eficácia da abordagem híbrida frente a modelos de modalidade única. Em seguida, avalia-se a robustez do modelo proposto a variações na qualidade da imagem, simulando condições de uso em cenários clínicos reais.

5.3.1 Avaliação Comparativa de Desempenho

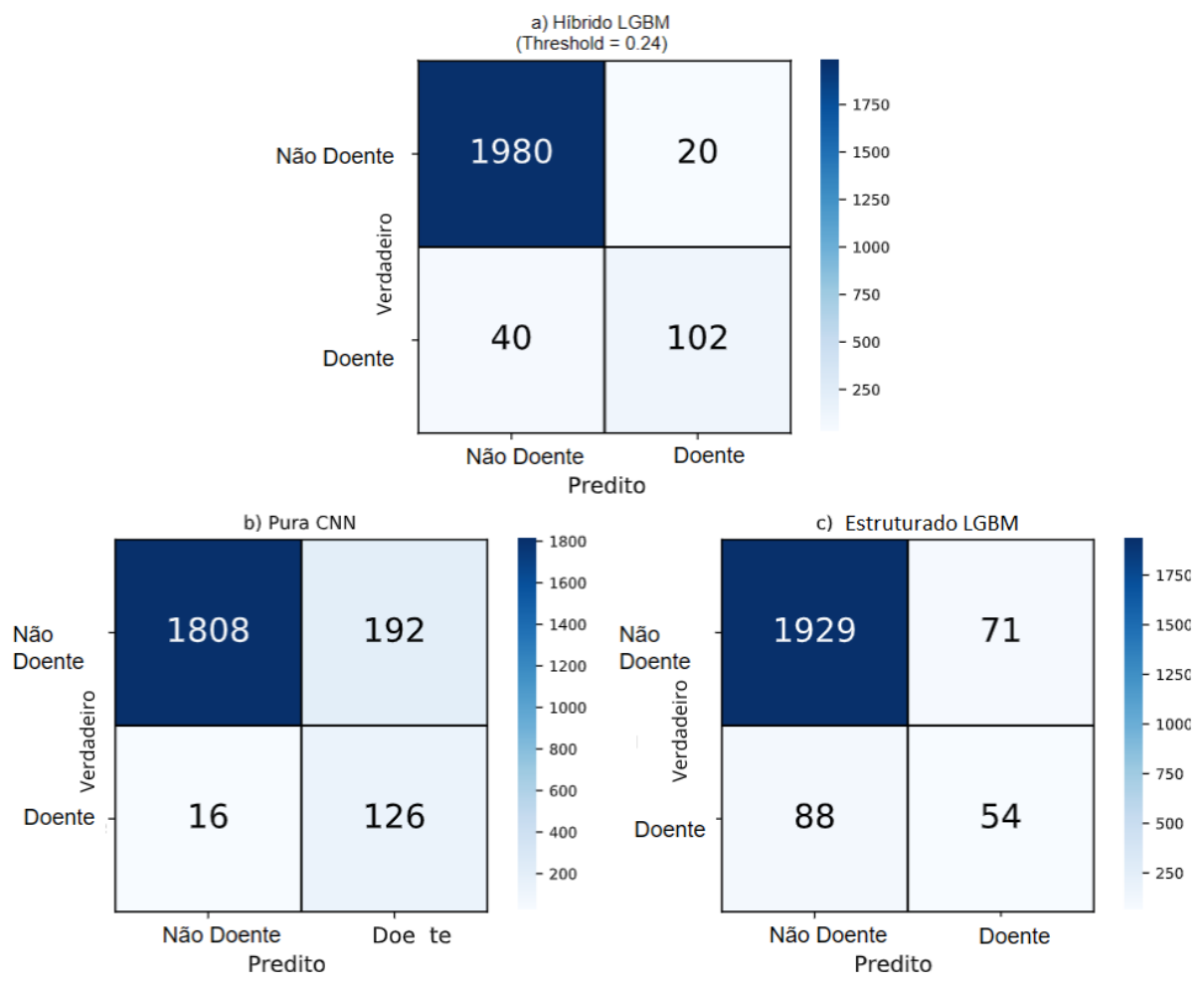
A avaliação dos três modelos no conjunto de teste demonstrou a superioridade da abordagem híbrida. Conforme a Tabela 4, o modelo Híbrido LGBM alcançou um AUC de 0.967 e um F1-score balanceado de 0.77 para a classe positiva (após otimização do limiar de decisão), superando significativamente os modelos de modalidade única.

Tabela 4 – Métricas de desempenho para os modelos avaliados no conjunto de teste de RD. C1 refere-se à classe RD-positivo (minoritária).

Modelo	Acurácia	Precisão (C1)	Revocação (C1)	F1-score (C1)	AUC
Pure CNN	90.3%	0.40	0.89	0.55	0.951
Structured LGBM	92.4%	0.43	0.37	0.40	0.833
Hybrid LGBM	97.0%	0.84	0.72	0.77	0.967

Fonte: Elaborado pelo autor (2025).

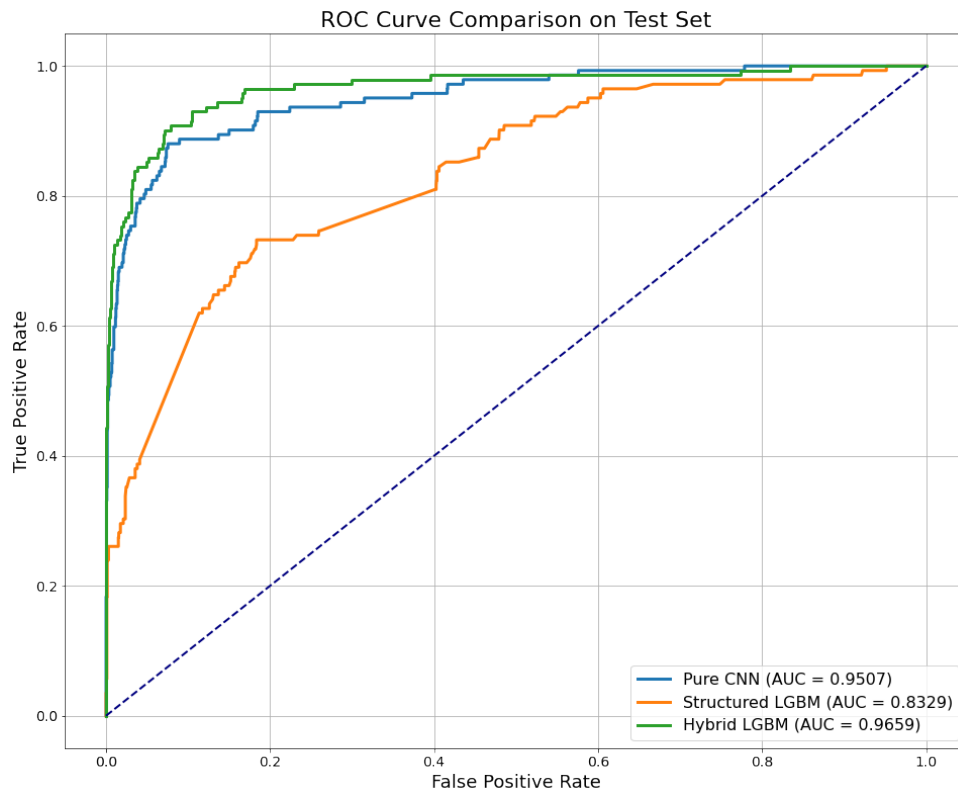
Figura 2 – Matrizes de confusão para os três modelos no conjunto de teste de RD: (a) Modelo Híbrido LGBM, (b) Modelo Pure CNN, e (c) Modelo Structured LGBM.



Fonte: Elaborado pelo autor (2025).

As matrizes de confusão (Figura 2) e as curvas ROC (Figura 3) ilustram essa superioridade. O modelo Híbrido (a) alcançou um excelente equilíbrio entre sensibilidade e especificidade, com um número reduzido de falsos negativos (40) em comparação com o modelo clínico (88), o que é crítico em um cenário de triagem médica. As curvas ROC confirmam visualmente a maior capacidade discriminativa do modelo combinado.

Figura 3 – Curvas ROC para os três modelos avaliados no conjunto de teste de RD.



Fonte: Elaborado pelo autor (2025).

5.3.2 Análise de Robustez contra Compressão JPEG

Para simular um cenário real onde a qualidade da imagem pode variar, foi conduzida uma análise de robustez. As imagens do conjunto de teste foram re-comprimadas em diferentes níveis de qualidade JPEG (90, 70 e 50), e o modelo Híbrido foi reavaliado sem qualquer retreinamento. Os resultados (Tabela 5) mostram que o modelo é altamente robusto, com apenas uma pequena queda no desempenho mesmo em níveis de compressão significativos, mantendo um AUC acima de 0.955.

Tabela 5 – Desempenho do Modelo Híbrido sob Compressão JPEG.

Qualidade JPEG	AUC	F1-score (RD)
Original (100)	0.967	0.77
90	0.965	0.76
70	0.961	0.75
50	0.955	0.74

Fonte: Elaborado pelo autor (2025).

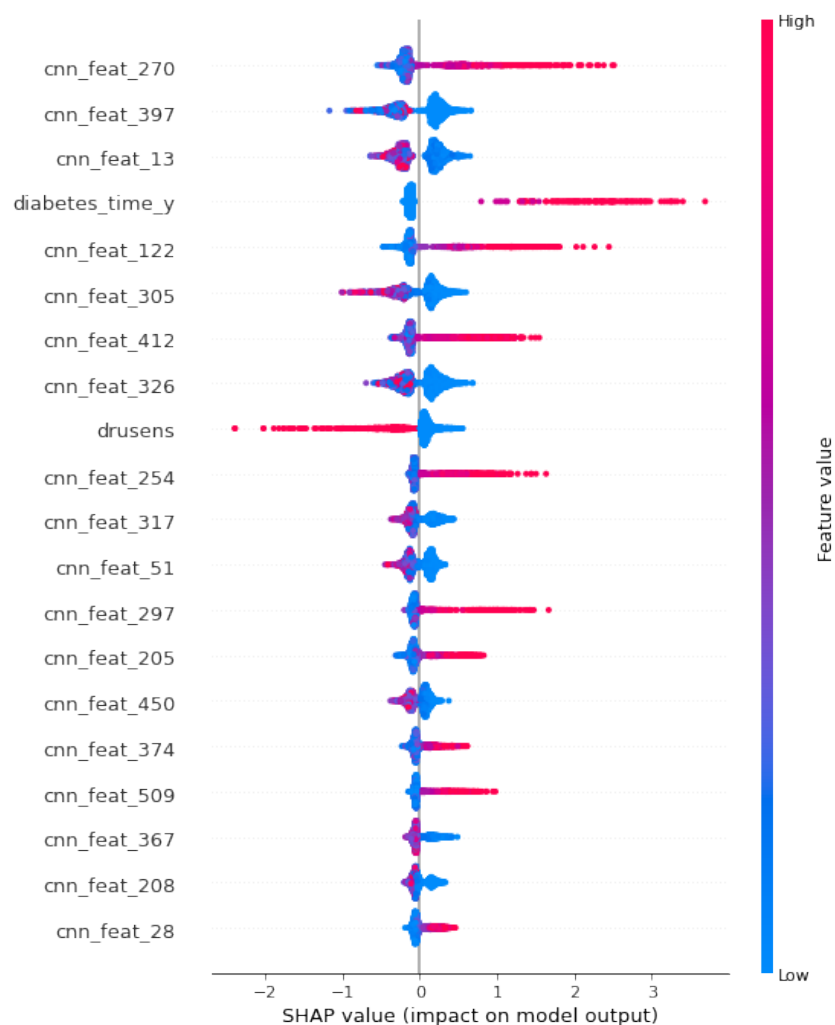
5.4 Análise de Explicabilidade (Resultados do XAI)

Para além das métricas de desempenho, esta seção investiga o processo decisório do modelo através do framework *eXAI*. A análise objetiva validar a coerência clínica das decisões do sistema, primeiro quantificando a importância de cada *feature* de entrada com a técnica SHAP. Em seguida, são explorados os padrões visuais que a *CNN* aprendeu a associar com a patologia, a fim de fornecer evidências visuais para o diagnóstico.

5.4.1 Análise de Importância de características com SHAP

A análise SHAP foi aplicada ao meta-classificador LGBM para compreender os fatores que influenciam suas previsões. O gráfico de resumo (Figura 4) ilustra a importância global das *características* de entrada.

Figura 4 – Gráfico de resumo SHAP para o modelo híbrido de RD, mostrando o impacto das *características* da CNN (e.g., *cnn_feat_270*) e das *características* clínicas.



Fonte: Elaborado pelo autor (2025).

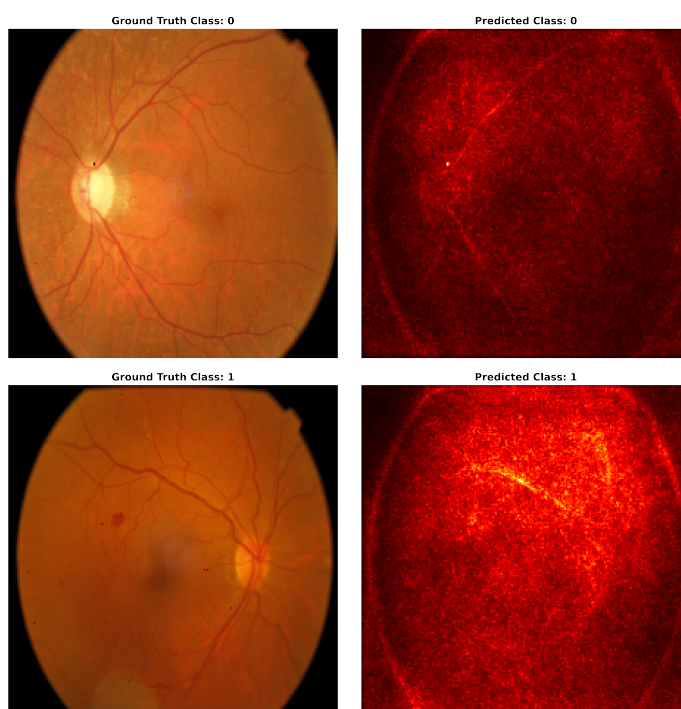
Observa-se que as *características* extraídas pela *CNN* (prefixo *cnn_feat_*) são, de longe, as mais influentes, confirmando o poder da informação visual. No entanto, variáveis clínicas como *diabetes_time_y* e a presença de drusens aparecem com destaque entre os preditores mais importantes, demonstrando seu valor complementar. Altos valores dessas *características* clínicas (indicados pela cor vermelha) consistentemente empurram a predição para a classe positiva (RD), alinhando-se com o conhecimento médico estabelecido.

Para validar quantitativamente essa observação, foi calculada a correlação de Spearman entre os valores absolutos do SHAP para as *características* de imagem e as principais variáveis de risco clínico. Foi encontrada uma correlação positiva e significativa com o tempo de diabetes (*diabetes_time_y*), indicando que o modelo depende mais das evidências visuais em pacientes com maior tempo de doença, um comportamento clinicamente coerente.

5.4.2 Padrões Visuais Aprendidos pela CNN

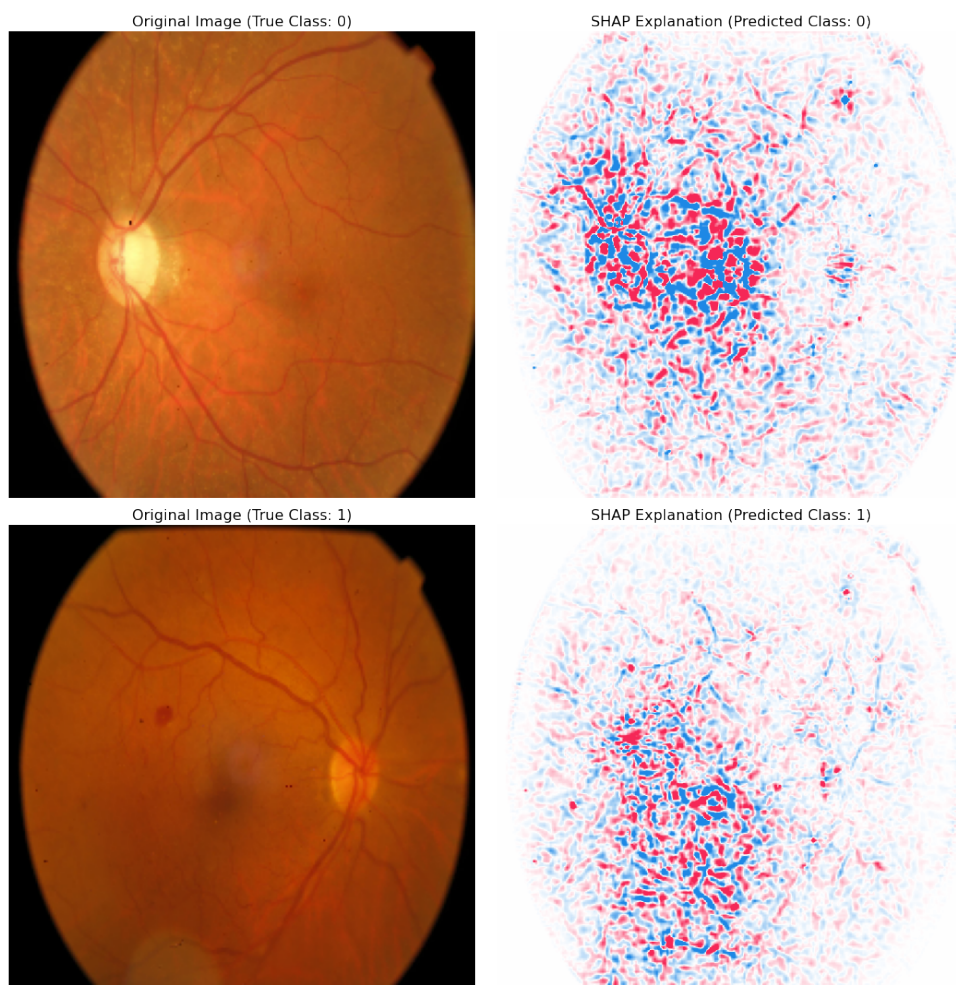
A técnica de Maximização de Ativação, aplicada à *CNN*, revelou que os filtros da rede aprenderam a identificar padrões visualmente consistentes com lesões de RD, como microaneurismas e exsudatos (Figura 5). Adicionalmente, visualizações SHAP em nível de imagem (Figura 6) destacam as regiões retinianas específicas que o modelo considerou mais importantes para cada diagnóstico individual, fornecendo uma base para o diálogo entre o sistema de IA e o médico.

Figura 5 – Visualização da Maximização de Ativação para camadas internas da *CNN*, revelando os padrões visuais aprendidos associados à Retinopatia Diabética.



Fonte: Elaborado pelo autor (2025).

Figura 6 – Visualização SHAP para a CNN, destacando regiões retinianas importantes para um caso negativo (topo) e um caso positivo (base).



Fonte: Elaborado pelo autor (2025).

5.4.3 Custo Computacional e Reprodutibilidade

O treinamento do extrator de *características* e do meta-classificador LGBM foi conduzido no ambiente de hardware descrito. O processo completo de treinamento levou aproximadamente 4 horas. Para garantir a total reprodutibilidade deste estudo de caso, o código-fonte, incluindo o *notebook* com o *fluxo de trabalho* completo, está publicamente disponível em um repositório no GitHub: <<https://github.com/migmiguel29/SeamCarvingeXai>>.

5.5 Discussão e Conclusão do Capítulo

Este estudo de caso validou com sucesso a aplicação da Abordagem 2 do framework eXAI. O modelo híbrido demonstrou desempenho de ponta ao combinar efetivamente a extração de *características* por *deep learning* com o contexto clínico do paciente. A etapa de curadoria

dos dados foi fundamental para garantir que o modelo fosse treinado com sinais visuais claros, o que se refletiu no alto desempenho.

A superioridade do modelo combinado, especialmente na métrica F1-score e no AUC, destaca a importância da abordagem multimodal. A análise XAI, tanto qualitativa quanto quantitativa, reforçou a validade clínica do modelo. O SHAP confirmou que o modelo toma decisões com base em fatores logicamente consistentes: a avaliação da imagem pela *CNN* é o fator mais importante, mas é modulada por fatores de risco clínicos bem conhecidos. Isso não apenas aumenta a confiança no modelo, mas também fornece uma base para o diálogo entre o sistema de IA e o especialista. A robustez demonstrada contra a compressão JPEG solidifica ainda mais sua potencial aplicabilidade em cenários do mundo real.

Os achados deste capítulo servem como uma forte evidência da eficácia da metodologia proposta. O próximo capítulo explorará a aplicação deste mesmo framework a um domínio distinto para testar sua flexibilidade.

6 Estudo de Caso 2: Detecção de Adulteração de Imagens por *Seam Carving*

6.1 Introdução ao Estudo de Caso

No campo da segurança da informação e forense digital, a detecção de manipulações em imagens é um desafio crescente. Uma técnica sofisticada de adulteração é o *seam carving*, um algoritmo de redimensionamento ciente do conteúdo que remove ou insere costuras de pixels de baixa energia, preservando as partes importantes da imagem e produzindo alterações muitas vezes imperceptíveis ao olho humano (AVIDAN; SHAMIR, 2007). Embora eficaz, este processo introduz artefatos estruturais sutis. A capacidade de detectar tais manipulações de forma confiável é, portanto, fundamental para garantir a integridade e a autenticidade de evidências digitais.

Modelos de *deep learning*, em particular as Redes Neurais Convolucionais (*CNNs*), demonstraram grande potencial na detecção dessas fraudes. Contudo, a sua natureza de "caixa-preta" limita a confiança em seu uso em domínios de alto risco como a forense, onde entender *como* uma decisão é tomada é tão importante quanto a decisão em si.

Para endereçar essa lacuna de interpretabilidade, este capítulo apresenta a segunda aplicação do framework *eXAI*, detalhado no Capítulo 4. O objetivo é desenvolver e validar um sistema que não apenas classifica imagens manipuladas por *seam carving* com alta acurácia, mas que também fornece uma compreensão profunda e multifacetada de seu processo decisório. A novidade desta abordagem, inspirada conceitualmente em *Ensemble Learning*, não reside na combinação de múltiplos modelos preditivos, mas sim na integração de um *ensemble* de técnicas de explicabilidade. Especificamente, combinamos o SHAP para atribuição de importância em nível de pixel com o Grad-CAM para localização de regiões de interesse, gerando uma interpretação mais holística e robusta das decisões de uma única *CNN*.

Para este problema, que se baseia unicamente em artefatos visuais, foi implementada a Abordagem 1 da metodologia (Modelo de Classificação Direta), conforme descrito na Seção 4.3.1, por não haver a necessidade de dados estruturados adicionais.

6.2 Materiais e Métodos Específicos

6.2.1 Construção do conjunto de dados e Preparação

Para criar um ambiente controlado, foi construído um *conjunto de dados* específico para o estudo. Utilizando o *Seam Carving JPEG Image Forgery conjunto de dados* (??),

derivado de coleções de imagens amplamente utilizadas em forense digital (LIU, 2019; LIU, 2017; CELEBI; HSU; LIU, 2022), foi gerado um conjunto de dados com um total de 10.300 imagens, perfeitamente balanceado entre as duas classes de interesse:

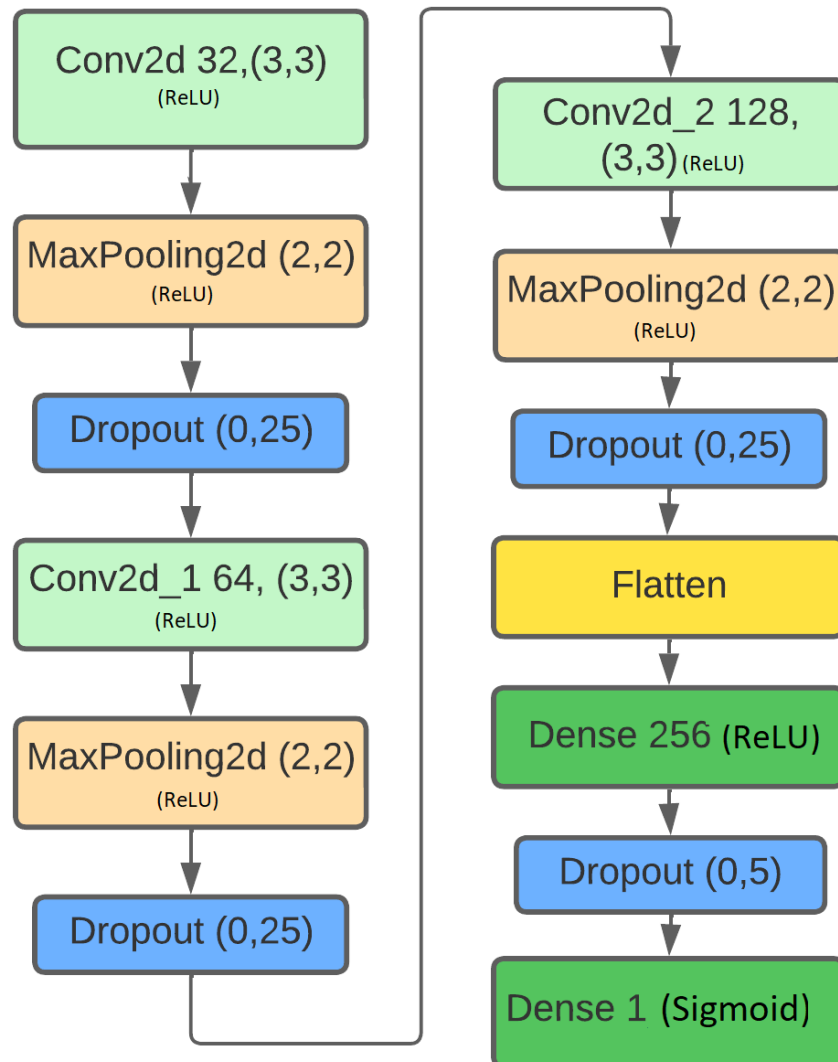
- **Uncompressed (classe 0):** 5.150 imagens originais, não adulteradas.
- **Seam Carving (classe 1):** 5.150 imagens correspondentes, manipuladas através da aplicação do algoritmo de *seam carving* (AVIDAN; SHAMIR, 2007).

O pré-processamento incluiu o redimensionamento das imagens para 256×256 pixels e a normalização dos valores dos pixels para o intervalo $[0, 1]$. O conjunto de dados foi dividido de forma estratificada em 75% para treinamento (7.725 imagens), 12.5% para validação (1.288 imagens) e 12.5% para teste (1.288 imagens). O aumento de dados em tempo real (*data augmentation*) foi aplicado exclusivamente ao conjunto de treinamento para mitigar o sobreajuste e melhorar a generalização do modelo.

6.2.2 Arquitetura e Treinamento do Modelo CNN

Foi projetada uma *CNN* customizada para a tarefa de classificação binária, utilizando a API Keras com *backend* TensorFlow. A detecção de artefatos de *seam carving* depende da identificação de perturbações estatísticas de baixo nível, que são não-semânticas por natureza. Por essa razão, uma arquitetura treinada do zero é mais apropriada do que modelos pré-treinados em conjunto de dados de objetos, como o ImageNet, pois evita o viés inerente a *características* de alto nível.

Figura 7 – Arquitetura da Rede Neural Convolucional customizada para a detecção de *Seam Carving*.



Fonte: Elaborado pelo autor (2025).

A arquitetura, detalhada na Figura 7, é composta por três blocos convolucionais seguidos por uma cabeça de classificação. A definição dos hiperparâmetros para o treinamento seguiu uma abordagem metodológica. Parâmetros como o otimizador Adam e a função de perda de entropia cruzada binária foram escolhidos por serem amplamente adotados e eficazes em tarefas de classificação de imagens. Outros valores, como a taxa de aprendizado, o número de épocas e as taxas de *dropout*, foram ajustados empiricamente, monitorando a convergência e o desempenho no conjunto de validação para evitar o sobreajuste. Os hiperparâmetros finais utilizados estão resumidos na Tabela 6.

Tabela 6 – Hiperparâmetros utilizados no estudo de caso de *Seam Carving*.

Parâmetro	Valor
<i>Arquitetura e Treinamento da CNN</i>	
Otimizador	Adam
Taxa de Aprendizagem	1e-4
Função de Perda	Binary Cross-entropy
Épocas Máximas	100
Tamanho do <i>Batch</i>	32
<i>Early Stopping</i> (Paciência)	10
Taxas de <i>Dropout</i>	0.25 (blocos conv.), 0.5 (densa)

Fonte: Elaborado pelo autor (2025).

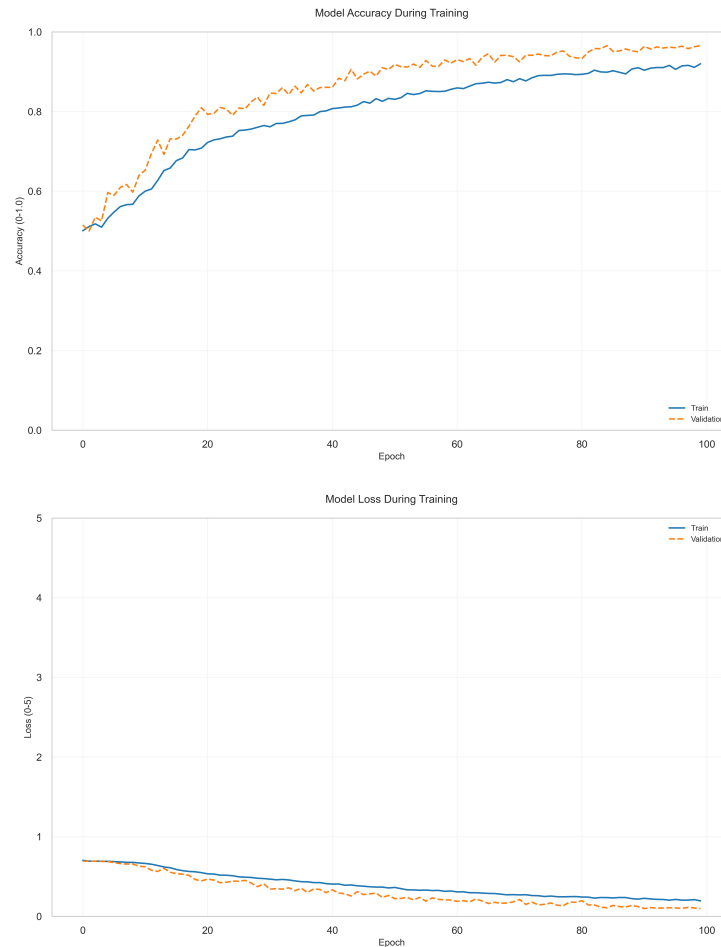
6.3 Resultados e Análises

A avaliação do modelo foi conduzida em múltiplas frentes: desempenho quantitativo, robustez, comparação com o estado da arte e análise qualitativa através do framework *eXAI*.

6.3.1 Análise do Treinamento e Desempenho no Teste

O monitoramento das métricas durante o treinamento (Figura 8) mostrou uma convergência estável das curvas de acurácia e perda para os conjuntos de treinamento e validação, indicando ausência de sobreajuste significativo.

Figura 8 – Curvas de acurácia e perda (*loss*) do modelo *CNN* durante o treinamento para o estudo de caso de *Seam Carving*.



Fonte: Elaborado pelo autor (2025).

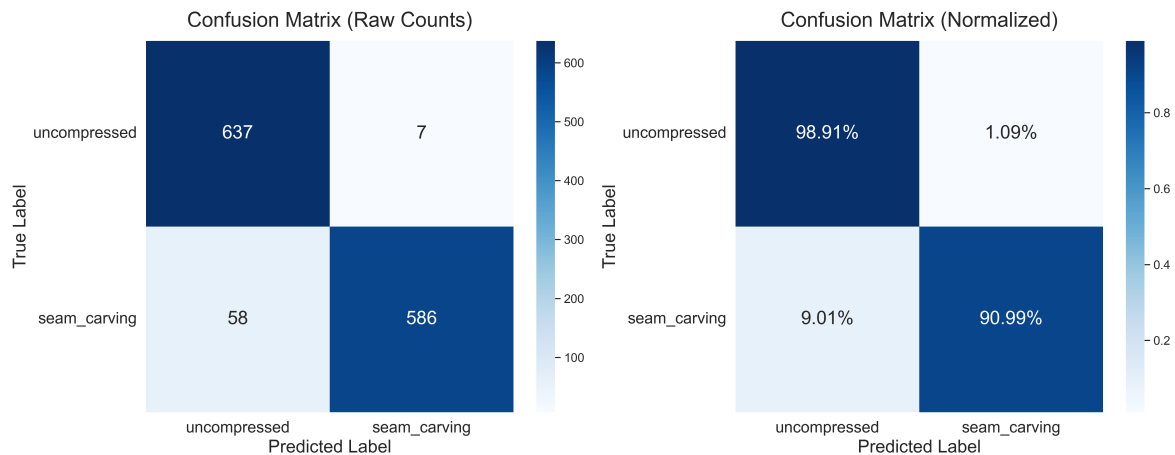
No conjunto de teste (1.288 imagens), o modelo alcançou uma acurácia geral de 95%, com uma precisão de 99% para a classe *seam_carving*. Esta alta precisão é crucial para aplicações forenses, pois minimiza a ocorrência de falsos positivos. A Tabela 7 e a Figura 9 detalham estes resultados. A principal fonte de erro foram os Falsos Negativos (58 casos), sugerindo que os artefatos em algumas imagens eram excepcionalmente sutis.

Tabela 7 – Métricas de desempenho detalhadas para o modelo *CNN* no conjunto de teste de *Seam Carving*.

Classe	Precisão	Revocação	F1-score	Suporte
Uncompressed (C0)	0.92	0.99	0.95	644
Seam Carving (C1)	0.99	0.91	0.95	644
Acurácia Média Ponderada	95%			

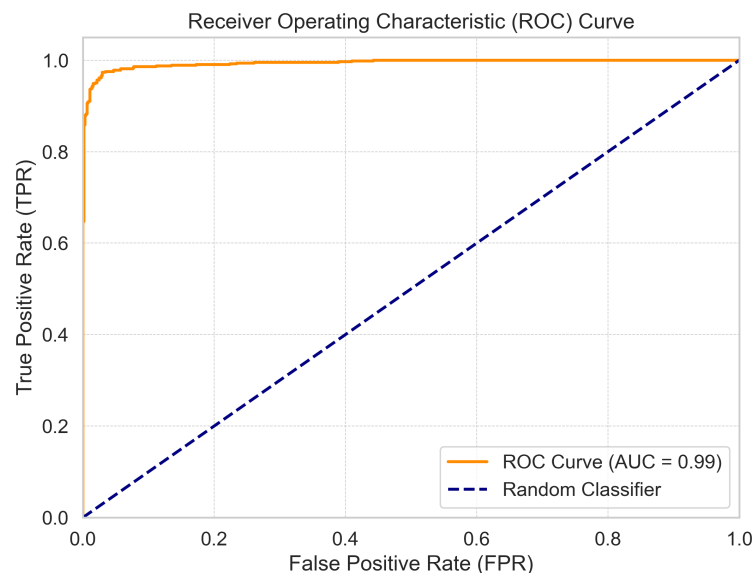
Fonte: Elaborado pelo autor (2025).

Figura 9 – Matrizes de confusão para o modelo CNN no conjunto de teste. A matriz à esquerda exibe as contagens brutas das previsões. A matriz à direita apresenta os resultados normalizados, onde cada linha é normalizada pelo total da classe verdadeira; os valores na diagonal representam, portanto, a revocação (taxa de verdadeiros positivos) para cada classe.



Fonte: Elaborado pelo autor (2025).

Figura 10 – Curva ROC (Receiver Operating Characteristic) para o modelo CNN no conjunto de teste. O alto valor de AUC (0.99) indica um excelente desempenho discriminativo.



Fonte: Elaborado pelo autor (2025).

A curva ROC (Receiver Operating Characteristic) e a métrica AUC (Area Under the Curve) confirmam a excelente capacidade discriminativa do modelo. Com um valor de AUC de 0.99, o modelo demonstra uma performance próxima à de um classificador ideal para distinguir entre as duas classes, independentemente do limiar de decisão.

6.3.2 Análise de Robustez e Comparação com o Estado da Arte

Para avaliar a robustez do modelo em cenários realistas, seu desempenho foi testado contra imagens do conjunto de teste re-comprimidas em diferentes níveis de qualidade JPEG. A Tabela 8 mostra uma degradação clara da acurácia conforme o nível de compressão aumenta, um comportamento esperado, já que a compressão pode destruir os artefatos sutis da manipulação.

Tabela 8 – Acurácia do modelo vs. Qualidade da Compressão JPEG.

Nível de Qualidade JPEG	Acurácia no Conjunto de Teste
100 (Original)	95.73%
90	93.71%
70	78.11%
50	61.80%

Fonte: Elaborado pelo autor (2025).

Adicionalmente, para contextualizar a performance, comparamos nosso modelo com métodos estado-da-arte (SOTA), conforme resumido na Tabela 9. Embora uma comparação direta seja complexa devido a variações nos conjunto de dados, nosso modelo customizado apresenta uma acurácia altamente competitiva, alinhada com os resultados de arquiteturas especializadas.

Tabela 9 – Comparação de performance com métodos SOTA para detecção de *Seam Carving*.

Método	Tipo de conjunto de dados	Acurácia Reportada
Nam et al. (2019) (NAM et al., 2019a)	Custom / Não comprimido	98.81%
Ye et al. (2019)	Custom / Não comprimido	97.80%
Celebi et al. (2022) (CELEBI; HSU; LIU, 2022)	JPEG Comprimido (QF95)	94% - 96%
Modelo Proposto	Custom / Balanceado	95% - 97%

Fonte: Elaborado pelo autor (2025).

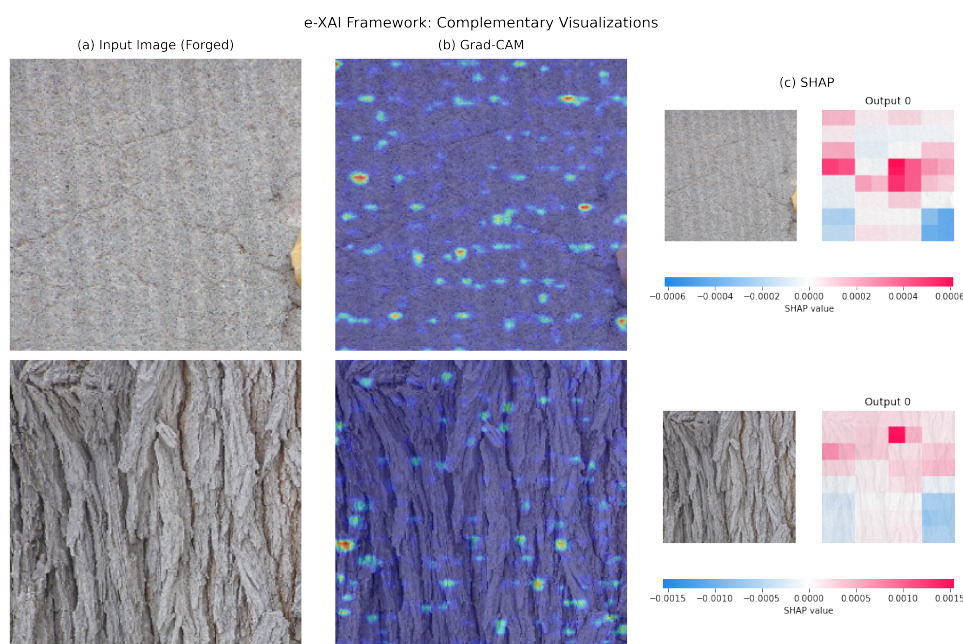
6.3.3 Análise de Explicabilidade (Resultados do eXAI)

Além do desempenho quantitativo, o framework eXAI foi aplicado para construir uma narrativa compreensível sobre o processo decisório do modelo.

6.3.3.1 Análise Qualitativa com Grad-CAM e SHAP

A Figura 11 ilustra a sinergia entre Grad-CAM e SHAP. Os mapas de calor do Grad-CAM revelam que o modelo consistentemente foca sua atenção em áreas de alta complexidade textural (como o chão rochoso ou o tecido estampado). Essas são precisamente as regiões onde o algoritmo de *seam carving* é forçado a tomar decisões mais intrincadas, deixando traços estatísticos mais detectáveis. O Grad-CAM, portanto, responde à pergunta: "Onde o modelo procurou por evidências?".

Figura 11 – Visualizações complementares do framework *eXAI*. (a) Imagem de entrada forjada. (b) Mapa de calor Grad-CAM, indicando regiões de interesse. (c) Mapa de atribuição SHAP, onde pixels vermelhos contribuíram para a classificação de fraude.



Fonte: Elaborado pelo autor (2025).

Complementarmente, os mapas de atribuição do SHAP respondem à pergunta: "Que evidência específica o modelo encontrou?". Os pixels em vermelho, que contribuem positivamente para a classificação de fraude, não formam objetos semânticos. Em vez disso, eles aparecem como um padrão distribuído de perturbações de baixo nível dentro das regiões de alta atenção identificadas pelo Grad-CAM. Juntas, as duas técnicas mostram que o modelo aprendeu a identificar regiões suscetíveis à manipulação e, dentro delas, a detectar os artefatos estatísticos que confirmam a fraude.

6.3.3.2 Análise Quantitativa dos Mapas de Explicabilidade

Para validar estatisticamente a observação de que o modelo foca em áreas complexas, investigamos a correlação entre a atenção do modelo e a complexidade visual. A complexidade local da imagem foi quantificada usando a variância do operador Laplaciano, uma métrica comum para detecção de textura e bordas. Foi calculado o coeficiente de correlação de Pearson entre a intensidade média do Grad-CAM e a variância Laplaciana em uma grade de recortes não sobrepostos da imagem.

A análise revelou uma correlação positiva e estatisticamente significativa ($r = 0.3220$, $p < 0.000$). Este resultado confirma quantitativamente que a *CNN* direciona seu foco para regiões de maior complexidade textural, validando as *insights* qualitativas obtidas com as visualizações do *eXAI*.

6.3.4 Custo Computacional e Reprodutibilidade

O treinamento do modelo *CNN* customizado, utilizando o critério de parada antecipada (*Early Stopping*), foi concluído em aproximadamente 4 horas no ambiente de hardware especificado. A inferência em uma única imagem é realizada em milissegundos, tornando o modelo prático para análises em lote. Para fomentar a transparência e permitir a verificação dos resultados, todo o código-fonte para a criação do conjunto de dados, treinamento do modelo e geração das análises de explicabilidade está disponível no GitHub: <<https://github.com/migmiguel29/SeamCarvingXai>>.

6.4 Discussão e Conclusão do Capítulo

Este estudo de caso validou com sucesso a aplicação da Abordagem 1 do framework *eXAI*. O modelo *CNN* customizado demonstrou ser altamente eficaz, com uma acurácia competitiva e uma precisão de 99% para a classe forjada, um requisito fundamental para aplicações forenses.

A principal contribuição, no entanto, vai além da acurácia. O framework *eXAI* forneceu *insights* transparentes e valiosos sobre o comportamento do modelo. A combinação sinérgica e a validação quantitativa das técnicas de explicabilidade confirmaram que o modelo aprendeu a identificar artefatos estruturais consistentes com a manipulação, em vez de atalhos espúrios baseados no conteúdo semântico. A análise de robustez também destacou uma limitação prática do modelo frente à compressão JPEG, apontando direções para trabalhos futuros, como o treinamento com dados aumentados incluindo imagens comprimidas.

Em suma, a combinação de alto desempenho com interpretabilidade validada torna esta abordagem particularmente valiosa para domínios de forense digital. Tendo validado com sucesso as duas abordagens do framework *eXAI*, o próximo capítulo procederá com uma discussão geral e comparativa dos achados de ambos os estudos de caso, sintetizando as contribuições centrais da tese.

7 Discussão Geral e Comparativa

Os capítulos anteriores apresentaram a aplicação e validação do framework *eXAI* em dois domínios distintos e de alta criticidade: o diagnóstico assistido por computador de Retinopatia Diabética (RD) e a detecção de adulteração de imagens por *seam carving* para fins de segurança da informação. Este capítulo visa sintetizar e comparar os achados desses dois estudos de caso, discutindo as implicações mais amplas da metodologia proposta e sua contribuição para o campo da Inteligência Artificial Explicável (XAI).

7.1 Análise Comparativa dos Resultados e Abordagens

A tese propôs um framework metodológico flexível, que foi instanciado de duas maneiras distintas para atender às necessidades específicas de cada problema, conforme recapitulado na Tabela 10.

Tabela 10 – Comparativo das abordagens nos dois estudos de caso.

Aspecto	Estudo de Caso 1 (RD)	Estudo de Caso 2 (Seam Carving)
Problema	Diagnóstico Médico (Segurança do Paciente)	Forense Digital (Segurança da Informação)
Dados de Entrada	Multimodal (Imagem + Dados Clínicos)	Unimodal (Apenas Imagem)
Arquitetura	Ensemble de <i>Stacking</i> (CNN + LGBM)	CNN de Classificação Direta
Desbalanceamento	Altamente desbalanceado (realista)	Perfeitamente balanceado (controlado)
Acurácia (Teste)	97.0%	95.7%
Principal Desafio	Minimizar Falsos Negativos (doença não detectada)	Detectar artefatos sutis e distribuídos

Fonte: Elaborado pelo autor (2025).

A análise comparativa revela que a flexibilidade do framework foi crucial. No diagnóstico de RD, um problema intrinsecamente multimodal, a integração de dados clínicos através da arquitetura de *Stacking* (Abordagem 2) foi essencial para alcançar um desempenho de ponta. Conforme demonstrado no Capítulo 5, o modelo híbrido não apenas superou significativamente os modelos de modalidade única, mas o fez com um F1-score balanceado superior, métrica crítica em cenários desbalanceados. A combinação das *características* extraídas pela CNN com o contexto do paciente permitiu uma redução drástica nos falsos negativos, um requisito fundamental para a segurança do paciente em aplicações de triagem médica.

Por outro lado, na detecção de *Seam Carving*, um problema unimodal onde a evidência reside unicamente nos dados da imagem, a abordagem de classificação direta com uma CNN

otimizada (Abordagem 1) provou ser suficiente e altamente eficaz, alcançando uma acurácia de aproximadamente 96%. Isso demonstra que o núcleo do framework, uma CNN bem treinada atuando como extratora de *características* ou geradora de predições, é robusto. A camada de *ensemble* de *stacking* pode, então, ser vista como um módulo adicional que pode ser acoplado ao núcleo da CNN quando dados multimodais estão disponíveis e são relevantes para o problema, confirmando a modularidade e adaptabilidade da metodologia proposta.

7.2 O Papel Contextual do XAI: Insights Distintos para Problemas Distintos

A contribuição central desta tese reside não apenas na acurácia dos modelos, mas na profundidade e, crucialmente, na natureza contextual dos *insights* gerados pela aplicação do framework eXAI. A análise comparativa da explicabilidade nos dois domínios demonstrou que o valor e o propósito do XAI são altamente dependentes do contexto da aplicação, um conceito alinhado com a literatura que defende que a interpretabilidade não é um fim em si mesma, mas um meio para alcançar um objetivo específico, como confiança, robustez ou justiça (DOSHI-VELEZ; KIM, 2017a).

No estudo de caso da Retinopatia Diabética, a análise *SHAP* do modelo híbrido serviu para validar o alinhamento do modelo com o conhecimento clínico pré-existente. Ela não apenas confirmou que as *características* extraídas pela CNN eram as mais importantes, mas também quantificou a influência de fatores de risco conhecidos, como *diabetes_time_y* e a presença de *drusens*. Isso transforma a caixa-preta em um "assistente de vidro", cujo processo de decisão é congruente com o raciocínio médico. Neste contexto, o XAI funciona como um construtor de confiança, um passo essencial para a adoção de sistemas de IA na prática clínica.

Em contrapartida, no estudo de caso de *Seam Carving*, o XAI serviu a um propósito diferente: revelar uma estratégia de detecção desconhecida e validar a robustez do modelo. A análise sinérgica de *SHAP* e *Grad-CAM* revelou que o modelo aprendeu a ignorar o conteúdo semântico das imagens e a focar em artefatos de textura de baixa-nível e distorções estatísticas imperceptíveis para humanos. Isso é crucial em forense digital, pois confirma que o modelo não está aprendendo atalhos ou vieses espúrios, mas sim os traços fundamentais da manipulação. Aqui, o XAI não valida um conhecimento pré-existente, mas sim revela a estratégia interna do detector, aumentando a confiança em sua generalidade e resistência a vieses.

Portanto, esta tese demonstra que o XAI não é apenas uma ferramenta para abrir a caixa-preta, mas um instrumento metodológico que, dependendo do domínio, pode:

1. Validar a relevância clínica de um modelo de diagnóstico, comparando suas decisões com o conhecimento de especialistas.

2. Revelar as estratégias de detecção de um modelo de segurança, garantindo que ele não se baseie em correlações espúrias.
3. Aumentar a confiança do usuário final, seja ele um médico ou um analista forense, ao fornecer justificativas inteligíveis para as previsões da IA, o que pode levar a uma melhor colaboração humano-máquina (ARRIETA et al., 2020).

7.3 Implicações e Contribuições Gerais para a Pesquisa

Esta pesquisa oferece contribuições teóricas e práticas. Teoricamente, ela propõe e valida um framework, o *eXAI*, que une de forma conceitual o *Ensemble Learning* com técnicas de XAI para criar sistemas de análise de imagem mais robustos e transparentes. A tese avança a noção de *ensemble* ao demonstrar que este conceito pode ser aplicado não apenas a modelos preditivos para melhorar a acurácia, mas também a técnicas de explicação para gerar uma compreensão mais holística e confiável.

Praticamente, os dois estudos de caso demonstram soluções eficazes para problemas reais. O modelo para RD apresenta um potencial significativo como ferramenta de triagem, capaz de otimizar o fluxo de trabalho de oftalmologistas com alta sensibilidade para a doença. O modelo para detecção de *Seam Carving* representa uma ferramenta valiosa para a área de forense digital, onde a detecção de manipulações sutis é de extrema importância.

A principal implicação deste trabalho é a reafirmação de que a busca por acurácia em modelos de *Deep Learning* não deve ser desassociada da busca por interpretabilidade, especialmente em aplicações críticas. O framework *eXAI* oferece um caminho pragmático para alcançar ambos os objetivos simultaneamente, alinhando a performance quantitativa com a validação qualitativa, um passo essencial para o desenvolvimento de uma Inteligência Artificial responsável.

7.4 Limitações da Pesquisa e Trabalhos Futuros

Apesar dos resultados promissores, esta pesquisa possui limitações que abrem caminhos para trabalhos futuros. A principal limitação é a validação dos modelos em conjuntos de dados específicos para cada problema. Embora os protocolos de teste tenham sido rigorosos, a generalização dos modelos para outros *conjunto de dados* com diferentes características precisa ser investigada. A validação em *conjunto de dados* públicos de referência, como EyePACS ou APTOS para RD, e contra outras variações de manipulação de imagem, seria um passo importante para *benchmark* externo.

Adicionalmente, no estudo de caso de RD, apenas um conjunto limitado de *características* clínicas foi utilizado. A incorporação de um leque mais amplo de dados do paciente,

como comorbidades ou uso de insulina, poderia aprimorar ainda mais o desempenho do modelo híbrido.

Por fim, embora análises quantitativas iniciais tenham sido realizadas para validar as saídas do XAI, a próxima etapa seria a validação das explicações geradas diretamente com especialistas. A condução de estudos de usuário com médicos e analistas forenses seria um passo importante para quantificar a utilidade clínica e prática do framework e o seu impacto real na tomada de decisão.

8 Conclusão

Esta tese propôs e validou um framework metodológico, denominado *eXAI* (*Ensemble Explainable Artificial Intelligence*), projetado para desenvolver sistemas de classificação de imagens que são simultaneamente precisos, robustos e interpretáveis. Através da aplicação deste framework em dois domínios críticos distintos, o diagnóstico de Retinopatia Diabética e a detecção de adulteração de imagens por *Seam Carving*, demonstrou-se sua flexibilidade e eficácia.

A principal contribuição deste trabalho reside na demonstração de que a integração de técnicas de *Deep Learning* (*CNNs*), *Ensemble Learning* — particularmente o *stacking* para dados multimodais — e XAI pode superar os desafios da opacidade dos modelos de caixa-preta. Conforme discutido no Capítulo 7, os resultados dos estudos de caso validam a hipótese central desta tese: é possível construir modelos de IA para aplicações críticas que não apenas alcançam alta performance, mas cujas decisões são transparentes e alinhadas ao conhecimento do especialista. A análise XAI, em particular, provou ser uma ferramenta de valor contextual, primeiramente ao validar o raciocínio clínico em um cenário de diagnóstico e, em segundo lugar, ao revelar a estratégia de detecção de artefatos em um cenário forense.

Em suma, esta pesquisa contribui com um framework metodológico pragmático e validado, reforçando que a interpretabilidade não deve ser vista como um adendo, mas como um componente central no desenvolvimento de sistemas de IA confiáveis e prontos para a adoção em cenários do mundo real.

8.1 Limitações e Trabalhos Futuros

Apesar dos resultados promissores, esta pesquisa possui limitações que, ao serem reconhecidas, abrem avenidas importantes para trabalhos futuros.

Uma primeira limitação refere-se à validação dos modelos em conjuntos de dados específicos para cada problema. Embora os protocolos de teste tenham sido rigorosos, a generalização dos modelos para outros *datasets* com diferentes características precisa ser investigada. Um passo futuro crucial seria a validação em *datasets* públicos de referência, como EyePACS ou APTOS para RD, para realizar um *benchmark* externo e avaliar a robustez do framework em cenários mais amplos.

Do ponto de vista metodológico, conforme apontado durante a avaliação deste trabalho, a robustez estatística das conclusões poderia ser aprofundada. A validação dos modelos foi baseada em uma única divisão dos dados, e trabalhos futuros poderiam empregar técnicas como a validação cruzada (*k-fold*) ou múltiplas execuções com sementes aleatórias distintas.

Adicionalmente, a aplicação de testes estatísticos formais, como o teste de DeLong para a comparação de curvas ROC, permitiria uma quantificação mais rigorosa da significância das diferenças de desempenho entre os modelos.

No que tange à explicabilidade, um esforço foi feito para ir além da análise qualitativa, incorporando validações quantitativas iniciais para validar a coerência das saídas do XAI. No estudo de *Seam Carving*, demonstrou-se estatisticamente a correlação entre a atenção do modelo (Grad-CAM) e a complexidade textural da imagem. No estudo de RD, foi comprovada a correlação entre a importância das características visuais (SHAP) e os fatores de risco clínicos. No entanto, a estabilidade das próprias explicações não foi formalmente testada. Uma direção valiosa para pesquisas futuras, alinhada às sugestões recebidas, seria a implementação de testes de robustez para as explicações, como os de inserção e deleção de *features*, para quantificar a confiabilidade dos valores SHAP frente a perturbações na entrada.

Finalmente, outras direções de pesquisa incluem a expansão da integração de dados clínicos no estudo de RD e o desenvolvimento de interfaces de XAI mais interativas para validar a utilidade prática do framework através de estudos de usuário com especialistas (médicos e analistas forenses), quantificando o impacto real das explicações geradas na tomada de decisão.

Referências

- ADADI, A.; BERRADA, M. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, v. 6, p. 52138–52160, 2018.
- ARRIETA, A. B.; DÍAZ-RODRÍGUEZ, N.; SER, J. D.; BENNETOT, A.; TABIK, S.; BARBADO, A.; GARCÍA, S.; GIL-LÓPEZ, S.; MOLINA, D.; BENJAMINS, R.; CHATILA, R.; HERRERA, F. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, v. 58, p. 82–115, 2020.
- AVIDAN, S.; SHAMIR, A. Seam carving for content-aware image resizing. *ACM Trans. Graph.*, ACM, v. 26, n. 3, p. 10, 2007.
- BAKALO, R.; GOLDBERGER, J.; BEN-ARI, R. Weakly and semi supervised detection in medical imaging via deep dual branch net. *Neurocomputing*, Elsevier BV, v. 421, p. 15–25, Jan 2021. ISSN 0925-2312. Disponível em: <<http://dx.doi.org/10.1016/j.neucom.2020.09.037>>.
- BIACH, F.; IALA, I.; LAANAYA, H.; MINAOUI, K. Encoder-decoder based convolutional neural networks for image forgery detection. *Multimedia Tools and Applications*, 01 2021.
- BIOLCHINI, J.; MIAN, P. G.; NATALI, A. C. C.; TRAVASSOS, G. H. Systematic review in software engineering. *System Engineering and Computer Science Department COPPE/UFRJ, Technical Report ES*, v. 679, n. 05, p. 45, 2005.
- BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, p. 5–32, 2001.
- BUDA, M.; MAKI, A.; MAZUROWSKI, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, Elsevier, v. 106, p. 249–259, 2018.
- CAI, Y.; KONG, X.; WANG, X. Temporal action detection with long action seam mechanism. *ACM Digital Library*, v. 2, p. 165–171, 2018.
- CELEBI, N.; HSU, T.; LIU, Q. A comparison study to detect seam carving forgery in JPEG images with deep learning models. *Journal of Surveillance, Security and Safety*, v. 3, n. 2, p. 88–100, 2022.
- CHEN, C.; GRAUMAN, K. Subjects and their objects: Localizing interactees for a person-centric view of importance. *International Journal of Computer Vision*, v. 126, 2018. Disponível em: <<https://doi.org/10.1007/s11263-016-0958-6>>.
- CHENG, H.-J.; WEI, J.-D.; LIN, C.-Y.; YE, J. Detecting seam-carved image by extreme learning machines using patch analysis method, jury voting, and combinatorial fusion. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, v. 50, n. 5, p. 1850–1864, 2018.
- CHITTAJALLU, D. R.; DONG, B.; TUNISON, P.; COLLINS, R.; WELLS, K.; FLESHMAN, J.; SANKARANARAYANAN, G.; SCHWARTZBERG, S.; CAVUOTO, L.; ENQUOBAHRIE, A. Xai-cbir: Explainable ai system for content based retrieval of video frames from minimally invasive surgery videos. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. [S.l.: s.n.], 2019. p. 66–69.

CHOI, J.; EUN, H.; KIM, C. Boosting proximal dental caries detection via combination of variational methods and convolutional neural network. *J. Signal Process. Syst.*, Kluwer Academic Publishers, USA, v. 90, n. 1, p. 87-97, jan. 2018. ISSN 1939-8018. Disponível em: <<https://doi.org/10.1007/s11265-016-1214-6>>.

CIESLAK, L. F. S.; COSTA, K. Augusto Pontara da; PAPA, J. P. Seam carving detection using convolutional neural networks. In: *2018 IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI)*. [S.l.: s.n.], 2018. p. 000195-000200.

DAVIS, B.; GLENSKI, M.; SEALY, W.; ARENDT, D. *Measure Utility, Gain Trust: Practical Advice for XAI Researcher*. 2020.

DIETTERICH, T. G. Ensemble methods in machine learning. In: SPRINGER. *International workshop on multiple classifier systems*. [S.l.], 2000. p. 1-15.

DOSHI-VELEZ, F.; KIM, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

DOSHI-VELEZ, F.; KIM, B. *Towards A Rigorous Science of Interpretable Machine Learning*. arXiv, 2017. Disponível em: <<https://arxiv.org/abs/1702.08608>>.

ERHAN, D.; BENGIO, Y.; COURVILLE, A.; VINCENT, P. Visualizing higher-layer features of a deep network. In: *University of Montreal*. [S.l.: s.n.], 2009. v. 1341, n. 3, p. 1.

ERTORK, A.; ERTORK, S. Anomaly preserving content-aware hyperspectral image size reduction. In: *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*. [S.l.: s.n.], 2018. p. 6368-6371.

FURUTA, R.; TSUBAKI, I.; YAMASAKI, T. Fast volume seam carving with multipass dynamic programming. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 28, n. 5, p. 1087-1101, 2018.

FUTIA, G.; VETRO, A. On the integration of knowledge graphs into deep learning models for a more comprehensible ai—three challenges for future research. *Information*, v. 11, n. 2, 2020. ISSN 2078-2489. Disponível em: <<https://www.mdpi.com/2078-2489/11/2/122>>.

IQBAL; CHEN, M.; FU, L.; LIN, H.; YUN. Seam carve detection using convolutional neural networks. In: LIU SHUAI AND YANG, G. (Ed.). *Advanced Hybrid Information Processing*. Cham: Springer International Publishing, 2019. p. 392-407. ISBN 978-3-030-19086-6.

JAGTAP, M. T.; TRIPATHI, R. A novel 3-d image retargeting by using stereo seam carving with disparity map acquisition (dma) algorithm. In: *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*. [S.l.: s.n.], 2019. p. 200-203.

JAM, J.; KENDRICK, C.; WALKER, K.; DROUARD, V.; HSU, J. G.-S.; YAP, M. H. A comprehensive review of past and present image inpainting methods. *Computer Vision and Image Understanding*, v. 203, p. 103147, 2021. ISSN 1077-3142. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1077314220301661>>.

JOUIS, G.; MOUCHÈRE, H.; PICAROUGNE, F.; HARDOUIN, A. Anchors vs attention: Comparing xai on a real-life use case. In: _____. [S.l.: s.n.], 2021. p. 219-227. ISBN 978-3-030-68795-3.

KE, G.; MENG, Q.; FINLEY, T.; WANG, T.; CHEN, W.; MA, W.; YE, Q.; LIU, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. In: *Advances in neural information processing systems* 30. [S.l.: s.n.], 2017.

KEANE, M.; KENNY, E. How case-based reasoning explains neural networks: A theoretical analysis of xai using post-hoc explanation-by-example from a survey of ann-cbr twin-systems. In: _____. [S.l.: s.n.], 2019. p. 155–171. ISBN 978-3-030-29248-5.

KENNY, E. M.; FORD, C.; QUINN, M.; KEANE, M. T. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in xai user studies. *Artificial Intelligence*, v. 294, p. 103459, 2021. ISSN 0004-3702. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0004370221000102>>.

KITCHENHAM, B.; CHARTERS, S. Guidelines for performing systematic literature reviews in software engineering. Citeseer, 2007.

LAMY, J.-B.; SEKAR, B.; GUEZENNEC, G.; BOUAUD, J.; SéROUSSI, B. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial Intelligence in Medicine*, v. 94, p. 42–53, 2019. ISSN 0933-3657. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0933365718304846>>.

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, leee, v. 86, n. 11, p. 2278–2324, 1998.

LI, C.; HU, R.; LIANG, C.; XIAO, C.; RUAN, W. Faster seam carving for video retargeting. In: *2018 25th IEEE International Conference on Image Processing (ICIP)*. [S.l.: s.n.], 2018. p. 823–827.

LIU, Q. An approach to detecting JPEG down-recompression and seam carving forgery under recompression anti-forensics. *Pattern Recognition*, v. 65, p. 35–46, 2017.

LIU, Q. An improved approach to exposing JPEG seam carving under recompression. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 29, n. 7, p. 1907–1918, 2019.

LIU, Q. *Seam Carving JPEG Image Forgery Dataset*. 2021. Kaggle. Accessed: [Data em que você acessou, e.g., 15 May 2024]. Disponível em: <<https://www.kaggle.com/datasets/liuqzsc/seam-carving-jpeg-image-forgery-dataset>>.

LUNDBERG, S.; LEE, S.-I. *A Unified Approach to Interpreting Model Predictions*. 2017. Disponível em: <<https://arxiv.org/abs/1705.07874>>.

MYERS-DEAN, J.; WEHRWEIN, S. Semantic pixel distances for image editing. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. [S.l.: s.n.], 2020. p. 2267–2274.

NAIR, V.; HINTON, G. E. Rectified linear units improve restricted boltzmann machines. In: *Proceedings of the 27th International Conference on Machine Learning (ICML'10)*. Madison, WI, USA: Omnipress, 2010. p. 807–814. ISBN 9781605589077.

NAKAYAMA, L. F.; RESTREPO, D.; MATOS, J.; RIBEIRO, L. Z.; MALERBI, F. K.; CELI, L. A.; REGATIERI, C. S. Brset: a brazilian multilabel ophthalmological dataset of retina fundus photos. *medRxiv*, Cold Spring Harbor Laboratory Preprints, 2024.

NAM, S.-H.; AHN, W.; MUN, S.-M.; PARK, J.; KIM, D.; YU, I.-J.; LEE, H.-K. Content-aware image resizing detection using deep neural network. In: *2019 IEEE International Conference on Image Processing (ICIP)*. [S.l.: s.n.], 2019. p. 106–110.

NAM, S.-H.; AHN, W.; MUN, S.-M.; PARK, J.; KIM, D.; YU, I.-J.; LEE, H.-K. Content-aware image resizing detection using deep neural network. In: *2019 IEEE International Conference on Image Processing (ICIP)*. [S.l.: s.n.], 2019. p. 106–110.

NAM, S.-H.; AHN, W.; YU, I.-J.; KWON, M.-J.; SON, M.; LEE, H.-K. Deep convolutional neural network for identifying seam-carving forgery. *IEEE Transactions on Circuits and Systems for Video Technology*, Institute of Electrical and Electronics Engineers (IEEE), v. 31, n. 8, p. 3308–3326, Aug 2020. ISSN 1558-2205. Disponível em: <<http://dx.doi.org/10.1109/TCSVT.2020.3037662>>.

NAZARI, H.; AKGÜN, D. A deep learning model for image retargeting level detection. In: *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*. [S.l.: s.n.], 2020. p. 1–4.

NAZIH, W.; ASEERI, A. O.; ATALLAH, O. Y.; EL-SAPPAGH, S. Vision transformer model for predicting the severity of diabetic retinopathy in fundus photography-based retina images. *IEEE Access*, IEEE, v. 11, p. 117546–117561, 2023.

NGUYEN, T. V.; KANKANHALLI, M. As-similar-as-possible saliency fusion. *Multimedia Tools Appl.*, Kluwer Academic Publishers, USA, v. 76, n. 8, p. 10501–10519, abr. 2017. ISSN 1380-7501. Disponível em: <<https://doi.org/10.1007/s11042-016-3615-8>>.

NIU, Y.; GU, L.; ZHAO, Y.; LU, F. Explainable diabetic retinopathy detection and retinal image generation. *IEEE Journal of Biomedical and Health Informatics*, v. 26, n. 1, p. 44–55, 2022.

QUELLEC, G.; AI Hajj, H.; LAMARD, M.; CONZE, P.-H.; MASSIN, P.; COCHENER, B. Explain: Explanatory artificial intelligence for diabetic retinopathy diagnosis. *Medical Image Analysis*, v. 72, p. 102118, 2021. ISSN 1361-8415.

ROTHER, C.; KOLMOGOROV, V.; BLAKE, A. "grabcut": Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, Association for Computing Machinery, New York, NY, USA, v. 23, n. 3, p. 309–314, aug 2004. ISSN 0730-0301. Disponível em: <<https://doi.org/10.1145/1015706.1015720>>.

ROY, S.; DASGUPTA, T.; PRADHAN, T. An improved and optimized content-aware resizing algorithm for images with densely situated foreground objects. In: *2018 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*. [S.l.: s.n.], 2018. p. 1–6.

RUSSELL, S.; NORVIG, P. *INTELIGÊNCIA ARTIFICIAL*. 2ª. ed. [S.l.]: Campus, 2004.

SASAKI, K.; NAGAHAMA, Y.; ZE, Z.; IIZUKA, S.; SIMO-SERRA, E.; MOCHIZUKI, Y.; ISHIKAWA, H. Adaptive energy selection for content-aware image resizing. In: *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*. [S.l.: s.n.], 2017. p. 852–857.

SCHLEGEL, U.; ARNOUT, H.; EL-ASSADY, M.; OELKE, D.; KEIM, D. A. Towards a rigorous evaluation of XAI methods on time series. *CoRR*, abs/1909.07082, 2019. Disponível em: <<http://arxiv.org/abs/1909.07082>>.

SENAPATI, A.; TRIPATHY, H. K.; SHARMA, V.; GANDOMI, A. H. Artificial intelligence for diabetic retinopathy detection: a systematic review. *Informatics in Medicine Unlocked*, Elsevier, p. 101445, 2024.

SHORTEN, C.; KHOSHGOFTAAR, T. M. A survey on image data augmentation for deep learning. *Journal of big data*, Springer, v. 6, n. 1, p. 1–48, 2019.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. In: *International Conference on Learning Representations*. [S.l.: s.n.], 2015.

SONG, E.; LEE, M.; LEE, S. Carvingnet: Content-guided seam carving using deep convolution neural network. *IEEE Access*, v. 7, p. 284–292, 2019.

SUN, J.-Y.; KIM, S.-W.; LEE, S.-W.; KO, S.-J. A novel contrast enhancement forensics based on convolutional neural networks. *Signal Processing: Image Communication*, v. 63, p. 149–160, 2018. ISSN 0923-5965. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0923596518301036>>.

TAN, T. F.; LIM, G.; TING, D. S. W. Explainable artificial intelligence in ophthalmology. *Current Opinion in Ophthalmology*, Lippincott Williams & Wilkins, v. 34, n. 5, p. 369–375, 2023.

WALIA, S.; KUMAR, K. An eagle-eye view of recent digital image forgery detection methods. In: BHATTACHARYYA, P.; SASTRY, H. G.; MARRIBOYINA, V.; SHARMA, R. (Ed.). *Smart and Innovative Trends in Next Generation Computing Technologies*. Singapore: Springer Singapore, 2018. p. 469–487. ISBN 978-981-10-8660-1.

WANG, X.; NIU, S.; WANG, H. Image inpainting detection based on multi-task deep learning network. *IETE Technical Review*, Taylor & Francis, v. 38, n. 1, p. 149–157, 2021. Disponível em: <<https://doi.org/10.1080/02564602.2020.1782274>>.

WEBER, R. O.; JOHS, A. J.; LI, J.; HUANG, K. Investigating textual case-based xai. In: COX, M. T.; FUNK, P.; BEGUM, S. (Ed.). *CaseBased Reasoning Research and Development*. Cham: Springer International Publishing, 2018. p. 431–447. ISBN 978-3-030-01081-2.

WEI, D.-Y.; CHOU, Y.-C.; SU, P.-C. A multi-operator retargeting scheme for compressed videos. In: *2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*. [S.l.: s.n.], 2018. p. 1–2.

WEI, J.; CHENG, H.; CHANG, C. Hopfield network-based approach to detect seam-carved images and identify tampered regions. *Neural Comput. Appl.*, v. 31, n. 10, p. 6479–6492, 2019. Disponível em: <<https://doi.org/10.1007/s00521-018-3463-8>>.

WEN, B.; SU, G.-M. Transim: Transfer image local statistics across eotfs for hdr image applications. In: *2018 IEEE International Conference on Multimedia and Expo (ICME)*. [S.l.: s.n.], 2018. p. 1–6.

YAU, J. W.; ROGERS, S. L.; KAWASAKI, R.; LAMOUREUX, E. L.; KOWALSKI, J. W.; BEK, T.; CHEN, S.-J.; DEKKER, J. M.; FLETCHER, A.; GRAUSLUND, J. et al. Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care*, American Diabetes Association, v. 35, n. 3, p. 556–564, 2012.

YE, J.; SHI, Y.; XU, G.; SHI, Y.-Q. A convolutional neural network based seam carving detection scheme for uncompressed digital images. In: YOO, C. D.; SHI, Y.-Q.; KIM, H. J.; PIVA, A.; KIM, G. (Ed.). *Digital Forensics and Watermarking*. Cham: Springer International Publishing, 2019. p. 3–13. ISBN 978-3-030-11389-6.

YEOM, S.-K.; SEEGERER, P.; LAPUSCHKIN, S.; WIEDEMANN, S.; MÜLLER, K.-R.; SAMEK, W. Pruning by explaining: A novel criterion for deep neural network pruning. 12 2019.

ZHAO, Y. Fast image blending for high-quality panoramic images on mobile phones. *Multimedia Tools and Applications*, v. 80, p. 499–516, 2021.