UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

GABRIEL ARNAUD DE MELO FRAGOSO

Inteligência Artificial Explicável para Detecção de Fissuras em Concreto Armado: Uma Abordagem Pós-Hoc Combinando Técnicas de XAI em Redes Neurais Convolucionais

Recife

2025

GABRIEL ARNAUD DE MELO FRAGOSO

Inteligência Artificial Explicável para Detecção de Fissuras em Concreto Armado: Uma Abordagem Pós-Hoc Combinando Técnicas de XAI em Redes Neurais Convolucionais

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Mestre Acadêmico em Ciência da Computação.

Área de Concentração: Inteligência Artificial Explicável

Orientadora: Teresa Bernarda Ludermir

Recife

2025

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Fragoso, Gabriel Arnaud de Melo.
Inteligência artificial explicável para detecção de fissuras em concreto armado: uma abordagem pós-hoc combinando técnicas de XAI em eedes neurais convolucionais / Gabriel Arnaud de Melo Fragoso. - Recife, 2025.
73 f.: il.

Dissertação (Mestrado) - Universidade Federal de Pernambuco, Centro de Informática, Programa de Pós-Graduação em Ciência da Computação, 2025.
Orientação: Teresa Bernarda Ludermir.
Inclui referências.

1. Redes Neurais Convolucionais (CNN); 2. Inteligência Artificial Explicável (XAI); 3. Engenharia civil assistida por inteligência artificial. I. Ludermir, Teresa Bernarda. II. Título.

UFPE-Biblioteca Central

Gabriel Arnaud de Melo Fragoso

“Inteligência Artificial Explicável para Detecção de Fissuras em Concreto Armado: Uma Abordagem Pós-Hoc Combinando Técnicas de XAI em Redes Neurais Convolucionais”

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação. Área de Concentração: Computação Inteligente.

Aprovado em: 17/07/2025.

BANCA EXAMINADORA

_____

Profa. Dra. Teresa Bernarda Ludermir
Centro de Informática / UFPE
(**orientadora**)

_____

Prof. Dr. Tiago Pessoa Ferreira de Lima
Departamento de Análise e Desenvolvimento de Sistemas/IFPE

_____

Prof. Dr. Felipe Costa Farias
Escola Politécnica de Pernambuco/UPE

AGRADECIMENTOS

Quero agradecer, em primeiro lugar, à minha família, que sempre foi meu alicerce. À minha mãe e ao meu pai, as pessoas que mais acreditaram em mim e que, desde o começo, estiveram ao meu lado em cada passo. Aos meus irmãos, pelo apoio, pelas risadas e por tornarem cada desafio mais leve.

Ao meu namorado, minha fonte diária de força, por me lembrar todos os dias que eu sou capaz e por me apoiar mesmo nos momentos mais difíceis.

E, é claro, à minha gata Lila, minha companheira fiel, que esteve ali nos momentos de estudo, me animou com seus miados e roubou meu colo quando eu mais precisava relaxar.

Um agradecimento especial à minha orientadora, Teresa, pela oportunidade, pela paciência, pela orientação e por ter caminhado comigo em cada etapa desse processo.

Também quero expressar minha gratidão aos colegas, amigos e a todos que, de alguma forma, contribuíram para essa conquista. Cada palavra de incentivo fez diferença.

RESUMO

Este trabalho propõe uma abordagem inovadora para aumentar a interpretabilidade de redes neurais convolucionais (CNNs) na tarefa de detecção de fissuras em estruturas de concreto, integrando algoritmos de explicabilidade baseados em visualização, como o Grad-CAM, com técnicas de segmentação não supervisionada, a exemplo do algoritmo K-means. A metodologia emprega aprendizado por transferência com arquiteturas consagradas (VGG16, VGG19 e ResNet) alcançando acurácia superior a 99% nos conjuntos de treinamento e teste. Foram avaliadas estratégias de explicabilidade fundamentadas tanto em perturbação do espaço de entrada quanto nos pesos internos das camadas convolucionais. Os resultados indicam que a combinação entre Grad-CAM e K-means aprimora não apenas a acurácia na detecção de fissuras, mas também a transparência do processo decisório, aspecto crítico para aplicações reais em monitoramento estrutural. Para mensurar objetivamente o grau de explicabilidade dos modelos, foi proposta uma nova métrica baseada na sobreposição entre as máscaras geradas pelos mapas de ativação das técnicas de explicabilidade e os segmentos resultantes da clusterização, permitindo avaliar a coerência espacial entre as regiões de interesse destacadas pelo modelo e as áreas efetivamente associadas às fissuras. Embora o desempenho da segmentação tenha sido majoritariamente satisfatório, foram identificadas limitações em imagens com alta complexidade visual. A seleção automatizada de camadas mais relevantes para análise foi validada por especialistas em engenharia civil, reforçando a viabilidade prática da proposta. Ressalta-se o ineditismo da abordagem com foco explícito em explicabilidade no contexto da classificação de fissuras em concreto, com código-fonte disponibilizado abertamente no GitHub. Como desdobramento natural, recomenda-se a investigação de técnicas de segmentação mais robustas e a ampliação do conjunto de dados para abranger maior diversidade de padrões fissurais.

Palavras-chaves: Redes Neurais Convolucionais (CNN). Inteligência Artificial Explicável (XAI). Detecção de Fissuras. Grad-CAM. K-means. Deep Learning. Engenharia Civil Assistida por Inteligência Artificial.

ABSTRACT

This study proposes an innovative approach to enhance the interpretability of Convolutional Neural Networks (CNNs) in the task of crack detection in concrete structures, by integrating visualization-based explainability algorithms, such as Grad-CAM, with unsupervised image segmentation techniques, exemplified by the K-means clustering algorithm. The methodology employs transfer learning using well-established architectures (VGG16, VGG19, and ResNet), achieving accuracy rates above 99% on both training and testing datasets. Explainability strategies were evaluated based on both input perturbation methods and the internal weights of convolutional layers. The results indicate that the combination of Grad-CAM and K-means not only improves the accuracy of crack detection but also enhances the transparency of the decision-making process—an essential aspect for real-world applications in structural health monitoring. To objectively quantify model explainability, a novel metric was proposed based on the overlap between the activation maps generated by explainability techniques and the image segments obtained through clustering. This metric enables the assessment of spatial coherence between the model-highlighted regions of interest and the areas actually associated with cracks. Although the segmentation performance was generally satisfactory, limitations were observed in cases involving images with high visual complexity. The automatic selection of the most relevant layers for analysis was validated by civil engineering experts, reinforcing the practical applicability of the proposed method. This work is noteworthy for being one of the first public contributions focused on explainability in the context of concrete crack classification, with its source code openly available on GitHub. For future work, the adoption of more robust segmentation techniques and the expansion of the dataset to include a greater variety of crack patterns are recommended.

Keywords: Convolutional Neural Networks. Explainable AI. Artificial Intelligence. Grad-CAM. K-means.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CNN	Convolutional Neural Network
DL	Deep Learning
GBP	Guided Backpropagation
GDPR	General Data Protection Regulation
Grad-CAM	Gradient-weighted Class Activation Mapping
IA	Inteligência Artificial
LGPD	Lei de Proteção de Dados Pessoais
LIME	Local Interpretable Model-agnostic Explanations
LRP	Layer-wise Relevance Propagatio
ML	Machine Learning
NLP	Natural Language Processing
PDI	Processamento Digital de Imagens
ResNet	Redes Residuais Profundas (do inglês, Residual Neural Network)
SHAP	SHapley Additive exPlanations
VGG	Visual Geometry Group
XAI	Inteligência Artificial Explicável (do inglês, Explainable Artificial Intelligence)

SUMÁRIO

1 INTRODUÇÃO

Vivemos atualmente uma nova era de transformação industrial, impulsionada pelo avanço de tecnologias sofisticadas, dentre as quais destaca-se a Inteligência Artificial (IA). As máquinas deixaram de realizar apenas funções manuais, passando a executar tarefas racionais que envolvem processos intelectuais, tradicionalmente associados à inteligência humana (LUDERMIR, 2021). Dessa maneira, a inteligência artificial promove a integração entre humanos e máquinas, por meio da incorporação de sistemas inteligentes artificiais em dispositivos como próteses cerebrais, braços biônicos, células artificiais e joelhos inteligentes. Além disso, viabiliza a interação entre humanos e máquinas enquanto agentes distintos, conectados por meio de aplicativos e algoritmos inteligentes (KAUFMAN, 2019).

A IA abrange uma ampla variedade de subcampos, contemplando desde tarefas gerais, como aprendizado e percepção, até atividades específicas, incluindo jogos de xadrez, demonstração de teoremas matemáticos, produção poética, condução autônoma em vias urbanas e diagnóstico médico. Assim, a inteligência artificial apresenta-se como um campo universal, relevante para qualquer atividade intelectual (RUSSELL; NORVIG, 2004).

Entre as aplicações bem-sucedidas da IA, destacam-se os sistemas de chat inteligentes, que por meio do Processamento de Linguagem Natural ou Natural Language Processing (NLP) são capazes de aprender, raciocinar e se comunicar de forma eficiente e contextualizada, utilizando linguagem natural de maneira fluida e conversacional (ORYNBAY et al., 2025). Igualmente importantes são os Sistemas de Recomendação, utilizados por plataformas como Amazon (sugestão de produtos), Netflix (sugestão de filmes e séries) e Spotify (sugestão de músicas), cujas recomendações são alinhadas às preferências dos usuários por meio de técnicas de inteligência artificial (LUDERMIR, 2021).

No âmbito da visão computacional, as Convolutional Neural Network (CNN) (redes neurais convolucionais) destacam-se como algoritmos de aprendizado profundo que simulam o funcionamento do córtex visual dos animais, possibilitando o reconhecimento e a classificação de padrões em imagens digitais (YAMASHITA et al., 2018). Essas redes são amplamente adotadas devido à sua capacidade de aprender automaticamente características relevantes, tais como texturas, formas e cores, a partir dos dados de entrada (ROZENDO et al., 2023).

A trajetória da inteligência artificial remonta a um período anterior ao uso do termo. Em 1943, foi introduzido o conceito de neurônio artificial, o perceptron, um modelo matemático

inspirado no funcionamento dos neurônios biológicos, que fomentou avanços significativos na área e originou as Redes Neurais (MCCULLOCH; PITTS, 1943). O termo "Inteligência Artificial" foi oficialmente utilizado pela primeira vez em 1956, no evento "Dartmouth Summer Research Project on Artificial Intelligence", realizado no Dartmouth College, EUA, embora a concepção integral da IA tenha sido inicialmente discutida no artigo seminal "Computing Machinery and Intelligence", de Alan Turing, em 1950 (TURING, 2009).

Embora diversas técnicas de IA tenham surgido desde a década de 1950, somente recentemente tornou-se viável resolver problemas complexos graças ao aumento do poder computacional e à disponibilidade massiva de dados. O desenvolvimento das GPUs e a expansão das infraestruturas computacionais possibilitaram a aplicação de técnicas avançadas, capazes de enfrentar desafios cada vez mais sofisticados (LUDERMIR, 2021).

Os modelos de Machine Learning (ML) podem ser categorizados em duas classes principais (FELLOUS et al., 2019). A primeira engloba os modelos "caixa-preta", cujas decisões são de difícil compreensão para humanos, como nos casos típicos de Deep Learning (DL). Esses modelos, embora complexos, geralmente apresentam desempenho superior. A segunda classe corresponde aos modelos "caixa-branca", que possuem processos internos transparentes e interpretáveis, permitindo uma análise detalhada das decisões, como exemplificado pelas árvores de decisão (CAMACHO et al., 2018).

A explicabilidade dos modelos de IA constitui um desafio contínuo, especialmente na interpretação da relação entre os recursos de entrada e as decisões do modelo. Há uma demanda crescente por métodos que convertam essas "caixas pretas" em "caixas brancas", proporcionando interpretações biologicamente significativas e acessíveis (CAMACHO et al., 2018).

O avanço da inteligência artificial, com impacto significativo em múltiplas áreas, evidenciou a necessidade de transparência e explicabilidade nos sistemas automatizados. Neste contexto, surgiram regulamentações importantes, como o General Data Protection Regulation (GDPR) na União Europeia e a Lei de Proteção de Dados Pessoais (LGPD) no Brasil, que estabelecem diretrizes para a proteção de dados pessoais e a governança ética da IA (FEITOSA, 2024).

Embora muitos modelos de ML apresentem alta acurácia, frequentemente carecem de transparência e explicabilidade, fatores essenciais para a confiança nos sistemas de IA. A Inteligência Artificial Explicável (do inglês, Explainable Artificial Intelligence) (XAI) surge como um conjunto relativamente novo de técnicas que alia algoritmos sofisticados a métodos explicativos, promovendo soluções interpretáveis que se mostraram eficazes em diversos domínios (FELLOUS et al., 2019).

Todavia, a confiança nas abordagens atuais de explicabilidade pode ser ilusória, pois usuários podem avaliar inadequadamente a qualidade das decisões baseados em explicações locais. As técnicas contemporâneas frequentemente oferecem descrições genéricas do funcionamento dos modelos, mas revelam-se pouco confiáveis para justificar decisões individuais (GHASSEMI; OAKDEN-RAYNER; BEAM, 2021; FEITOSA, 2024).

No campo da visão computacional, a oferta de explicações claras é particularmente crucial, sobretudo para identificar as regiões específicas da imagem que mais influenciaram a decisão do modelo. Compreender quais pixels são determinantes permite aprimorar a eficácia dos modelos, além de aumentar a confiança e a transparência, viabilizando aplicações mais seguras e robustas em múltiplos setores (MINAEE et al., 2021).

1.1 OBJETIVOS E CONTRIBUIÇÕES

Considerando o atual avanço tecnológico e o crescente uso de técnicas de inteligência artificial, esta dissertação tem como objetivo principal analisar a explicabilidade de métodos de inteligência artificial explicável (XAI) aplicados a modelos de classificação de imagens no contexto da engenharia civil. Para tanto, diferentes métodos explicáveis foram combinados e hibridizados, buscando identificar a abordagem mais eficaz, baseada na acurácia e na coerência das predições dos modelos de IA. Dessa forma, o estudo busca responder às seguintes questões centrais:

- Quais abordagens podem ser adotadas para aprimorar os métodos de XAI, de modo que as explicações geradas estejam alinhadas e sejam representativas das predições dos modelos de inteligência artificial?
- Quais critérios e métricas quantitativas podem ser empregadas para avaliar a fidelidade e a consistência dos métodos de XAI na explicação dos processos subjacentes às predições dos modelos?

A partir dessas indagações, torna-se imprescindível a análise criteriosa dos resultados obtidos por cada método individual e híbrido. Essa análise permitirá a seleção e o aprimoramento das técnicas mais promissoras, por meio da combinação estratégica dos métodos explicáveis, com o objetivo de maximizar a interpretabilidade sem comprometer o desempenho preditivo. Para a concretização desse objetivo, os seguintes passos metodológicos foram adotados:

- Implementação e treinamento de modelos avançados de inteligência artificial para classificação de imagens;
- Aplicação e adaptação de métodos XAI específicos para modelos de visão computacional;
- Análise qualitativa e quantitativa das explicações geradas por cada método, visando compreender suas características e limitações;
- Identificação do método explicável mais consistente e adequado ao domínio da engenharia civil;
- Desenvolvimento de estratégias para a hibridização dos métodos selecionados, objetivando a melhoria dos resultados explicativos;
- Definição e aplicação de métricas robustas para a avaliação do desempenho das estratégias híbridas em termos de explicabilidade.

As contribuições científicas deste trabalho são sintetizadas em:

- A realização de uma análise detalhada da eficácia de diversos métodos XAI na interpretação das decisões de modelos de visão computacional aplicados à engenharia civil;
- A proposição e validação de métodos híbridos que potencializam a explicabilidade dos modelos de inteligência artificial neste domínio específico;
- A utilização e adaptação de métricas específicas para avaliar a qualidade e a confiabilidade das explicações produzidas pelos métodos XAI em tarefas de classificação de imagens.

Os experimentos foram implementados utilizando a linguagem Python, com o suporte da plataforma Google Colaboratory. Todo o código-fonte desenvolvido encontra-se disponível para consulta pública em um repositório no Github (FRAGOSO, 2025), promovendo a transparência e a reprodutibilidade dos resultados apresentados.

A originalidade desta proposta reside na integração inédita entre métodos de inteligência artificial explicável (XAI) e métricas de avaliação derivadas da teoria psicométrica, aplicadas ao domínio da engenharia civil. Diferentemente de abordagens convencionais que utilizam avaliações predominantemente qualitativas ou subjetivas, este trabalho propõe um arcabouço

quantitativo rigoroso baseado em métricas como precisão, revocação, especificidade e F1-score aplicadas a mapas de saliência, possibilitando uma análise sistemática da fidelidade e consistência das explicações geradas. Além disso, a hibridização estratégica de métodos XAI, aliada à validação inter-método e inter-instância, representa uma contribuição inovadora ao campo, ampliando a interpretabilidade dos modelos sem comprometer sua acurácia preditiva, especialmente em aplicações críticas como a detecção automatizada de fissuras em estruturas de concreto.

1.2 TRABALHOS RELACIONADOS

A detecção automatizada de manifestações patológicas em estruturas de concreto, particularmente fissuras, tem sido objeto de significativo interesse na engenharia civil, dada a sua relevância para a integridade estrutural e a segurança das edificações. Nos últimos anos, modelos baseados em redes neurais convolucionais (CNNs) demonstraram notável desempenho na classificação de imagens de concreto fissurado e não fissurado. (PEREIRA et al., 2024) avaliaram arquiteturas como VGG16, VGG19 e ResNet50, todas com transferência de aprendizado, alcançando acurácia superior a 99%, utilizando gradientes visuais (Grad-CAM) para verificar a robustez e interpretabilidade dos modelos empregados.

Contudo, embora tais abordagens ofereçam elevada performance preditiva, elas carecem de mecanismos explícitos de interpretabilidade. Dada a natureza *caixa-preta* das CNNs, torna-se necessário incorporar métodos de Inteligência Artificial Explicável (XAI) para elucidar os fatores determinantes na tomada de decisão do modelo. Nesse sentido, técnicas como Gradient-weighted Class Activation Mapping (Grad-CAM), Local Interpretable Model-agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP) e Layer-wise Relevance Propagatio (LRP) vêm sendo amplamente investigadas, fornecendo mapas de saliência para destacar regiões relevantes da imagem com relação à predição realizada (FRESZ; LÖRCHER; HUBER, 2024).

Entretanto, avaliar a qualidade dessas explicações permanece um desafio, principalmente devido à ausência de uma verdade de referência (*ground truth*) pixel a pixel. Para mitigar esse problema, (FRESZ; LÖRCHER; HUBER, 2024) expandiram o conceito do *Focus Score* proposto por Arias-Duart et al., introduzindo uma estrutura de avaliação fundamentada em testes psicométricos e métricas como precisão, revocação, especificidade e *F1-score* para caracterizar a "fidelidade explicativa" dos métodos de saliência.

Na área da construção civil, diversas iniciativas também exploram a segmentação semântica

como abordagem alternativa à classificação, possibilitando a localização precisa das fissuras em nível de pixel. (PEREIRA, 2023) empregou segmentação semântica para detectar fissuras, rachaduras e trincas, evidenciando sua aplicabilidade em inspeções visuais automatizadas com redução significativa de tempo de análise. De maneira semelhante, (COTRIM; SHIMIZU; ROSA, 2023) compararam o desempenho de redes como Mask R-CNN e YOLOv8 para detecção e geolocalização de fissuras em estruturas de concreto, obtendo acurácias superiores a 90%.

Complementarmente, (OTTONI; AL., 2022) propôs métodos automatizados para recomendação de hiperparâmetros em CNNs voltadas à classificação de imagens na construção civil, com foco na otimização de desempenho e robustez. A metodologia *AutoHyperTuningSK* demonstrou impacto significativo na performance de classificadores, alcançando acurácia de até 99,48% na classificação de fissuras, com uso combinado de ajuste de taxa de aprendizado, otimizadores e técnicas de *data augmentation*.

Recentemente, estudos vêm avançando na integração entre detecção de fissuras e explicabilidade. Por exemplo, um método híbrido que combina ResNet-50 com transformada curvelet, otimização por LDA e classificação via XGB, utiliza LIME e Grad-CAM++ para explicitar a porcentagem de fissura detectada, atingindo taxas de acerto superiores a 99% (KAVITHA; BASKARAN; DHANAPRIYA, 2023). Forest et al. (2024) propuseram uma abordagem que extrai segmentações a partir de explicações geradas por classificadores XAI sob supervisão fraca, avaliando sua eficácia para monitoramento de crescimento de fissuras (FOREST et al., 2024).

Apesar do progresso substancial na detecção automatizada de fissuras por meio de redes neurais convolucionais e técnicas de Inteligência Artificial Explicável (XAI), persistem lacunas metodológicas críticas na literatura. Em particular, muitas abordagens permanecem predominantemente qualitativas, baseando-se em visualizações subjetivas de mapas de saliência sem estabelecer critérios objetivos para medir a fidelidade explicativa das regiões destacadas em relação às estruturas reais da imagem (SAMEK; WIEGAND; MÜLLER, 2017; SELVARAJU et al., 2017). Além disso, uma parcela significativa dos estudos não utiliza qualquer forma de ground truth para validação das explicações, tornando inviável aferir a correção das atribuições de importância feitas pelos métodos XAI (ADEBAYO et al., 2018). Embora alguns autores proponham avaliações com base em julgamentos humanos ou estudos de caso anedóticos (HOLZINGER; CARRINGTON; MÜLLER, 2020), essas abordagens são suscetíveis a viés interpretativo e carecem de reprodutibilidade. Existem, no entanto, tentativas de incorporar ground truth explícito à avaliação de saliência: o Pointing Game (ZHANG et al., 2018) verifica se o ponto de máxima ativação recai dentro da máscara anotada da classe-alvo; o RISE (PETSIUK; DAS; SAENKO,

2018) mede o impacto de inserções e remoções em regiões salientes.

Recentemente, Fresz et al. (FRESZ; LÖRCHER; HUBER, 2024) e Arias-Duart et al. (ARIAS-DUART et al., 2022) introduziram estratégias que empregam mosaicos de imagens e métricas derivadas da psicometria (como precisão, revocação e especificidade) para contornar a ausência de ground truth pixel a pixel, propondo uma estrutura formal de avaliação. Ainda assim, tais metodologias permanecem restritas a contextos experimentais e raramente são adotadas nos domínios aplicados da engenharia civil, o que evidencia a necessidade de padronização e validação externa de métricas de explicabilidade em cenários reais.

Tais evidências indicam que, apesar dos avanços na acurácia de modelos para detecção de fissuras, ainda há uma lacuna crítica no que se refere à avaliação sistemática da qualidade das explicações geradas por métodos de XAI. Esta dissertação busca contribuir nesse cenário ao aplicar CNNs em tarefas de detecção e integrar técnicas de explicabilidade cuja avaliação será realizada por métricas formais de saliência, como as propostas por (FRESZ; LÖRCHER; HUBER, 2024). Com isso, pretende-se oferecer uma abordagem quantitativa e objetiva para aferir a coerência das explicações visuais no contexto da inspeção automatizada de estruturas de concreto.

1.3 ORGANIZAÇÃO DO DOCUMENTO

A estrutura deste documento está organizada da seguinte maneira: o Capítulo2, intitulado Fundamentação Teórica, apresenta uma revisão abrangente dos conceitos e fundamentos necessários para o entendimento da pesquisa, incluindo temas como aprendizado de máquina, classificação de imagens, redes neurais convolucionais (CNN), inteligência artificial explicável (XAI) e algoritmos de segmentação de imagens.

No Capítulo3, intitulado Metodologia, são descritos os procedimentos adotados na condução dos experimentos, abrangendo a base de dados, os modelos de classificação e segmentação utilizados, bem como os métodos de explicabilidade aplicados.

O Capítulo4, denominado Método Proposto, detalha o pipeline desenvolvido nesta dissertação, incluindo os fluxos específicos para abordagens baseadas em Grad-CAM (análise dos pesos internos do modelo) e em perturbação de entrada (SHAP e LIME), além da integração com segmentação por K-means.

No Capítulo5, os resultados dos experimentos são apresentados e discutidos, com ênfase nas métricas de desempenho dos classificadores e na avaliação qualitativa e quantitativa das

explicações geradas.

Por fim, o Capítulo6 apresenta as considerações finais do trabalho, discute as limitações identificadas e propõe direções para trabalhos futuros que visem a ampliação e a melhoria da abordagem desenvolvida.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta a fundamentação teórica que sustenta a presente pesquisa, abordando os conceitos e técnicas essenciais para o desenvolvimento do método proposto. Inicialmente, são discutidos os princípios da aprendizagem de máquina, com ênfase em problemas de classificação, destacando os mecanismos subjacentes e as abordagens mais utilizadas. Em seguida, são detalhadas as redes neurais convolucionais (CNNs), enfatizando arquiteturas consagradas, como Visual Geometry Group (VGG) e Redes Residuais Profundas (do inglês, Residual Neural Network) (ResNet), reconhecidas por sua robustez e desempenho superior em tarefas de reconhecimento e classificação de padrões visuais.

Adicionalmente, o capítulo explora os fundamentos da inteligência artificial explicável (XAI), abordando metodologias para a interpretação e transparência dos modelos preditivos, bem como técnicas para avaliação da interpretabilidade das decisões automatizadas.

2.1 FUNDAMENTOS DE IMAGENS

O Processamento Digital de Imagens (PDI) é uma tarefa complexa que envolve um conjunto de etapas interconectadas. O primeiro passo efetivo nesse processo é comumente conhecido como pré-processamento. Para a análise e identificação de objetos, é necessária uma cadeia mais ampla de processos, na qual características ou atributos das imagens, como bordas, texturas e vizinhanças, precisam ser extraídos (QUEIROZ; GOMES, 2006).

Geralmente, as imagens são representadas em escala de cinza (grayscale) ou coloridas (OTTONI; AL., 2022) e devem passar pela etapa de aquisição que tem como função converter uma imagem em uma representação numérica adequada para o subsequente processamento digital (FILHO; NETO, 1999).

Os computadores não podem processar imagens contínuas diretamente; em vez disso, trabalham com arrays de números digitais. Portanto, é necessário representar imagens como arranjos bidimensionais de pontos (QUEIROZ; GOMES, 2006). Uma imagem monocromática pode ser matematicamente representada por uma função f(x,y) que descreve a intensidade luminosa. O valor dessa função em qualquer ponto com coordenadas espaciais (x,y) é proporcional ao brilho da imagem naquele ponto (RUSS, 2006; FILHO; NETO, 1999). A função f(x,y) é o resultado da interação entre a iluminância i(x,y), que indica a quantidade de luz

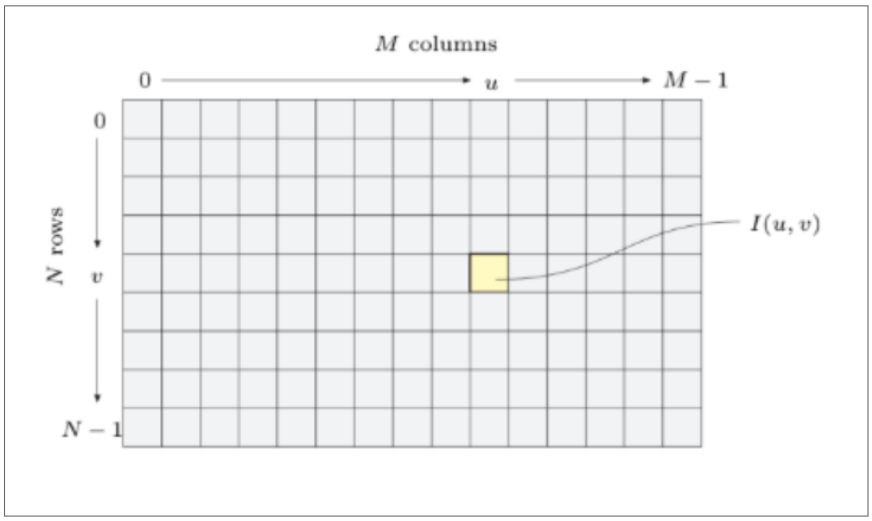
que incide sobre o objeto, e as propriedades de reflectância ou transmitância do objeto, representadas pela função r(x,y). Esta última expressa a fração de luz incidente que o objeto vai refletir ou transmitir para o ponto (x,y) (FILHO; NETO, 1999). A Equação 2.1 representa matematicamente esta função.

$$f(x, y) = i(x, y)r(x, y) \tag{2.1}$$

Em procedimentos computacionais, como a classificação, as imagens são utilizadas no formato digital. Dessa forma, a função f(x,y) é convertida para a forma discreta, resultando em uma imagem digital representada como uma grade de pixels (PEDRINI; SCHWARTZ, 2008). Cada ponto na grade bidimensional que compõe a imagem digital é denominado elemento de imagem ou pixel (QUEIROZ; GOMES, 2006).

No processamento digital de imagens, é comum adotar um sistema de coordenadas em que a origem (u = 0, v = 0) está localizada no canto superior esquerdo. As coordenadas u e v representam, respectivamente, as colunas e as linhas da imagem. Para uma imagem com dimensões M x N, o número máximo de colunas é u max = M - 1, e o número máximo de linhas é v max = N - 1 (BURGER; BURGE, 2022), conforme ilustrado na Figura 1.

Figura 1 – Coordenada em Imagens Digitais Processadas



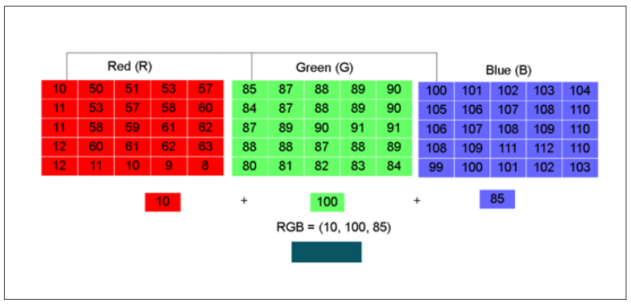
Fonte: BURGER; BURGE (2022)

Por outro lado, a quantização é responsável por definir o número inteiro L de níveis possíveis para cada pixel na imagem (PEDRINI; SCHWARTZ, 2008). Por exemplo, quando L = 256, cada pixel pode ser associado a valores entre 0 e 255. Nesse caso, uma possível representação

remete a atribuir: 0 para pixel preto, 255 para pixel branco e valores intermediários são os demais níveis da escala de cinza (ELGENDY, 2020).

Outro formato comum de representação de imagens é o sistema RGB, onde as imagens são descritas por três canais: vermelho (R - red), verde (G - green) e azul (B - blue) (OTTONI; AL., 2022). No caso de uma imagem que possui informações em diferentes canais ou bandas de frequência, é necessária uma função f(x,y) para cada banda (FILHO; NETO, 1999). Portanto, essas imagens devem ser representadas em três matrizes distintas, como ilustrado na Figura 2.

Figura 2 – Representação de imagens no sistema RGB



Fonte: OTTONI; AL. (2022)

2.2 APRENDIZAGEM DE MÁQUINA

Com o crescente volume de dados gerados diariamente, a capacidade de interpretar e extrair informações relevantes tornou-se um desafio significativo. Em muitos casos, a simples visualização dos dados não é suficiente para compreender padrões ou tendências subjacentes. É nesse contexto que o aprendizado de máquina (machine learning) emerge como uma ferramenta essencial e tem se tornado cada vez mais demandada, especialmente devido à abundância de conjuntos de dados disponíveis (MAHESH et al., 2020). Essas técnicas permitem que indústrias e organizações extraiam insights valiosos de grandes volumes de informações, auxiliando na tomada de decisões e na otimização de processos.

O aprendizado de máquina é um campo em rápido crescimento, destacando-se por seus algoritmos avançados e capacidade de reconhecer padrões complexos por meio de modelos e abordagens estatísticas (TANWAR, 2024). Essa disciplina combina estatística, teoria da informação e computação, formando uma base científica sólida com fundamentos matemáticos robustos (OSISANWO et al., 2017). Essa interdisciplinaridade proporciona ferramentas pode-

rosas capazes de lidar com problemas complexos, como classificação, regressão, clustering e previsão.

Vale destacar que o aprendizado de máquina não se limita a replicar o processo de aprendizagem humano. O aprendizado, nesse contexto, refere-se à identificação de regularidades estatísticas ou padrões nos dados, muitas vezes de forma distinta da abordagem humana (AYODELE, 2010). Embora os algoritmos possam não imitar diretamente a cognição humana, eles oferecem insights sobre a complexidade de diferentes ambientes de aprendizagem, permitindo a criação de soluções adaptativas e eficientes.

Os algoritmos de aprendizado de máquina são organizados em uma taxonomia com base no resultado desejado (AYODELE, 2010). Os principais tipos incluem:

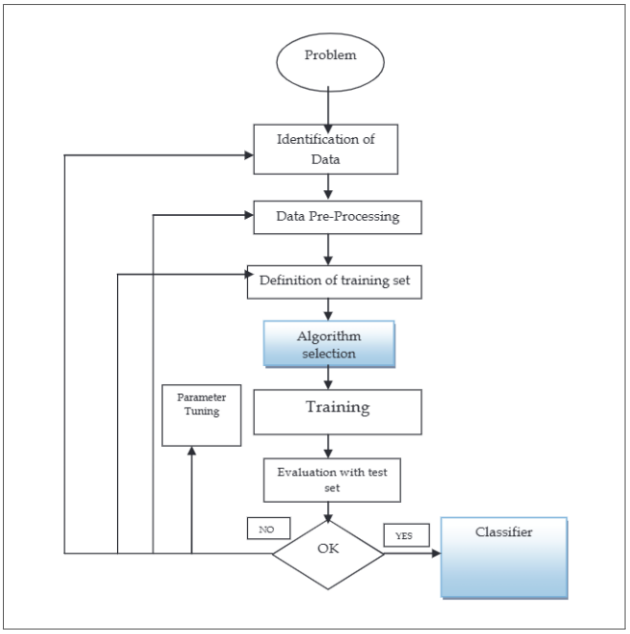
- Aprendizado Supervisionado: Mapeia entradas para saídas, usando exemplos rotulados.
- Aprendizado Não Supervisionado: Identifica padrões em dados sem rótulos.
- Aprendizado por reforço: Aprende ações com base no feedback do ambiente.

Em resumo, o aprendizado de máquina é um processo que depende fundamentalmente de dados, sejam eles textos, imagens ou estatísticas, para funcionar. Esses dados são coletados, preparados e utilizados como base para o treinamento de modelos, que, por sua vez, aprendem a identificar padrões ou fazer previsões de forma autônoma. Os programadores têm o papel crucial de selecionar o modelo adequado, fornecer os dados necessários e ajustar os parâmetros ao longo do tempo, refinando o sistema para que ele entregue resultados cada vez mais precisos e confiáveis (GEORGE, 2024).

No aprendizado de máquina supervisionado, os algoritmos são treinados com base em conjuntos de dados rotulados, onde cada item possui etiquetas que caracterizam suas informações. Uma "chave de respostas"é enviada junto aos dados para orientar os algoritmos sobre como interpretá-los, como no exemplo de fotos de frutas etiquetadas, que permitem ao algoritmo identificar os tipos de frutas em novas imagens (GEORGE, 2024). Essa abordagem é amplamente utilizada para desenvolver modelos de classificação e previsão, sendo a técnica mais comum para treinar redes neurais. No entanto, o modelo depende da disponibilidade completa dos dados de entrada, pois a ausência de alguns valores impossibilita a inferência sobre as saídas. Além disso, o aprendizado de classificação é aplicável a problemas em que deduzir uma classificação é útil e fácil de determinar (AYODELE, 2010). Assim, o aprendizado supervisionado se consolida como uma ferramenta essencial para tarefas que exigem precisão e clareza

na interpretação de dados. O processo de aprendizado de máquina supervisionado é ilustrado na Figura 3.

Figura 3 – Processo de Aprendizado de Máquina Supervisionado



Fonte: AYODELE (2010)

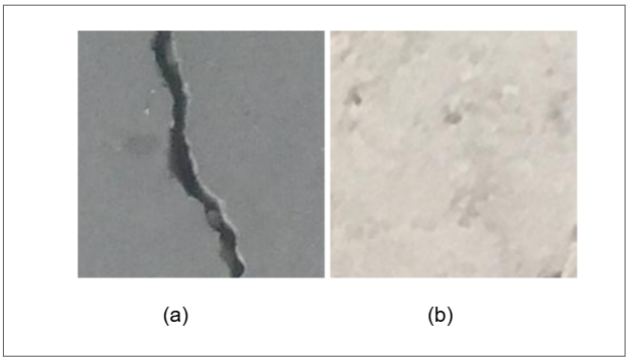
2.3 CLASSIFICAÇÃO DE IMAGENS

Os sistemas de visão computacional e processamento digital de imagens são cruciais para identificar padrões e objetos a partir de percepções visuais (ELGENDY, 2020)). A visão computacional destaca-se por fornecer respostas, seja por predição ou classificação, a partir de dados gráficos como imagens ou vídeos. Para isso, são utilizados dispositivos de processamento da informação equipados com algoritmos de inteligência artificial (OTTONI; AL., 2022).

Por outro lado, o processamento digital de imagens envolve a aplicação de transformações nas imagens, resultando em outra imagem. No entanto processar uma imagem é diferente de interpretá-la (ELGENDY, 2020). Para isso, os algoritmos de machine learning devem incluir uma etapa adicional de visão computacional, permitindo a classificação ou detecção de padrões (OTTONI; AL., 2022). A classificação de imagem é realizada utilizando algoritmos de aprendizado supervisionado de máquina, sendo assim é possível estabelecer conexões entre as propriedades das amostras com os respectivos rótulos. Um exemplo de aplicação dessa técnica na perspectiva da engenharia de estruturas é a classificação para averiguação do estado de

imagens de concreto simples em que a fissura é o principal agente patológico (JUNIOR et al., 2024). A Figura 4 representa um exemplo de imagens de concreto com e sem fissuras.

Figura 4 – Aplicação de classificação de imagens em patologias na construção. (a) Superfície com fissura. (b) Superfície sem patologias aparentes



Fonte: ÖZGENEL (2019)

2.4 REDES NEURAIS CONVOLUCIONAIS

Redes Neurais Convolucionais (CNN) são um tipo de modelo de aprendizado profundo amplamente utilizado para processamento de imagens e reconhecimento de padrões (OTTONI; AL., 2022). Essas redes foram fundamentais para o avanço do deep learning, demonstrando como os conhecimentos adquiridos a partir do estudo do cérebro podem ser efetivamente aplicados em tarefas de aprendizado de máquina (GOODFELLOW; BENGIO; COURVILLE, 2016). As CNNs são organizadas em camadas que processam os dados de entrada e geram classificações na saída. As camadas principais de uma CNN incluem a camada de entrada, camadas convolucionais, camadas de pooling, camadas totalmente conectadas e a camada de saída (BARCELLOS; GONZAGA, 2021). Cada uma dessas camadas desempenha um papel essencial na extração e organização das características da imagem, permitindo que a rede aprenda de forma eficiente padrões complexos.

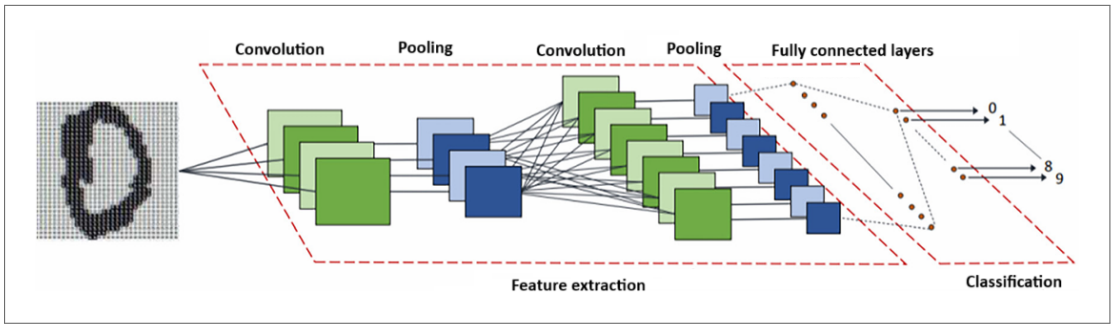
As camadas convolucionais são responsáveis pela extração de características das imagens, desempenhando papel fundamental no processamento das Redes Neurais Convolucionais (CNN) (ELGENDY, 2020). Matematicamente, a convolução é uma operação entre duas funções que resulta em uma terceira função modificada, representada pela equação 2.2 (GOODFELLOW; BENGIO; COURVILLE, 2016):

$$s(t) = (x * w)(t) \tag{2.2}$$

onde t representa o instante de tempo, x é a função de entrada, w é a função de peso, e s é o resultado da convolução. Essa operação permite que a rede identifique padrões relevantes nas imagens, extraindo informações essenciais para tarefas como classificação e detecção de objetos (GOODFELLOW; BENGIO; COURVILLE, 2016).

Entretanto, uma CNN pode também incluir outros tipos de camadas e operações, tais como camadas totalmente conectadas, camadas de pooling e operações de dropout (DUCHI; HAZAN; SINGER, 2011). As camadas totalmente conectadas representam um sistema tradicional de Perceptron Multicamadas (MLP) dentro da arquitetura da CNN (ELGENDY, 2020). A adição de mais camadas convolucionais e de pooling permite que a rede explore padrões cada vez mais específicos. Dessa forma, a saída da última camada de pooling pode ser entendida como uma representação resumida do conjunto de características extraídas da imagem original (PEREIRA et al., 2024). A Figura 5 ilustra a arquitetura de uma Rede Neural Convolucional (CNN).

Figura 5 – Arquitetura de uma CNN, adaptado



Fonte: BURGER; BURGE (2022)

A capacidade das CNNs de extrair automaticamente características relevantes e hierarquicamente organizadas permite a construção de modelos robustos e precisos, especialmente em domínios que demandam alta acurácia (LECUN; BENGIO; HINTON, 2015). Ademais, a combinação de diferentes tipos de camadas e técnicas, como pooling e dropout, contribui para a generalização dos modelos, reduzindo o risco de sobreajuste e ampliando sua aplicabilidade prática em cenários reais (SRIVASTAVA et al., 2014). Dessa forma, as CNNs configuram-se como ferramentas essenciais no avanço das soluções baseadas em inteligência artificial para reconhecimento e classificação de imagens, consolidando seu papel como referência no campo da visão computacional (KRIZHEVSKY; SUTSKEVER; HINTON, 2012).

2.4.1 Modelo Visual Geometry Group (VGG)

A arquitetura VGG foi proposta por Simonyan e Zisserman em 2014, no trabalho seminal "Very Deep Convolutional Networks for Large-Scale Image Recognition"(SIMONYAN; ZISSERMAN, 2014). Seu principal objetivo foi demonstrar que redes convolucionais profundas, utilizando exclusivamente filtros convolucionais pequenos (3×3), poderiam alcançar desempenho superior em tarefas de reconhecimento visual, mantendo uma estrutura simples e regular.

Existem duas versões principais dessa arquitetura: a VGG16 e a VGG19. A VGG16 possui 13 camadas convolucionais e 3 camadas totalmente conectadas, enquanto a VGG19 é composta por 16 camadas convolucionais e 3 camadas totalmente conectadas. A largura das camadas convolucionais (número de canais) é relativamente reduzida, iniciando em 64 canais na primeira camada e dobrando após cada operação de max-pooling, até atingir 512 canais. A função de ativação utilizada em todas as camadas convolucionais é a ReLU (Rectified Linear Unit), e a redução da dimensão espacial é realizada através do max-pooling ao final de cada bloco convolucional. O modelo VGG16 conta com aproximadamente 138 milhões de parâmetros, enquanto o VGG19 possui cerca de 144 milhões (SIMONYAN; ZISSERMAN, 2014).

2.4.2 Modelo Residual Network (ResNet)

A arquitetura ResNet (Residual Network) foi introduzida por He et al. em 2015 com o artigo "Deep Residual Learning for Image Recognition"(HE et al., 2016). O principal avanço dessa arquitetura é a introdução dos blocos residuais, que permitem a construção de redes muito profundas, com centenas de camadas, sem que ocorram problemas severos de degradação do desempenho e do desaparecimento do gradiente. Esses blocos utilizam conexões de atalho (skip connections) que facilitam o fluxo do gradiente durante o treinamento, aumentando a estabilidade e a eficiência do aprendizado (HE et al., 2016).

O modelo ResNet-50, uma das versões mais populares, é composto por 50 camadas que incluem blocos residuais com convoluções 1×1 e 3×3. A arquitetura inicia com uma camada convolucional 7×7, seguida por quatro estágios contendo blocos residuais empilhados, e termina com uma camada de pooling global e uma camada totalmente conectada para classificação (HE et al., 2016).

2.5 INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL

A segmentação semântica e a classificação de imagens impulsionadas por redes neurais convolucionais (CNNs) demonstraram uma capacidade extraordinária de processar grandes volumes de dados, permitindo-lhes abordar tarefas complexas como tradução de linguagem e reconhecimento de objetos com elevada acurácia (HORTA et al., 2021). Entretanto, apesar de seu desempenho impressionante, as CNNs são frequentemente percebidas como "caixas pretas"devido à complexidade e opacidade de seus processos decisórios. Essa falta de transparência dificulta a compreensão de como entradas específicas influenciam as previsões, podendo reduzir a confiança dos usuários e dificultar sua adoção em áreas críticas onde a interpretabilidade é essencial (DHORE et al., 2024).

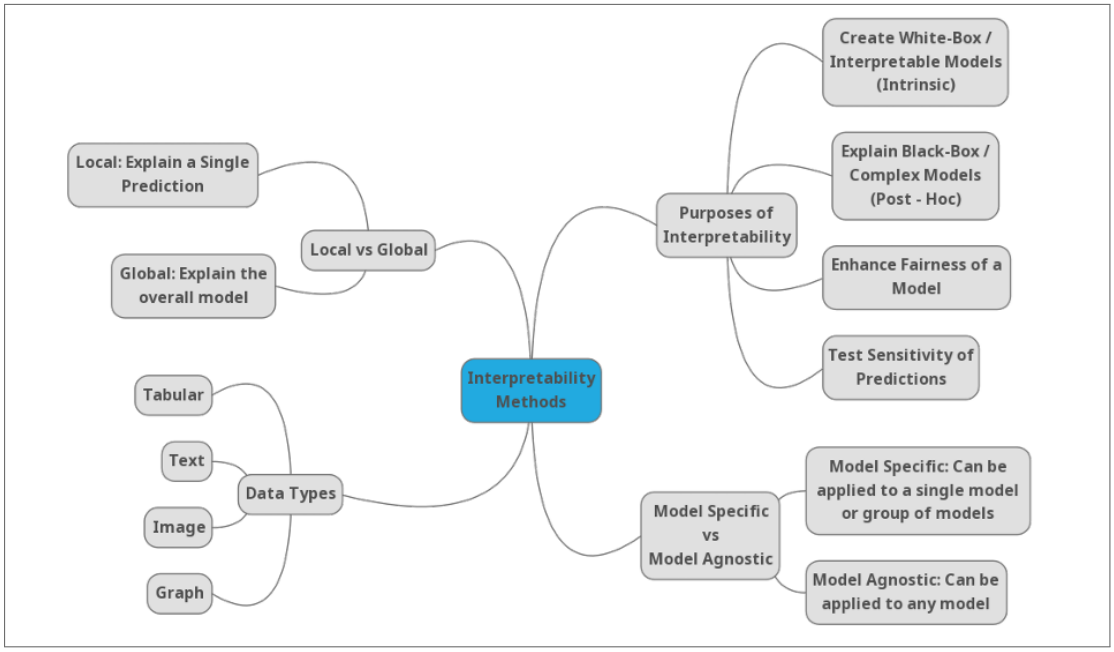
Diante desse desafio, emergiu o campo da Inteligência Artificial Explicável (XAI), oferecendo um conjunto diverso de técnicas destinadas a tornar os modelos de aprendizado de máquina mais compreensíveis. Os métodos de XAI buscam fornecer insights sobre como os modelos operam, esclarecer seus resultados e aumentar a interpretabilidade para os usuários finais, especialmente os decisores humanos (GIPIŠKIS; TSAI; KURASOVA, 2024). Ao invés de tratar a IA como um sistema indecifrável, pesquisas em XAI concentram-se no desenvolvimento de explicações técnicas e funcionais que revelam o funcionamento interno dos algoritmos (BRYAN-KINNS, 2024).

Uma distinção importante na XAI refere-se aos modelos inerentemente interpretáveis, projetados desde o início para serem transparentes, e às explicações pós-hoc, que são aplicadas a modelos já treinados para ajudar os usuários a compreenderem melhor seus processos de decisão (RUDIN, 2019). A Figura 6 ilustra um mapa mental sintetizado que organiza os principais critérios utilizados para classificar técnicas de interpretabilidade em aprendizado de máquina. A consideração desses critérios é fundamental para que pesquisadores e profissionais possam selecionar a abordagem mais apropriada às exigências do seu problema específico(LINARDATOS; PAPASTEFANOPOULOS; KOTSIANTIS, 2020).

Dentre as técnicas post-hoc (métodos aplicados após o treinamento ou a predição do modelo) para classificação de imagens, destacam-se abordagens baseadas em perturbação de pixels, como LIME e SHAP, bem como técnicas baseadas nos pesos internos dos modelos, como Grad-CAM, LRP e Guided Backpropagation (GBP).

O LIME gera explicações locais perturbando regiões da imagem original e criando modelos simplificados e interpretáveis que aproximam localmente o comportamento do modelo original

Figura 6 – Mapa mental taxonômico das técnicas de interpretabilidade em aprendizado de máquina



Fonte: LINARDATOS; PAPASTEFANOPOULOS; KOTSIANTIS (2020)

(RIBEIRO; SINGH; GUESTRIN, 2016). Já o SHAP, fundamentado na teoria dos valores de Shapley, atribui importâncias específicas aos pixels com base em suas contribuições marginais médias para a previsão, oferecendo explicações consistentes e robustas (LUNDBERG; LEE, 2017). Por outro lado, o Grad-CAM utiliza os gradientes das ativações internas das CNNs para criar mapas visuais que destacam regiões importantes da imagem que influenciaram significativamente na decisão do modelo (SELVARAJU et al., 2016; SELVARAJU et al., 2017).

Embora esses métodos sejam valiosos para melhorar a confiança nas previsões, mensurar objetivamente a qualidade das explicações em modelos visuais permanece um desafio significativo. Tal dificuldade surge devido à falta de métricas universalmente aceitas que avaliem simultaneamente fidelidade, relevância e compreensibilidade das explicações geradas (SHAH; SHEPPARD, 2020; HOOKER et al., 2019; GILPIN et al., 2018).

Dessa forma, a XAI desempenha um papel crucial não apenas em melhorar a transparência e interpretabilidade das decisões baseadas em IA, mas também em facilitar a adoção responsável e segura dessas tecnologias em áreas críticas como medicina, segurança pública e engenharia (GILPIN et al., 2018).

2.6 ALGORITMOS DE SEGMENTAÇÃO DE IMAGENS

A segmentação de imagens é uma etapa fundamental no processamento de imagens que visa particionar a imagem em regiões homogêneas, atribuindo rótulos semânticos a cada pixel de uma imagem, facilitando a análise e interpretação de estruturas relevantes, permitindo a compreensão detalhada da cena analisada (REAUNGAMORNRAT et al., 2022; DONG et al., 2025). Nos últimos anos, diversos estudos exploraram a aplicação da segmentação semântica em variados domínios, evidenciando seu potencial para tarefas complexas em áreas como medicina, agricultura e engenharia (MARQUES et al., 2024; FERNANDES et al., 2023).

Entre os métodos clássicos utilizados na segmentação, destacam-se os algoritmos baseados em agrupamento, como o k-means. Este algoritmo particiona um conjunto de dados em k clusters com base na proximidade dos pontos aos centróides pré-definidos. Inicialmente, o número de clusters k é selecionado, e os centróides são inicializados aleatoriamente. Em cada iteração, os pontos são atribuídos ao centróide mais próximo, que é então recalculado pela média dos pontos associados. Esse processo iterativo prossegue até que um critério de convergência seja satisfeito (AKHTAR; AHMAD, 2015; QURESHI; AHAMAD, 2018).

Dentre os diversos métodos, os algoritmos baseados em agrupamento (clustering) destacam-se por sua simplicidade e eficácia, especialmente em contextos não supervisionados. O algoritmo k-means, introduzido originalmente por MacQueen em 1967 (MACQUEEN, 1967). Ele particiona os pixels em K grupos, minimizando a soma das distâncias quadráticas entre cada pixel e o centróide do seu grupo, com base em características como cor, intensidade e textura.

O funcionamento do k-means consiste em inicializar K centróides, atribuir cada pixel ao centróide mais próximo e atualizar iterativamente as posições dos centróides até a convergência do critério de minimização.

Apesar de sua popularidade, o algoritmo k-means apresenta uma série de limitações que comprometem sua aplicabilidade em tarefas complexas de segmentação de imagens. Entre os principais obstáculos, destacam-se:

- **Sensibilidade à inicialização dos centróides:** diferentes inicializações podem levar a soluções distintas, uma vez que o algoritmo tende a convergir para mínimos locais, afetando a reprodutibilidade dos resultados.
- **Assunção de clusters esféricos e de variância semelhante:** o k-means pressupõe que os dados formam grupos aproximadamente circulares e com densidade homogênea, o

que raramente se verifica em imagens reais com regiões de morfologia e textura variáveis.

- **Fronteiras de decisão lineares:** por utilizar distância euclidiana como métrica de agrupamento, o algoritmo induz divisões lineares no espaço de características, dificultando a segmentação de regiões com separações não lineares ou contornos complexos.
- **Baixa robustez a ruídos e outliers:** a média aritmética utilizada para atualização dos centróides é sensível a valores extremos, o que pode distorcer significativamente o agrupamento em presença de ruídos estruturais ou artefatos nas imagens.
- **Desconsideração da informação espacial:** o algoritmo agrupa pixels com base apenas em suas propriedades espectrais (como cor ou intensidade), desconsiderando a posição espacial dos mesmos, o que pode resultar em segmentações fragmentadas ou semanticamente inconsistentes.

Nesse sentido, trabalhos futuros podem considerar a adoção de abordagens mais robustas, como algoritmos de agrupamento baseados em densidade, métodos espectrais ou técnicas baseadas em aprendizado profundo, que incorporam características espaciais e morfológicas com maior expressividade.

Apesar dessas limitações, o algoritmo k-means demonstrou capacidade de resolver o problema proposto com precisão satisfatória, sobretudo devido à baixa complexidade e heterogeneidade restrita da base de dados utilizada, que favoreceu a formação de clusters bem definidos. Além disso, por se tratar de uma abordagem não supervisionada, o método dispensou a necessidade de conjuntos de treinamento previamente rotulados, o que constitui vantagem prática em cenários com restrições de anotação. Outro fator relevante refere-se ao seu reduzido uso de memória RAM durante a execução, resultante de sua estrutura computacional simples baseada em operações vetoriais e iterativas de baixa complexidade, tornando-o uma solução eficiente para o problema em questão, considerando as limitações de processamento disponíveis, especialmente no que diz respeito à alocação de memória.

3 METODOLOGIA

Este capítulo descreve de forma detalhada os procedimentos metodológicos adotados para o desenvolvimento deste estudo, incluindo a caracterização dos dados utilizados, a definição dos modelos implementados, as técnicas empregadas para treinamento e validação, bem como as métricas de avaliação utilizadas para mensurar o desempenho do método proposto.

3.1 BASE DE DADOS

O presente estudo utiliza o conjunto de dados denominado Concrete Crack Images for Classification (ÖZGENEL, 2019), composto por imagens de superfícies de concreto com e sem fissuras, coletadas em diversas edificações no campus da Middle East Technical University (METU). Este dataset é amplamente empregado em pesquisas voltadas à detecção automática de falhas estruturais por meio de redes neurais convolucionais (CNNs), dada sua extensão, qualidade e relevância prática.

O conjunto contém 40.000 imagens divididas igualmente entre duas classes: com fissura (positiva) e sem fissura (negativa), totalizando 20.000 amostras por categoria. As imagens possuem resolução de 227 $\times$ 227 pixels e são codificadas em três canais (RGB). A base foi construída a partir de 458 imagens de alta resolução (4032 $\times$ 3024 pixels), representando uma diversidade significativa em termos de textura superficial e condições de iluminação, elementos essenciais para avaliar a robustez de algoritmos de visão computacional em ambientes reais.

É importante destacar que, na versão original do conjunto, não foram aplicadas técnicas de aumento de dados (data augmentation), como rotações aleatórias, espelhamento ou ruído sintético. A ausência dessas técnicas pode limitar a capacidade de generalização do modelo em contextos mais variáveis (SHORTEN; KHOSHGOFTAAR, 2019). Para padronização e redução da complexidade computacional, as imagens foram convertidas para escala cinza e redimensionadas para 128 $\times$ 128 pixels.

Cada imagem foi rotulada de acordo com a presença ou ausência de fissuras e organizada em um formato estruturado. Para evitar vieses na distribuição dos dados e garantir aleatoriedade no processo de aprendizagem, neste estudo as imagens foram embaralhadas antes de serem separadas em variáveis de entrada (X) e rótulos de saída (y), conforme recomendado em práticas modernas de aprendizado de máquina (GOODFELLOW et al., 2016).

Após o pré-processamento, os dados foram convertidos em arrays NumPy e formatados para o padrão exigido pela biblioteca TensorFlow: (número de imagens, 128, 128, 1). Em seguida, o conjunto de dados foi armazenado em um arquivo HDF5 com compressão gzip, visando otimização de espaço e eficiência no carregamento durante as etapas de treinamento e validação do modelo.

Os experimentos utilizando as bandas no espaço de cores RGB não puderam ser realizados devido a limitações técnicas relacionadas à capacidade de memória RAM disponível. O processamento simultâneo das três bandas exigiu recursos computacionais superiores aos suportados pelo ambiente de execução, inviabilizando a realização das análises propostas nesse espaço de cores.

3.2 MODELOS DE CLASSIFICAÇÃO

Para a detecção de fissuras em estruturas de concreto, este trabalho emprega uma abordagem baseada em Redes Neurais Convolucionais (CNNs). Foram utilizadas versões adaptadas das arquiteturas VGG16, VGG19 (SIMONYAN; ZISSERMAN, 2014) e ResNet50 (HE et al., 2016), amplamente reconhecidas e consolidadas na literatura para tarefas de classificação de imagens.

O conjunto de dados empregado consiste em imagens de superfícies de concreto, tanto com presença quanto sem presença de fissuras, armazenadas no formato HDF5. Previamente ao treinamento, as imagens foram normalizadas para o intervalo [0,1], mediante a divisão dos valores dos pixels por 255, o que favorece a estabilidade numérica durante o processo de otimização.

As arquiteturas foram construídas com base em pesos pré-treinados no conjunto de dados ImageNet (DENG et al., 2009), de modo a aproveitar as representações de baixo nível previamente aprendidas. Durante o treinamento, as camadas convolucionais das arquiteturas originais (VGG16, VGG19 e ResNet50) permaneceram congeladas, ou seja, seus pesos não foram atualizados. Essa estratégia visa preservar as características extraídas de domínios genéricos, permitindo que apenas as camadas adicionais aprendam os padrões específicos do problema de detecção de fissuras.

A arquitetura final dos modelos foi composta pelos seguintes componentes:

- Camadas convolucionais pré-treinadas e não treináveis da arquitetura escolhida (VGG16, VGG19 ou ResNet50);

- Uma camada de achatamento (flattening) para converter a saída das camadas convolucionais em um vetor unidimensional;
- Uma camada densa com 256 neurônios e função de ativação ReLU, responsável por introduzir não linearidade ao modelo;
- Uma camada de saída com um único neurônio e função de ativação sigmoide, apropriada para tarefas de classificação binária (presença ou ausência de fissura).

O modelo foi compilado com a função de perda *binary crossentropy* e otimizado via *Adam*, com taxa de aprendizado de $\eta = 10^{-3}$. O treinamento foi realizado por 10 épocas, com tamanho de lote (*batch size*) igual a 16 para a arquitetura VGG e 32 para a ResNet, utilizando divisão de 70% para treinamento e 30% para validação. As imagens de entrada foram normalizadas para o intervalo [0,1] e adaptadas ao formato RGB, com dimensões de $128 \times 128 \times 3$ pixels. Os experimentos foram executados no ambiente L4 do Google Colab, que oferece uma infraestrutura computacional composta por uma NVIDIA L4 Tensor Core GPU com 22.5GB de memória dedicada e 53GB de memória RAM, garantindo desempenho adequado para treinamento de redes neurais convolucionais profundas.

3.3 MODELO DE SEGMENTAÇÃO

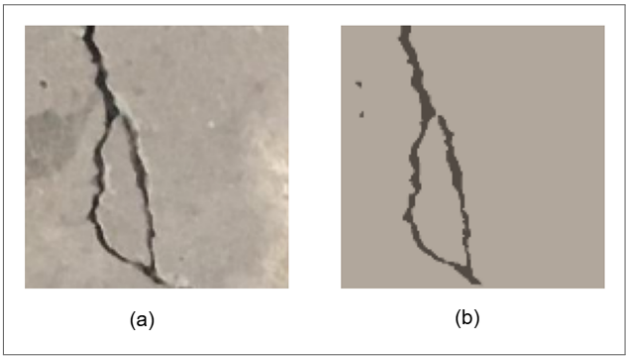
Neste trabalho, o algoritmo de segmentação não supervisionada k-means foi empregado com o objetivo de realizar a detecção de fissuras em imagens de superfícies estruturais. A técnica foi aplicada para segmentação binária das imagens, distinguindo pixels pertencentes a regiões de trincas daqueles que compõem o fundo. Para garantir uniformidade na análise, todas as imagens de entrada foram previamente redimensionadas para uma resolução fixa de 128×128 pixels. Em seguida, cada imagem foi convertida para um array tridimensional do tipo NumPy, no qual cada pixel é representado por três canais de cor (vermelho, verde e azul).

A segmentação foi realizada considerando $k=2$, de forma a dividir os pixels em dois agrupamentos distintos. Após a execução do algoritmo, os rótulos atribuídos a cada pixel foram convertidos em valores binários: 0 (preto) representando o fundo e 255 (branco) representando as regiões segmentadas como trincas. Essa abordagem permitiu a obtenção de máscaras binárias que destacam automaticamente as fissuras nas superfícies analisadas.

Optou-se pelo uso do k-means em virtude de sua natureza não supervisionada e baixo custo computacional, o que o torna uma alternativa viável para tarefas de detecção preliminar em contextos com recursos computacionais limitados, além de dispensar a necessidade de um conjunto de dados rotulado para treinamento. Estudos recentes têm demonstrado o potencial do k-means em tarefas de segmentação em diferentes domínios, incluindo inspeção visual automatizada e análise de materiais (DONG et al., 2025; QURESHI; AHAMAD, 2018; MARQUES et al., 2024; FERNANDES et al., 2023; AKHTAR; AHMAD, 2015)

Um exemplo visual do resultado da segmentação binária obtida com o algoritmo k-means, aplicado à detecção de trincas em uma imagem de concreto, é apresentado na Figura 7. Esta ilustração evidencia a capacidade do método em isolar automaticamente as regiões fissuradas a partir da distribuição de cores na imagem original.

Figura 7 – Segmentação de Fissura com K-means



Fonte: Autor

3.4 MÉTODOS DE EXPLICABILIDADE

No presente estudo foram aplicadas técnicas de Inteligência Artificial Explicável (XAI) com o propósito de interpretar e tornar compreensíveis as decisões tomadas por modelos de redes neurais convolucionais em tarefas de classificação de imagens. Serão considerados cinco algoritmos representativos de abordagens distintas no campo da XAI: SHAP, LIME, Grad-CAM, LRP e GBP.

Os métodos selecionados contemplam tanto técnicas baseadas em perturbação dos dados de entrada quanto abordagens que exploram diretamente os pesos e gradientes internos do modelo. Essa diversidade metodológica permite uma análise comparativa entre diferentes paradigmas de explicação, considerando tanto a perspectiva da variação da entrada quanto a da estrutura interna da rede.

Nas subseções a seguir, são apresentados os fundamentos técnicos de cada um dos métodos de interpretabilidade considerados neste trabalho. Os procedimentos de implementação correspondentes serão detalhados no capítulo 4, com ênfase na sua aplicação ao contexto da engenharia civil, onde a explicabilidade dos modelos é um requisito fundamental para assegurar a confiabilidade e a validação técnica das decisões automatizadas.

3.4.1 Abordagens baseadas em perturbação de pixels

As técnicas de Inteligência Artificial Explicável (XAI) baseadas em perturbações de pixels representam uma classe distinta de métodos voltados à interpretação de redes neurais profundas, notadamente em tarefas de visão computacional. Diferentemente das abordagens que analisam os gradientes internos dos modelos, os métodos por perturbação operam diretamente na entrada do sistema, modificando sistematicamente os pixels de uma imagem a fim de avaliar o impacto dessas alterações sobre a saída do modelo. Tal estratégia visa inferir a importância de regiões específicas da entrada ao correlacionar variações na predição com as perturbações introduzidas (ZEILER; FERGUS, 2014).

A lógica subjacente a essas técnicas está na premissa de que, ao suprimir ou modificar partes da imagem original é possível identificar quais regiões são mais relevantes para a decisão da rede. Este tipo de abordagem fornece uma forma intuitiva e visualmente acessível de explicação, o que é particularmente valioso em aplicações sensíveis (RIBEIRO; SINGH; GUESTRIN, 2016).

Assim, métodos baseados em perturbações de pixel ocupam um papel complementar às técnicas baseadas em gradientes, oferecendo meios acessíveis de interpretação que não requerem conhecimento intrusivo sobre a estrutura interna do modelo.

3.4.1.1 Local Interpretable Model-agnostic Explanations (LIME)

O método LIME, originalmente proposto por Ribeiro, Singh e Guestrin (2016), é uma técnica voltada à geração de explicações locais para modelos preditivos complexos, tratando-os como caixas-pretas (RIBEIRO; SINGH; GUESTRIN, 2016). Embora sua formulação seja genérica e aplicável a diversos tipos de dados, seu uso tem sido adaptado com eficácia para o domínio de imagens, com o objetivo de identificar visualmente quais regiões de uma imagem contribuíram de forma mais significativa para uma determinada predição.

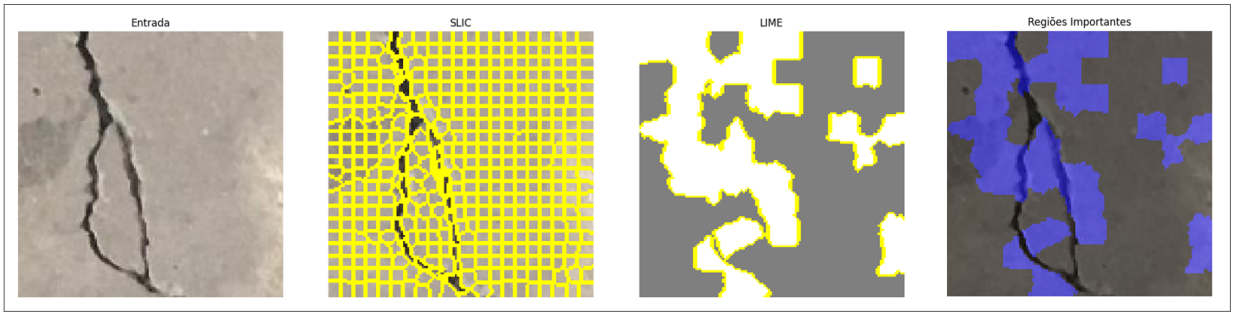
No contexto da classificação de imagens, o LIME atua segmentando a imagem original em

regiões coesas denominadas superpixels, utilizando algoritmos de segmentação como o SLIC (Simple Linear Iterative Clustering)(ACHANTA et al., 2012). Em seguida, a imagem é perturbada por meio da ativação e desativação aleatória desses superpixels, normalmente substituindo as regiões desativadas por uma cor neutra ou média da imagem, o que gera um conjunto de variações artificiais da entrada original. Cada versão perturbada da imagem é processada pelo modelo original, que retorna as respectivas probabilidades associadas às classes previstas (RIBEIRO; SINGH; GUESTRIN, 2016; DAI et al., 2021).

A explicação gerada pelo LIME é tipicamente apresentada como uma visualização em que os superpixels mais relevantes para a decisão do modelo são destacados sobre a imagem original. Essa representação visual tem sido eficaz em aplicações críticas, como diagnóstico por imagem e análise de defeitos visuais, contribuindo para uma melhor compreensão do comportamento do modelo por especialistas humanos (SAMEK; MÜLLER, 2019).

A Figura 8 apresenta uma visualização resultante da aplicação do LIME, destacando os superpixels mais relevantes para a predição. As regiões evidenciadas indicam as áreas da imagem que mais influenciaram a decisão do modelo, facilitando sua interpretação.

Figura 8 – Mapa explicativo do LIME evidenciando as regiões da imagem que mais contribuíram para a classificação realizada pela rede neural.



Fonte: Autor

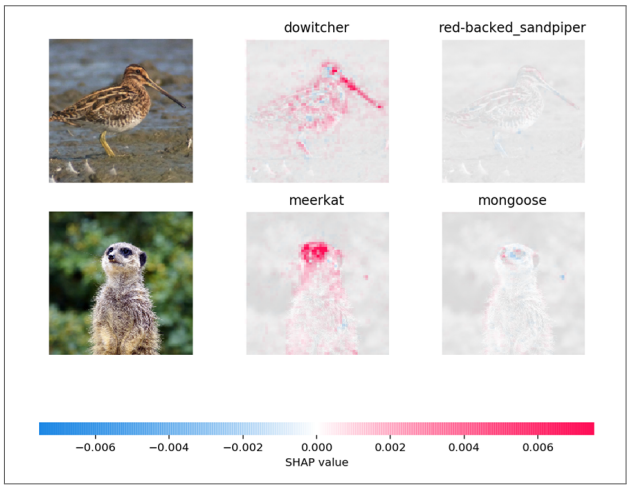
3.4.1.2 Shapley Additive Explanations

O método SHAP constitui uma abordagem de interpretabilidade fundamentada na Teoria dos Jogos, cuja proposta central é atribuir, de forma consistente e localmente precisa, valores de importância às variáveis de entrada com base em sua contribuição marginal para a predição de um modelo. Introduzido por Lundberg e Lee (2017), o SHAP formaliza uma classe de métodos aditivos de atribuição de características, sendo o único, segundo os autores, a satisfazer simultaneamente as propriedades de consistência e exatidão local (LUNDBERG; LEE, 2017).

Aplicado ao domínio da classificação de imagens, o SHAP adapta seu arcabouço formal para lidar com variáveis espaciais contínuas, como pixels ou superpixels. A metodologia consiste na geração de um conjunto de instâncias perturbadas da imagem original por meio da ativação seletiva de segmentos, simulando a presença ou ausência de informações visuais. As predições correspondentes são então utilizadas para estimar os valores de Shapley, calculando a contribuição marginal de cada segmento à saída do modelo em diferentes coalizões de entrada.

A saída gerada pelo SHAP assume geralmente a forma de um mapa de importância visual, no qual regiões com contribuições positivas ou negativas para a classe predita são evidenciadas (LUNDBERG et al., 2018).A Figura 9 ilustra a predições para duas imagens de entrada, os pixels em vermelho indicam valores de SHAP positivos que aumentam a probabilidade da classe, enquanto os pixels em azul representam valores negativos que reduzem essa probabilidade.

Figura 9 – Visualização das predições explicadas pelo SHAP para duas imagens de entrada.



Fonte: LUNDBERG et al. (2018)

3.4.2 Abordagem baseadas nos pesos internos dos modelos

As abordagens de Inteligência Artificial Explicável (XAI) fundamentadas na análise dos gradientes internos dos modelos constituem uma importante linha de pesquisa voltada à interpretação de redes neurais profundas, especialmente em aplicações de visão computacional que fazem uso extensivo de redes neurais convolucionais (CNNs). O crescente interesse por essas técnicas está associado à complexidade e opacidade dos modelos modernos de aprendizado profundo, os quais, embora altamente precisos, carecem de transparência em seus processos decisórios (LINARDATOS; PAPASTEFANOPOULOS; KOTSIANTIS, 2020; SAMEK; MÜLLER, 2019).

Esses métodos têm sua origem na necessidade de compreender de que maneira as entradas influenciam as saídas de um modelo, a partir da observação do comportamento interno da rede. Para isso, utilizam os gradientes calculados durante o processo de retropropagação, que indicam o quanto a saída é sensível às ativações de determinadas camadas intermediárias. Por meio dessa análise, torna-se possível identificar os componentes internos que mais contribuem para uma determinada predição, como filtros convolucionais ou agrupamentos de neurônios (MONTAVON et al., 2017).

Essas abordagens foram inicialmente desenvolvidas no contexto da visão computacional, com destaque para os saliency maps propostos por Simonyan et al. (2013), que calculam o gradiente da saída do modelo com respeito à entrada para identificar os pixels mais sensíveis à decisão preditiva (SIMONYAN; VEDALDI; ZISSERMAN, 2013).

A evolução desse paradigma resultou em técnicas mais refinadas, como o Grad-CAM (Gradient-weighted Class Activation Mapping) (SELVARAJU et al., 2016), que se tornou uma das abordagens mais utilizadas para gerar visualizações localizadas da atenção do modelo em classificações de imagem.

3.4.2.1 Gradient-weighted Class Activation Mapping (Grad-CAM)

O Grad-CAM é uma técnica amplamente utilizada para gerar mapas de localização discriminativos de classe em redes neurais convolucionais (CNNs). Essa abordagem se enquadra nas técnicas baseadas em informações internas do modelo, discutidas anteriormente, como parte das metodologias model-specific aplicadas à explicação de predições em visão computacional (SELVARAJU et al., 2016; SIMONYAN; VEDALDI; ZISSERMAN, 2013).

Para uma arquitetura genérica baseada em CNNs, o mapa de localização discriminativo de classe, denotado por $L^c_{\text{Grad-CAM}} \in R^{u \times v}$, é calculado a partir dos seguintes passos (SELVARAJU et al., 2016):

Primeiramente, calcula-se o gradiente da saída do modelo para a classe alvo y^c em relação aos mapas de ativação A de uma camada convolucional selecionada, representado como $\frac{\partial y^c}{\partial A^k_{ij}}$. Esses gradientes são então agregados por média global, produzindo os pesos α^c_k, definidos pela equação 3.1 :

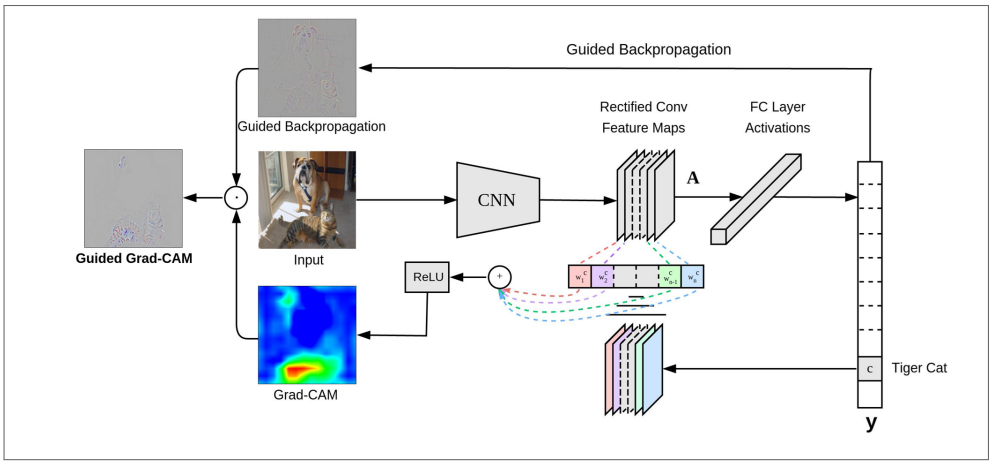
$$\alpha^c_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}} \tag{3.1}$$

Em seguida, o mapa de localização é obtido por meio expressão representada pela equação 3.2:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha^c_k A^k\right) \tag{3.2}$$

A função ReLU atua filtrando os valores negativos, de modo a preservar apenas as regiões mais relevantes para a decisão do modelo. Isso resulta em mapas visuais que destacam as áreas mais responsáveis pela predição, sendo uma representação coerente com os mecanismos internos das CNNs e altamente interpretável por humanos.

Figura 10 – Esquema do Guided Grad-CAM, que combina Grad-CAM com Guided Backpropagation para gerar visualizações de alta resolução das regiões da imagem mais relevantes para a predição



Fonte: SELVARAJU et al. (2017)

3.4.2.2 Layer-wise Relevance Propagation (LRP)

O objetivo do LRP é atribuir a cada pixel da imagem de entrada do modelo uma pontuação de relevância que quantifique sua contribuição para a predição realizada. Essa redistribuição da relevância é feita em cada uma das camadas, seguindo o fluxo da saída até a entrada e respeitando uma regra de conservação que garante a interpretabilidade local da explicação.

Formalmente, a relevância R_j de um neurônio j na camada l é obtida a partir das relevâncias R_k dos neurônios k na camada seguinte $l+1$ por meio da equação 3.4.2.2 (BACH et al., 2015)

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{j'} a_{j'} w_{j'k} + \epsilon \cdot \text{sign}(\sum_{j'} a_{j'} w_{j'k})} R_k$$

(3.3)

onde a_j é a ativação do neurônio j, w_{jk} é o peso da conexão entre os neurônios j e k, e ϵ é um fator de regularização que previne instabilidades numéricas.

A principal vantagem do LRP sobre métodos puramente baseados em gradientes é sua robustez frente a redes profundas com funções de ativação não lineares. Segundo Montavon et al.(MONTAVON; SAMEK; MÜLLER, 2018), o LRP é eficaz na geração de mapas de relevância que são compatíveis com a semântica da tarefa, sendo aplicável em diversas áreas.

Essa abordagem também permite interpretações mais precisas sobre o comportamento interno dos modelos, ao preservar as contribuições positivas e negativas de maneira explícita. Samek et al (SAMEK; WIEGAND; MÜLLER, 2017) destacam que o LRP oferece maior fidelidade às decisões do modelo em comparação com visualizações baseadas unicamente em gradientes de primeira ordem.

3.4.2.3 Guided Backpropagation (GBP)

Guided Backpropagation (GBP) é uma técnica de explicabilidade introduzida por Springenberg et al. (SPRINGENBERG et al., 2014) com o intuito de gerar visualizações mais interpretáveis dos padrões de ativação em redes convolucionais profundas. O método consiste em uma modificação do algoritmo de retropropagação tradicional, com a introdução de uma restrição que zera os gradientes negativos tanto na propagação para trás quanto nas ativações ReLU.

Durante o processo de retropropagação, o gradiente de entrada é modulado por dois filtros: um sobre os gradientes e outro sobre as ativações, resultando na regra demonstrada na equação 3.4.2.3.

$$\delta_{\mathsf{GBP}}(x) = \delta(x > 0) \cdot \delta_{\mathsf{ReLU}}(x > 0)$$

(3.4)

Com isso, são preservadas apenas as contribuições positivas, eliminando ruídos e destacando estruturas visuais que são realmente discriminativas para a classe predita. De acordo

com Zeiler e Fergus (ZEILER; FERGUS, 2014), o GBP gera mapas visuais de alta resolução que facilitam a interpretação humana, sendo especialmente úteis em tarefas como classificação de imagens e detecção de objetos.

Apesar de sua simplicidade, o GBP é uma técnica dependente do modelo (*model-specific*), pois requer acesso direto à arquitetura e aos pesos da rede. Adebayo et al.(ADEBAYO et al., 2018) alertam, entretanto, que métodos baseados em gradientes, como o GBP, devem ser avaliados criticamente por meio de testes de sanidade, já que podem ser sensíveis a alterações na arquitetura do modelo sem refletir mudanças reais nas decisões do sistema.

4 MÉTODO PROPOSTO

Neste capítulo, apresenta-se o método proposto para a aplicação de técnicas de Inteligência Artificial Explicável (XAI) em modelos de redes neurais convolucionais utilizados em tarefas de classificação de imagens com fissuras em concreto armado. Serão detalhados, de forma sistemática, os procedimentos adotados para a implementação dos algoritmos explicativos selecionados, abrangendo tanto as abordagens baseadas em perturbação dos dados de entrada quanto aquelas fundamentadas nos pesos e gradientes das camadas internas dos modelos.

Inicialmente, são descritos os passos necessários para a preparação dos dados e o carregamento do modelo preditivo previamente treinado. Na sequência, são abordados os métodos que exploram a estrutura interna do modelo (LRP, GBP e Grad-CAM), com ênfase no Grad-CAM, que utiliza informações dos mapas de ativação e dos gradientes para gerar visualizações localmente discriminativas. Em seguida, apresenta-se o fluxo de execução para os métodos baseados em perturbação, como SHAP e LIME, os quais operam por meio da modificação controlada das entradas para inferir a importância de diferentes regiões da imagem.

O objetivo deste capítulo é fornecer uma descrição clara e reprodutível das etapas envolvidas na geração das explicações, de modo a possibilitar sua replicação em aplicações reais e facilitar a análise interpretativa dos resultados por especialistas da área.

4.1 ABORDAGEM BASEADAS NOS PESOS INTERNOS DOS MODELOS

Os métodos Layer-wise Relevance Propagation (LRP) e Guided Backpropagation (GBP) não requerem aplicação individual em cada camada convolucional do modelo, pois operam de forma global sobre toda a arquitetura. O LRP propaga a relevância da saída do modelo até a entrada, respeitando uma regra de conservação ao longo de todas as camadas, enquanto o GBP retropropaga os gradientes modificados desde a saída até a entrada, suprimindo gradientes negativos nas ativações ReLU. Aplicar esses métodos isoladamente em camadas intermediárias comprometeria sua integridade conceitual, uma vez que ambos dependem da cadeia completa de ativação e propagação para fornecer explicações consistentes e semanticamente interpretáveis.

Os mapas de saliência gerados pelos métodos de explicabilidade Layer-wise Relevance Propagation (LRP) e Guided Backpropagation (GBP), durante o desenvolvimento experimental,

apresentaram baixo contraste entre as regiões de interesse (fissura) e o fundo, presença de ruído textural, natureza difusa e pouco estruturada das ativações produzidas por ambos os métodos. Tais características comprometeram a aplicação de máscara de referência com precisão suficiente para gerar máscaras confiáveis. Em razão disso, optou-se por não calcular métricas quantitativas baseadas em segmentação, como o F1-Score, que requer a sobreposição entre a segmentação prevista e uma máscara de referência para avaliar acurácia. Assim, os resultados visuais obtidos por LRP e GBP estão apresentados no capítulo de resultados para fins qualitativos, sem, contudo, a correspondente avaliação numérica via F1-Score.

Diferentemente do LRP e do GBP, o Grad-CAM (Gradient-weighted Class Activation Mapping) opera com base na análise de mapas de ativação espacial oriundos de camadas convolucionais específicas, combinando essas ativações com os gradientes da classe-alvo. Essa abordagem permite que o Grad-CAM seja aplicado diretamente em diferentes camadas convolucionais da rede, produzindo mapas de saliência que refletem diferentes níveis de abstração. Camadas mais profundas, por exemplo, tendem a capturar padrões semânticos mais globais, enquanto camadas iniciais preservam detalhes locais e texturais. Assim, a escolha da camada em que o Grad-CAM é aplicado influencia diretamente a granularidade e o tipo de informação destacada na explicação visual, o que permite uma análise mais flexível da hierarquia representacional do modelo.

Neste sentido, este trabalho propõe uma abordagem híbrida de interpretabilidade para redes neurais convolucionais aplicadas à detecção de fissuras em imagens. A proposta consiste na combinação entre o Grad-CAM (Gradient-weighted Class Activation Mapping), que fornece mapas de ativação discriminativos baseados nos gradientes internos do modelo, e o algoritmo de segmentação K-means, que gera regiões com contornos bem definidos. Essa combinação visa superar limitações de cada método isolado: enquanto o Grad-CAM destaca regiões relevantes para a predição, ele pode produzir mapas difusos; por outro lado, o K-means oferece boa delimitação espacial, mas não reflete diretamente o raciocínio do modelo.

Dessa forma, as regiões de maior ativação identificadas pelo Grad-CAM são refinadas por meio da segmentação com K-means, resultando em uma explicação visual mais detalhada. Essa abordagem proporciona uma representação mais precisa e interpretável das fissuras, ampliando a confiaça nas prediões do modelo por meio de evidência visual estruturada.

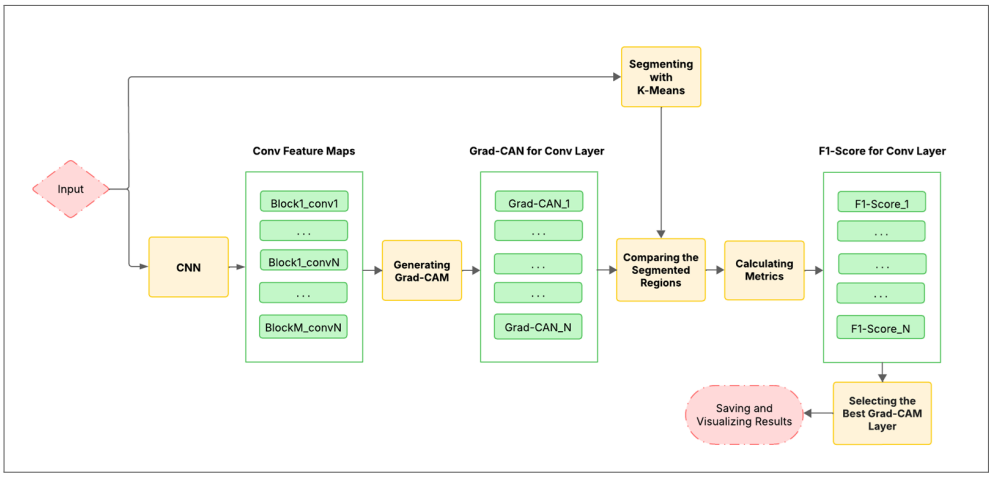
Além disso, considerando que métodos de explicabilidade muitas vezes carecem de critérios objetivos para avaliar a qualidade das explicações, propõe-se o uso de uma métrica quantitativa para mensurar a sobreposição entre os mapas gerados e as regiões segmentadas, promovendo

uma análise mais robusta da interpretabilidade.

O processo metodológico proposto pode ser descrito em sete etapas principais, conforme ilustrado na Figura 11:

- Carregamento do Modelo e Preparação das Imagens
- Geração do Grad-CAM
- Segmentação com K-Means
- Comparacão das regiões segmentadas
- Cálculo de métricas
- Seleção da melhor camada do Grad-CAM
- Salvamento e visualização dos resultados

Figura 11 – Visão Geral do Método Proposto Para Algoritmos que Utilizam os Pesos internos dos Modelos



Fonte: Autor

A implementação do método proposto foi desenvolvida em linguagem Python, utilizando bibliotecas amplamente consolidadas para processamento de dados e aprendizado de máquina, tais como NumPy, TensorFlow, OpenCV e Scikit-Learn. As imagens são processadas em lotes, e para cada imagem de entrada executa-se um pipeline completo composto pelas seguintes etapas: geração do mapa Grad-CAM, segmentação via algoritmo K-means e cálculo de métricas de similaridade entre as regiões segmentadas e as regiões ativadas.

A seleção da camada convolucional mais representativa é realizada com base na maximização de uma métrica de sobreposição, assegurando que a camada escolhida reflita, de forma

mais informativa, as regiões relevantes para a predição. Os resultados finais são armazenados em formatos visuais e numéricos, possibilitando tanto a análise qualitativa das regiões explicadas quanto a avaliação quantitativa da qualidade da interpretação gerada.

As subseções a seguir apresentam, de forma detalhada, cada uma das etapas que compõem o método proposto.

4.1.1 Carregamento do Modelo e Preparação das Imagens

Nesta etapa, realiza-se o carregamento do modelo de aprendizado profundo desenvolvido neste trabalho. O modelo, baseado em uma arquitetura de rede neural convolucional, foi treinado previamente utilizando a base de dados de fissuras descrita na Seção 3.1, com o objetivo de realizar a classificação automática das imagens.

As imagens utilizadas no processo de explicabilidade são recuperadas a partir de um arquivo em formato HDF5, no qual foram previamente armazenadas após o treinamento. Em seguida, as imagens são normalizadas, assegurando que os valores dos pixels estejam compatíveis com a faixa esperada pela rede neural. Esse procedimento garante a integridade do fluxo de entrada e a consistência dos resultados nas etapas posteriores de geração de explicações visuais e segmentação.

4.1.2 Geração do Grad-CAM

Após o carregamento das imagens e do modelo, aplica-se a técnica Grad-CAM (Gradient-weighted Class Activation Mapping), a qual tem por objetivo destacar as regiões mais relevantes de uma imagem no processo decisório da rede neural convolucional. Essa técnica baseia-se na análise dos gradientes associados à predição da classe de interesse, com o intuito de gerar mapas de calor que evidenciem as áreas da imagem que mais influenciaram a decisão do modelo.

Diferentemente de abordagens que utilizam uma camada convolucional fixa, geralmente a última antes da camada densa, este trabalho adota uma estratégia mais abrangente. Cada camada convolucional da CNN é considerada individualmente, uma vez que diferentes camadas capturam distintos níveis de abstração: camadas iniciais tendem a preservar informações espaciais mais detalhadas, enquanto camadas mais profundas representam características semânticas de maior complexidade (SELVARAJU et al., 2017; CHATTOPADHAY et al., 2018; REYES

et al., 2020).

Assim, um laço de repetição percorre todas as camadas convolucionais da rede, aplicando os seguintes procedimentos em cada uma delas:

- Extração da saída da camada convolucional;
- Cálculo dos gradientes da predição em relação a essa camada;
- Ponderação das ativações pela média dos gradientes, resultando em um mapa de calor;
- Aplicação da função ReLU e sobreposição do mapa de calor à imagem original, evidenciando as regiões de maior contribuição para a predição.

Ao final desse processo, é gerado um conjunto de mapas Grad-CAM, cada um correspondente a uma camada convolucional. Para selecionar o mais representativo, este trabalho propõe uma etapa adicional de avaliação: cada mapa gerado é comparado, por meio de uma métrica de sobreposição, às regiões segmentadas por K-means. A camada que apresenta a maior correspondência com as estruturas relevantes é então selecionada como a mais adequada para fins explicativos.

4.1.3 Segmentação com K-Means

A segmentação é aplicada com o objetivo de distinguir as fissuras do fundo da imagem, utilizando o algoritmo K-Means. Essa técnica de agrupamento divide os pixels em dois grupos distintos:

- **Região da fissura:** Pixels pertencentes à fissura na estrutura de concreto;
- **Região de fundo:** Pixels que compõem o fundo da imagem, sem presença de fissuras.

Inicialmente, a imagem é convertida para escala de cinza e, em seguida, aplica-se o algoritmo K-Means com dois agrupamentos. O resultado é uma máscara binária, em que um dos grupos representa a fissura e o outro corresponde ao fundo da imagem.

O objetivo dessa etapa é fornecer uma segmentação automática das fissuras, a qual será posteriormente comparada com os mapas gerados pelo Grad-CAM, a fim de avaliar a consistência das explicações fornecidas pelo modelo.

4.1.4 Comparação das Regiões Segmentadas

Nesta etapa, a máscara gerada pelo Grad-CAM é comparada com aquela obtida por meio da segmentação com K-Means. A avaliação é realizada pixel a pixel, classificando cada ponto em uma das seguintes categorias:

- **Verdadeiro Positivo (TP):** Pixels corretamente identificados como fissura por ambas as técnicas;
- **Verdadeiro Negativo (TN):** Pixels corretamente classificados como fundo por ambas as abordagens;
- **Falso Positivo (FP):** Pixels identificados como fissura pelo Grad-CAM, mas classificados como fundo pela segmentação com K-Means;
- **Falso Negativo (FN):** Pixels classificados como fundo pelo Grad-CAM, mas que pertencem a uma fissura segundo a segmentação com K-Means.

Essa análise permite quantificar a consistência entre as duas técnicas e avaliar se o modelo está, de fato, concentrando-se nas regiões relevantes ao identificar fissuras.

4.1.5 Cálculo das Métricas

Para avaliar o grau de correspondência entre a máscara gerada pelo Grad-CAM e a segmentação obtida por K-Means na identificação de fissuras, utiliza-se a métrica F1-score. Essa métrica fornece uma avaliação equilibrada da capacidade do modelo em focar nas regiões mais relevantes das fissuras, evitando tendências que poderiam surgir ao considerar apenas a precisão ou a revocação de forma isolada (SASAKI et al., 2007).

O F1-score é particularmente útil para tarefas de classificação pixel a pixel em imagens, nas quais o desbalanceamento entre classes é frequente. Ao considerar simultaneamente a precisão e a revocação, essa métrica oferece uma medida mais robusta do desempenho do modelo (POWERS, 2020).

O F1-score é calculado a partir da combinação da precisão (Equação 4.1) e da revocação (Equação 4.2) em uma única métrica balanceada, conforme apresentado na Equação 4.3:

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{4.1}$$

$$\text{Revocação} = \frac{VP}{VP + FN} \tag{4.2}$$

$$F1 = \frac{2 \times \text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{4.3}$$

4.1.6 Seleção da Melhor Camada do Grad-CAM

Redes neurais convolucionais (CNNs) são constituídas por múltiplas camadas convolucionais, cada uma das quais pode gerar mapas de ativação distintos quando submetidas à técnica Grad-CAM (Gradient-weighted Class Activation Mapping). Dado que esses mapas podem diferir significativamente quanto à resolução espacial e ao nível de abstração das características detectadas, torna-se necessário determinar qual camada oferece a explicação mais representativa e precisa da decisão do modelo.

Neste trabalho, o procedimento para identificação da camada mais adequada envolve a aplicação sistemática do Grad-CAM em cada uma das camadas convolucionais disponíveis, seguido pelo cálculo da métrica F1-score para cada mapa gerado. A camada convolucional cujo mapa de ativação apresentar o maior valor de F1-score, refletindo a maior correspondência com as regiões segmentadas das fissuras, é selecionada como a mais apropriada para a representação das regiões relevantes e para a interpretação da decisão do modelo. Essa abordagem objetiva maximizar a qualidade das explicações visuais, garantindo uma interpretação robusta e detalhada do comportamento da CNN.

4.1.7 Salvamento e Visualização dos Resultados

Na etapa final, os resultados são salvos para análise posterior, com o código armazenando diversas informações-chave. Isso inclui as imagens originais, os mapas de calor gerados pelo Grad-CAM, as máscaras segmentadas pelo K-Means e a sobreposição entre o Grad-CAM e a segmentação. Além disso, a camada com melhor desempenho, baseada no maior valor de F1-Score, é registrada para avaliações futuras.

Os resultados também são apresentados graficamente, a fim de facilitar a interpretação visual do desempenho do modelo. Essa visualização permite verificar se os métodos Grad-CAM e K-Means estão alinhados na identificação das mesmas regiões, contribuindo para a validação

da explicabilidade do modelo e assegurando que ele esteja de fato concentrado nas áreas mais relevantes para a detecção de fissuras.

4.2 ABORDAGEM BASEADAS EM PERTUBAÇÃO DE PIXEL

Diferentemente das abordagens baseadas nos pesos internos do modelo, como o Grad-CAM, as técnicas de explicabilidade por perturbação da entrada, como SHAP e LIME, operam de maneira agnóstica à arquitetura da rede neural, não exigindo acesso direto às ativações ou aos gradientes das camadas convolucionais. Em vez disso, essas técnicas inferem a importância de diferentes regiões da imagem ao observar as variações na saída do modelo provocadas por modificações controladas na entrada.

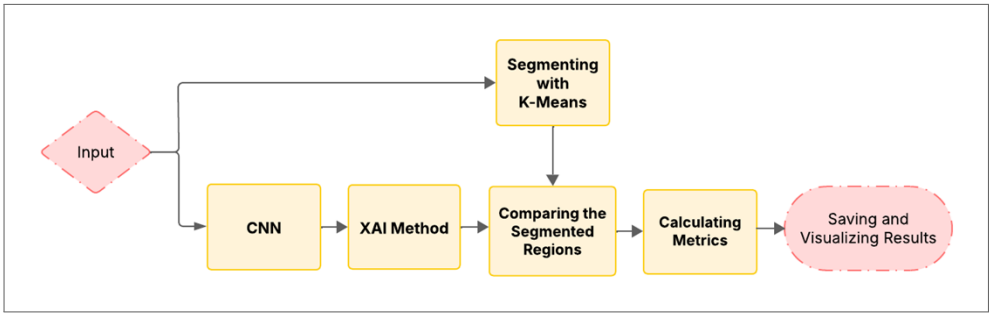
Nesse contexto, não se faz necessária a seleção de uma camada convolucional específica para a geração das explicações, uma vez que os resultados obtidos são independentes da hierarquia interna das representações aprendidas pelo modelo. O processo explicativo é construído com base em amostras geradas por mascaramento parcial da imagem (LIME) ou por coalizões de presença/ausência de regiões visuais (SHAP), sem recorrer às ativações intermediárias da rede.

Assim, neste trabalho, as explicações geradas pelas técnicas de perturbação foram aplicadas diretamente sobre as imagens de entrada, dispensando a etapa de seleção da melhor camada convolucional com base em métricas como o F1-score. Esse procedimento está alinhado com a natureza model-agnostic e seu processo metodológico pode ser descrito em 6 estapas e está ilustrado na Figura 12.

- Carregamento do Modelo e Preparação das Imagens
- Aplicação do Método de Explicabilidade
- Segmentação com K-Means
- Comparacão das Regiões Segmentadas
- Cálculo de Métricas
- Salvamento e Visualização dos Resultados

Tendo em vista que as etapas de Carregamento do Modelo e Preparação das Imagens, Segmentação com K-Means, Comparação das Regiões Segmentadas, Cálculo de Métricas e

Figura 12 – Visão Geral do Método Proposto para Algotimos que Trabalham com Pertubação de Pixels



Fonte: Autor

Salvamento e Visualização dos Resultados já foram detalhadas na seção anterior, esta seção se concentrará apenas na diferença metodológica principal: a técnica de explicabilidade aplicada. As demais etapas do pipeline permanecem inalteradas.

4.2.1 Aplicação do Método de Explicabilidade - SHAP

A técnica SHAP (Shapley Additive Explanations) foi empregada com o objetivo de interpretar localmente as decisões do modelo de classificação, fornecendo mapas visuais que evidenciam a contribuição de cada região da imagem para a predição realizada. Essa abordagem model-agnostic permite estimar a importância marginal de cada característica de entrada de forma consistente e equitativa (LUNDBERG; LEE, 2017).

No presente trabalho, a explicação é obtida utilizando o módulo shap.Explainer, que combina o modelo treinado com um masker do tipo Image, baseado na técnica de inpainting para preenchimento das regiões ocultadas. A função do masker é gerar múltiplas amostras da imagem com perturbações locais, permitindo ao SHAP estimar a variação na predição do modelo em função da presença ou ausência de determinadas regiões visuais.

As contribuições calculadas são representadas por mapas bidimensionais, onde pixels com valores positivos indicam influência positiva na predição da classe-alvo, e valores negativos indicam influência oposta. Esses mapas são posteriormente utilizados para gerar uma máscara binária com base em limiares definidos, a fim de permitir a comparação direta com as regiões segmentadas por K-Means.

4.2.2 Aplicação do Método de Explicabilidade - LIME

A aplicação do LIME (Local Interpretable Model-agnostic Explanations) neste trabalho foi realizada com foco na explicação local das decisões do modelo de classificação.

Para imagens, o LIME opera segmentando a entrada em regiões visuais coesas chamadas de superpixels. Na implementação adotada, a segmentação é realizada com o algoritmo SLIC (Simple Linear Iterative Clustering), de forma a dividir a imagem em um número fixo de regiões, em seguida, o LIME gera amostras sintéticas binárias a partir da imagem original, nas quais diferentes combinações de superpixels são mantidas ou ocultadas. Para cada amostra, a predição do modelo é registrada.

Com base no conjunto de predições obtidas por perturbação da entrada, é ajustado um modelo linear, cuja saída define o peso de importância atribuído a cada superpixel. A imagem resultante é gerada destacando os superpixels com maiores coeficientes positivos, os quais indicam maior contribuição para a classe predita.

Os mapas gerados são exportados em formato gráfico e posteriormente utilizados para gerar uma máscara binária com base em limiares definidos, a fim de permitir a comparação direta com as regiões segmentadas por K-Means.

5 RESULTADOS

Este capítulo apresenta e discute os resultados obtidos a partir da aplicação dos modelos e métodos propostos para a tarefa de detecção e explicação de fissuras em estruturas de concreto. Inicialmente, são reportadas as métricas de desempenho do modelo de classificação adotado, visando avaliar sua eficácia na identificação automática de fissuras. Em seguida, procede-se à análise do método automático de seleção de camadas, desenvolvido com o intuito de aprimorar a interpretabilidade da rede neural convolucional, por meio da identificação das regiões de interesse mais relevantes para a tomada de decisão do modelo.

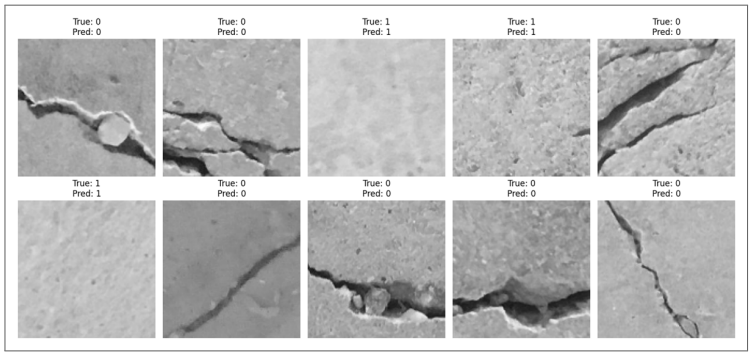
5.1 EVOLUÇÃO DOS MODELOS DE CLASSIFICAÇÃO

5.1.1 Visual Geometry Group - VGG 19

Durante o treinamento, observou-se uma rápida convergência, com a perda (loss) no conjunto de treino reduzindo-se de 0,05 para 0,01 ao longo das épocas. Paralelamente, a perda no conjunto de validação seguiu uma tendência semelhante, indicando ausência significativa de sobreajuste (overfitting).

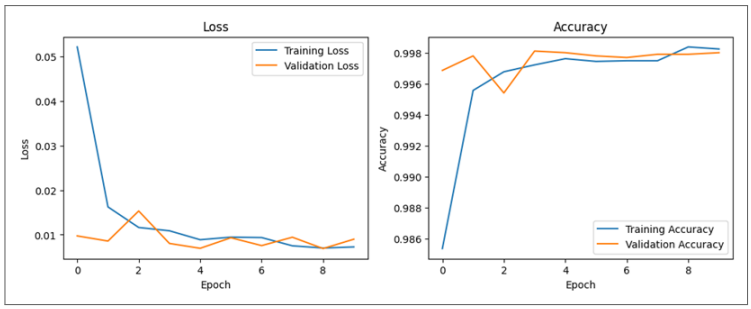
Os resultados demonstram claramente a robustez do modelo, com uma acurácia de 99,87% no conjunto de treino e acima de 99,2% no conjunto de validação. Ao ser avaliado no conjunto de teste, o modelo apresentou uma consistência notável, alcançando uma acurácia de 99,61% com uma perda de 0,0221. O relatório de classificação destacou valores próximos a perfeição para precisão (precision) e revocação (recall) em ambas as classes analisadas, confirmando a alta capacidade do modelo em distinguir com confiabilidade as fissuras. Adicionalmente, a matriz de confusão revelou apenas 31 classificações incorretas em um total de 8002 amostras, corroborando a eficácia da abordagem proposta na detecção automatizada de fissuras em estruturas de concreto. As representações gráficas da evolução das métricas de perda e acurácia ao longo do treinamento e validação encontram-se ilustradas na Figura 14 e a matriz de confusão na Tabela 1.

Figura 13 – Resultados da classificação de imagens de concreto pelo modelo VGG19. "True"indica a classe real da amostra (0 para fissura, 1 para intacto) e "Pred"indica a classe predita pelo modelo. Todos os exemplos exibidos representam classificações corretas.



Fonte: Autor

Figura 14 – Evolução das Métricas de Treinamento para o Modelo VGG 19



Fonte: Autor

Tabela 1 – Matriz de Confusão do Modelo VGG19

Verdadeiro / Predito	Sem Fissura	Com Fissura
Sem Fissura	3979	9
Com Fissura	23	3992

Fonte: Autor.

5.1.2 Visual Geometry Group - VGG 16

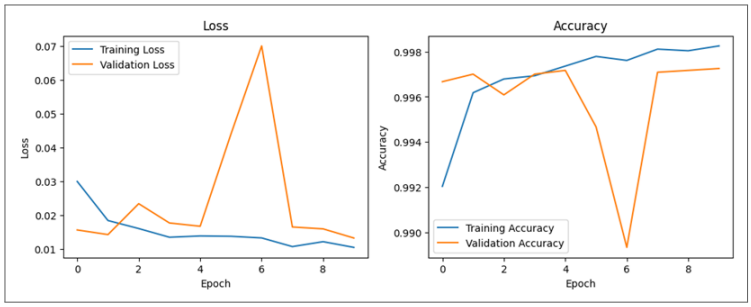
A matriz de confusão do modelo VGG 16, representada na Tabela 2, revelou um número reduzido de erros, com 3977 verdadeiros negativos e apenas 10 falsos positivos, além de 3996 verdadeiros positivos e 19 falsos negativos. Tal distribuição indica alta sensibilidade e especificidade do modelo, refletindo sua capacidade de distinguir com precisão as duas classes, minimizando falsos positivos e falsos negativos.

A Figura 15 representa o gráficos de perda (loss) e acurácia durante o treinamento e validação, e reforça a robustez do VGG16, mostrando convergência da função de perda para valores muito baixos, com a perda de treinamento estabilizando em torno de 0.0071 e a perda

de validação mantendo-se baixa e consistente. A acurácia, tanto no treinamento quanto na validação, permaneceu acima de 99%, evidenciando que o modelo aprendeu representações discriminativas sem apresentar sinais significativos de overfitting. Esse equilíbrio entre baixo erro e alta precisão é indicativo da adequada parametrização da arquitetura, do processo de otimização eficiente e da qualidade do conjunto de dados.

Além disso, os resultados no conjunto de teste confirmam a capacidade de generalização do VGG16, com uma acurácia de 99,64% e uma loss de 0,0167, demonstrando que o modelo mantém alta performance em dados não vistos durante o treinamento. Esses valores refletem a eficiência da arquitetura VGG16 em extrair características relevantes por meio de suas camadas convolucionais profundas.

Figura 15 – Evolução das Métricas de Treinamento para o Modelo VGG 16



Fonte: Autor

Tabela 2 – Matriz de Confusão do Modelo VGG16

Verdadeiro / Predito	**Sem Fissura**	**Com Fissura**
Sem Fissura	3977	10
Com Fissura	19	3996

Fonte: Autor.

5.1.3 Residual Neural Network - ResNet 50

Os resultados obtidos pelo modelo ResNet-50 evidenciam um desempenho robusto tanto durante o treinamento quanto na fase de teste. Observa-se uma acurácia de 98,16% durante o treinamento, acompanhada de um valor de loss relativamente baixo (0,0878), indicando boa capacidade do modelo em ajustar-se aos dados de treinamento. Na etapa de avaliação, a acurácia alcançada foi de 93,66%, com um loss de 0,2278, demonstrando uma leve degradação

do desempenho quando aplicado a dados não vistos anteriormente, o que é esperado em cenários reais de validação.

A matriz de confusão revela aspectos importantes para a análise qualitativa do modelo. O número de verdadeiros positivos (3512) e verdadeiros negativos (3983) é expressivo, confirmando a eficácia do modelo na identificação correta tanto de imagens com fissuras quanto sem fissuras. Contudo, destaca-se a presença de 503 falsos negativos, ou seja, casos em que o modelo classificou incorretamente imagens contendo fissuras como "sem fissura". Esse tipo de erro é particularmente crítico, pois pode acarretar a falha na detecção de danos estruturais, comprometendo a segurança e a integridade das construções analisadas.

Em síntese, o modelo ResNet-50 apresenta desempenho satisfatório para a tarefa de detecção de fissuras em concreto, com alta sensibilidade e especificidade. No entanto, a maior incidência de falsos negativos em comparação aos falsos positivos indica que futuras melhorias deveriam focar na redução desses erros.

Tabela 3 – Matriz de Confusão do Modelo VGG16

Verdadeiro / Predito	**Sem Fissura**	**Com Fissura**
Sem Fissura	3983	4
Com Fissura	503	3512

Fonte: Autor.

5.2 EVOLUÇÃO DOS MÉTODOS DE EXPLICALIDADE

Nesta seção, apresentam-se os resultados das abordagens de explicabilidade adotadas neste trabalho para modelos de classificação baseados em redes neurais convolucionais (CNN), aplicados à detecção de fissuras em estruturas de concreto. São analisadas, inicialmente, as técnicas Layer-wise Relevance Propagation (LRP) e Guided Backpropagation (GBP), ambas voltadas à geração de mapas de saliência que indicam as regiões da imagem mais relevantes para a decisão do modelo.

Durante os experimentos, constatou-se que os mapas gerados apresentaram limitações significativas do ponto de vista interpretativo. As visualizações resultantes exibiram baixo contraste entre as regiões de interesse (fissuras) e o fundo, presença acentuada de ruído textural, bem como ativação difusa e pouco estruturada. Esses fatores comprometeram a identificação precisa das áreas discriminativas da imagem.

Uma das principais dificuldades enfrentadas foi a definição de um valor de threshold adequado para segmentar as regiões relevantes dos mapas de saliência. A ausência de uma separabilidade clara entre os valores atribuídos aos pixels dificultou o estabelecimento de um limiar que evidenciasse, de forma consistente, as regiões associadas às fissuras. Em diversas amostras, pequenas variações no threshold resultaram em perdas abruptas de informação relevante ou, inversamente, na inclusão de ruído espúrio, revelando a instabilidade dos mapas gerados.

Esse comportamento pode estar relacionado ao elevado grau de confiança dos modelos de classificação, os quais apresentaram desempenho acima de 99% de acurácia. Como discutido por Fresz et al. em cenários nos quais o modelo apresenta separabilidade clara entre as classes, métodos de explicação baseados em saliência podem atribuir relevância de forma difusa, dificultando a obtenção de explicações localizadas (FRESZ; LÖRCHER; HUBER, 2024). Diante disso, optou-se por uma análise qualitativa para os dois métodos.

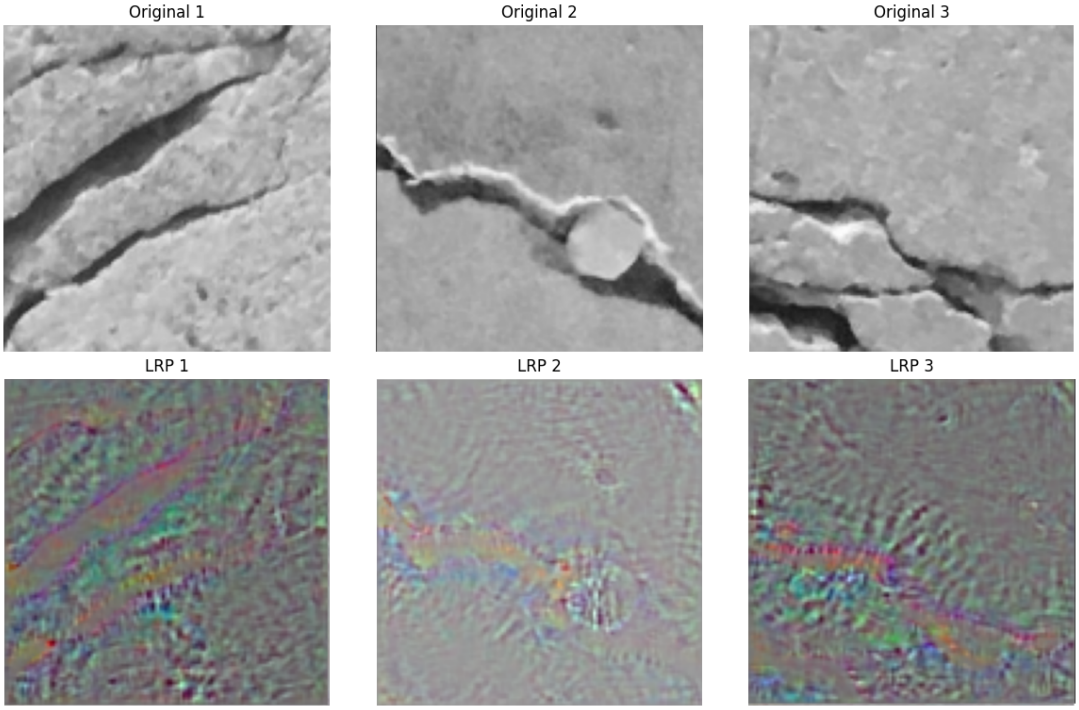


Figura 16 – Imagens originais de concreto fissurado (linha superior) e seus respectivos mapas de saliência gerados pelo método LRP (linha inferior).

A Figura 16 apresenta três exemplos representativos de imagens originais de superfícies de concreto com fissuras (linha superior) e seus respectivos mapas de saliência gerados por meio do método Layer-wise Relevance Propagation (LRP) (linha inferior). O objetivo desta visualização é avaliar a capacidade do método em evidenciar, de forma interpretável, as regiões discriminativas que contribuíram para a predição da classe “fissura”.

Observa-se que, embora os mapas gerados possuam padrões visuais de ativação distribuídos ao longo da imagem, as regiões relevantes atribuídas pelo LRP apresentam características de difusão espacial e baixo contraste estrutural. Em todos os três casos, há ativação sobre porções da imagem que não coincidem com a fissura em si, sugerindo uma distribuição pouco seletiva da importância atribuída aos pixels. Em particular, nota-se a presença de artefatos em padrão ondulatório que, apesar de indicarem atividade de rede, dificultam a associação direta com a morfologia da fissura presente na imagem original, ainda que parte da ativação se sobreponha à região fissurada, o que compromete a interpretabilidade.

Essas observações corroboram os apontamentos da literatura sobre as limitações de interpretabilidade em métodos baseados em decomposição como o LRP, especialmente quando aplicados em modelos com alta confiança preditiva e entradas com baixo grau de variação semântica local (FRESZ; LÖRCHER; HUBER, 2024). Em tais casos, a atribuição de relevância tende a ser espalhada ao longo da imagem, uma vez que o modelo pode aprender representações altamente distribuídas e redundantes.

Ressalta-se ainda a dificuldade operacional enfrentada na definição de um *threshold* adequado para a visualização dos mapas. Pequenas variações no limiar resultaram em representações excessivamente saturadas ou, ao contrário, visualmente esmaecidas, o que evidencia a instabilidade da representação e a sensibilidade do método a parâmetros de pós-processamento.

Portanto, com base na avaliação visual conduzida, conclui-se que os mapas gerados pelo método LRP, nas condições deste experimento, não forneceram explicações visuais suficientemente claras ou interpretáveis para embasar decisões técnicas sobre a detecção de fissuras.

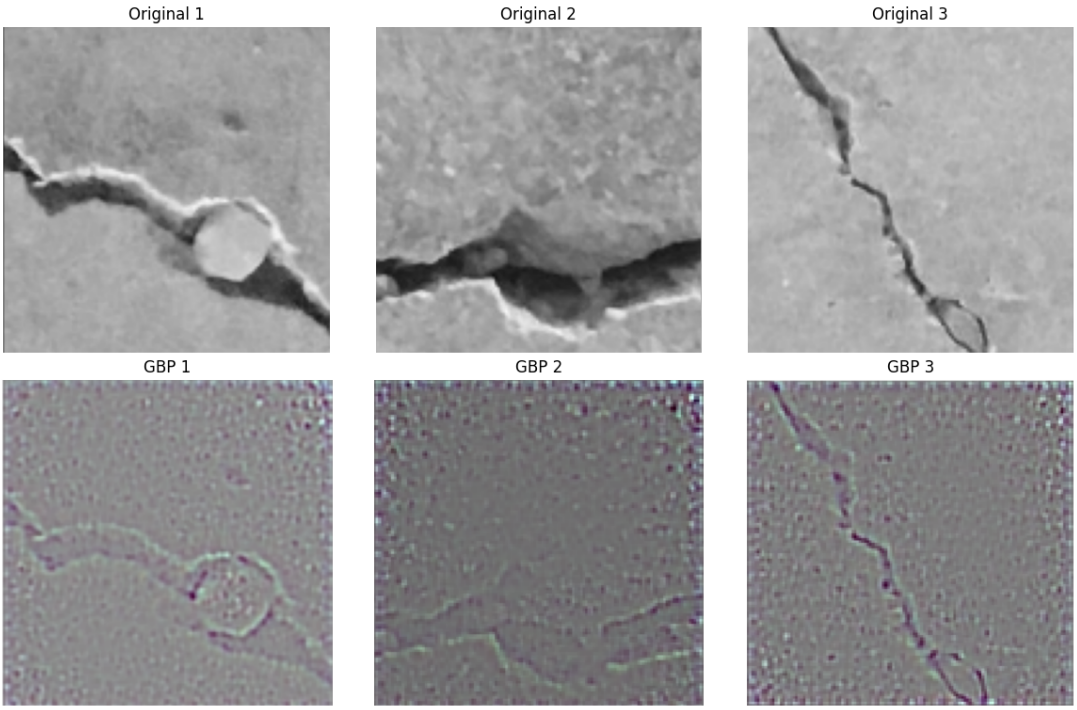


Figura 17 – Imagens originais de superfícies de concreto fissuradas (linha superior) e respectivos mapas de saliência gerados pelo método Guided Backpropagation (GBP) (linha inferior).

A Figura 17 apresenta três exemplos representativos de imagens de concreto com fissuras (linha superior) e os respectivos mapas de saliência gerados com a técnica Guided Backpropagation (GBP) (linha inferior). Esta abordagem é baseada no fluxo reverso de gradientes e tem como objetivo destacar, de forma mais precisa, as regiões da imagem que contribuíram positivamente para a decisão do modelo de classificação.

Em contraste com os resultados observados nos mapas produzidos pelo método LRP (Figura 16), o GBP apresenta visualizações consideravelmente mais limpas e mais estruturadas. As fissuras são realçadas com maior nitidez e linearidade. Nota-se que as regiões de interesse apresentam uma correspondência visual mais próxima com os traçados reais das fissuras nas imagens originais, o que favorece a interpretabilidade por parte do analista humano.

Adicionalmente, observa-se que o método GBP se mostrou particularmente eficaz na detecção de fissuras com menores espessuras, como demonstrado no exemplo 3. Nesse caso, mesmo diante de uma fissura estreita e de baixo contraste no plano original, a visualização gerada evidenciou com clareza seu traçado contínuo, preservando as características morfológicas da patologia. Esse comportamento indica maior sensibilidade do método para detalhes finos, o que é uma vantagem relevante no contexto da análise estrutural, onde microfissuras podem representar sintomas críticos de degradação.

No entanto, embora os resultados gerais do GBP sejam mais satisfatórios, ainda se verificam limitações na presença de ruído espacial, como exemplificado na "GBP 2". Nessa amostra, a fissura está parcialmente oculta por texturas irregulares e sombras no substrato, o que parece ter dificultado a ativação do mapa em regiões críticas, resultando em saliências fracas ou ausentes.

De forma geral, os resultados obtidos com o método GBP demonstram melhor desempenho qualitativo quando comparados ao LRP, tanto em termos de contraste visual quanto de alinhamento morfológico com a fissura original. A capacidade do GBP em destacar estruturas lineares e delgadas, associada à sua maior robustez na definição do limiar de visualização, o torna mais promissor para aplicações práticas em inspeção visual assistida por inteligência artificial, particularmente em tarefas de identificação precoce de manifestações patológicas em concreto.

Diferentemente de métodos baseados em gradientes diretos, como GBP, o Grad-CAM utiliza gradientes ponderados das ativações de uma camada convolucional específica, possibilitando a geração de mapas alinhados semanticamente com as estruturas discriminativas presentes na imagem, como as fissuras em concreto.

No entanto, a qualidade e a semântica dos mapas gerados pelo Grad-CAM são fortemente influenciadas pela escolha da camada convolucional analisada. Camadas mais superficiais capturam informações de baixo nível (como bordas e texturas), enquanto camadas mais profundas enfatizam abstrações semânticas. Essa variabilidade entre os níveis da rede torna necessária uma análise multiescala dos mapas gerados, a fim de identificar as camadas que produzem explicações mais informativas e confiáveis para a tarefa de classificação de fissuras estruturais.

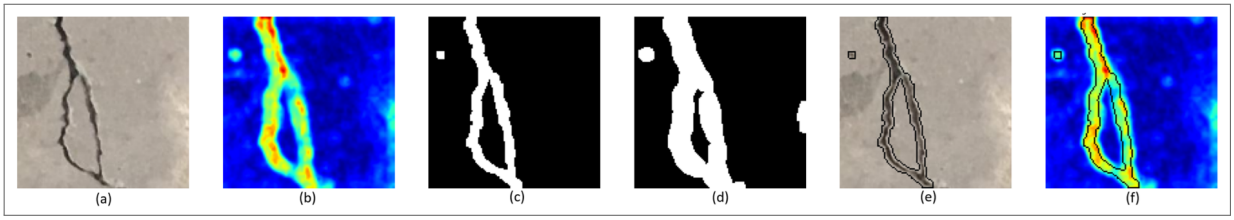
Na análise qualitativa dos mapas gerados pelo Grad-CAM, observou-se que não houve uma única camada convolucional capaz de produzir consistentemente as melhores explicações visuais ao longo das diversas amostras analisadas. A variabilidade nas representações obtidas em diferentes profundidades da rede indicou que a qualidade das explicações é altamente dependente do nível de abstração da camada considerada, sendo que, em alguns casos, camadas mais profundas enfatizavam regiões irrelevantes, enquanto camadas intermediárias ou superficiais forneciam mapas mais alinhados com as fissuras reais presentes nas imagens.

Diante dessa inconsistência, tornou-se necessário complementar a análise com uma etapa de validação especializada, na qual engenheiros civis atuaram como avaliadores para indicar, com base em seu conhecimento técnico, quais mapas apresentavam maior aderência aos padrões visuais típicos de fissuras estruturais. Para além dessa validação subjetiva, a seleção

da camada mais adequada também foi conduzida de forma quantitativa, seguindo a metodologia descrita no Capítulo 3, na qual métricas objetivas foram utilizadas para comparar a sobreposição dos mapas de saliência com regiões de interesse previamente segmentadas. Esse procedimento possibilitou uma avaliação sistemática da fidelidade explicativa das diferentes camadas e contribuiu para a definição do ponto ótimo de extração das explicações visuais via Grad-CAM.

A Figura 18 sintetiza esse processo, começando pela imagem original da fissura (a), seguida pelo mapa de ativação gerado pelo Grad-CAM (b). Inclui ainda a máscara de segmentação obtida pelo K-means (c) e a máscara gerada pelo Grad-CAM (d). A figura destaca também a imagem original com os contornos da segmentação (e) e a imagem sobreposta pelo Grad-CAM junto com os contornos da segmentação (f). Essa disposição facilita uma comparação visual da capacidade do modelo em identificar e segmentar fissuras.

Figura 18 – Ilustração do processo de explicação e segmentação de fissuras em concreto. (a) Imagem original; (b) Mapa de ativação gerado pelo Grad-CAM; (c) Máscara de segmentação obtida pelo K-means; (d) Máscara gerada pelo Grad-CAM; (e) Imagem original com contornos da segmentação; (f) Imagem com sobreposição do Grad-CAM e contornos segmentados.



Fonte: Autor

A camada convolucional selecionada automaticamente apresentou grande similaridade com aquela escolhida manualmente pelos especialistas, confirmando que o método automático é eficaz na identificação do nível mais adequado de abstração para explicar as decisões do modelo. Isso reforça a confiabilidade do método proposto e sua aplicação na classificação de fissuras em concreto.

Em relação ao desempenho, os resultados da métrica F1 indicam que o modelo apresentou desempenho satisfatório de forma geral, com média de 0,6488 e elevada variabilidade (desvio padrão = 0,2363). Contudo, ao considerar exclusivamente a classe contendo fissuras, o desempenho foi significativamente superior, com média de 0,7511, menor dispersão (desvio padrão = 0,2127) e mediana de 0,8198, indicando que a maioria das segmentações foi bem-sucedida. Por outro lado, foram identificados alguns valores atípicos inferiores, com escores F1 abaixo de 0,3, sugerindo menor precisão da segmentação em certos casos. Esses resultados indicam

que, embora o modelo tenha apresentado bom desempenho geral, ele pode apresentar dificuldades em imagens mais complexas, evidenciando a necessidade de ajustes para aprimorar a segmentação em situações mais desafiadoras.

Além dos métodos baseados em saliência aplicados diretamente à estrutura interna das redes neurais, como LRP, GBP e Grad-CAM, também foram analisadas abordagens baseadas em explicações agnósticas ao modelo, especificamente SHAP (SHapley Additive exPlanations) e LIME (Local Interpretable Model-agnostic Explanations). Diferentemente das abordagens anteriores, essas técnicas não dependem da arquitetura interna da rede, mas operam por meio de perturbações na entrada e análise do impacto dessas perturbações sobre a saída do modelo, visando construir uma explicação local aproximada da decisão.

Apesar da flexibilidade metodológica, a aplicação visual e qualitativa do SHAP e do LIME às imagens de concreto fissurado apresentou limitações consideráveis. Em ambos os casos, observou-se uma alta quantidade de regiões da imagem assinaladas como relevantes para a classificação, muitas das quais não apresentavam qualquer correspondência visual com a fissura propriamente dita. Esse comportamento compromete a interpretabilidade das explicações, uma vez que induz o observador a falsas associações entre padrões irrelevantes da imagem e a presença da patologia. Assim, os mapas gerados por SHAP e LIME revelaram-se menos informativos e mais ruidosos quando comparados aos métodos que exploram diretamente os gradientes ou ativações internas da rede neural, tornando difícil a identificação precisa das regiões críticas que de fato motivaram a predição da classe “fissura”.

Em comparação ao resultados quantitativos do Grad-CAM, o método baseado em SHAP, apresentou desempenho significativamente inferior na identificação das regiões relevantes das fissuras. A média do F1 foi de 0,2477, com elevado desvio padrão de 0,2357. Além disso, a mediana do F1 foi apenas 0,1909, e 25% das amostras obtiveram escores inferiores a 0,0396, indicando baixo alinhamento explicativo em grande parte dos casos. Embora valores máximos tenham alcançado 0,9417 em exemplos específicos, a ampla variabilidade e a baixa tendência central sugerem que o SHAP apresenta dificuldades para fornecer explicações consistentes e espacialmente precisas no contexto da segmentação de fissuras.

O método LIME demonstrou um desempenho consistente e estável na explicação das decisões do modelo, com uma média de F1 de 0,1947 e mediana de 0,2289. Embora esses resultados não sejam excepcionais, o LIME se destacou pela sua capacidade de fornecer explicações mais equilibradas (desvio padrão de 0,0948) em comparação ao SHAP, que apresentou uma variabilidade considerável.

A Tabela 4 exibe as estatísticas descritivas dos resultados quantitativos obtidos pelos métodos Grad-CAM, LIME e Grad-CAM aplicados à segmentação de fissuras em concreto. São apresentados os valores da média, desvio padrão e mediana para cada método, possibilitando uma comparação direta entre as abordagens avaliadas.

Tabela 4 – Estatísticas dos resultados dos métodos Grad-CAM, LIME e SHAP

Método	**Média**	**Desvio Padrão**	**Mediana**
Grad-CAM	0,6488	0,2362	0,8709
LIME	0,1947	0,0948	0,2289
SHAP	0,2477	0,2357	0,1909

Fonte: Autor

Esses achados reforçam que métodos baseados em perturbação de pixels, que tratam o modelo como uma caixa-preta, tendem a gerar explicações menos confiáveis e menos coerentes espacialmente, quando comparados a abordagens baseadas em gradientes. Em contrapartida, o método proposto, fundamentado no Grad-CAM e na seleção automática de camadas, demonstrou desempenho substancialmente superior, especialmente na classe fissura (média do F1 $=$ 0,7511), indicando maior capacidade de localizar e evidenciar regiões discriminativas. Tal resultado enfatiza a importância de utilizar informações internas do modelo para produzir explicações visuais mais fidedignas e interpretáveis em tarefas de classificação de imagens baseadas em CNN.

6 CONCLUSÃO

Nesta dissertação, foi proposto um método inovador de seleção automática de camadas em combinação com um modelo de classificação baseado em aprendizado por transferência, para detecção e classificação de fissuras em concreto. Os resultados obtidos evidenciaram alto desempenho dos modelos, alcançando precisão superior a 99% tanto no conjunto de treinamento quanto no de testes, confirmando assim a eficácia da abordagem adotada.

A integração das técnicas Grad-CAM e segmentação K-means revelou-se especialmente útil para fornecer insights sobre o processo decisório do modelo, contribuindo significativamente para sua interpretabilidade. Essa capacidade de explicação das predições constitui um avanço importante para aplicações práticas em monitoramento estrutural, nas quais a confiabilidade e a transparência das decisões automatizadas são fundamentais.

Destaca-se também o desempenho qualitativo do método Guided Backpropagation (GBP), que apresentou mapas de saliência mais limpos e coerentes do ponto de vista morfológico em comparação a outros métodos analisados. O GBP mostrou-se particularmente eficaz na detecção de fissuras finas e bem definidas, preservando traçados lineares relevantes para diagnósticos visuais. Além disso, sua maior estabilidade na definição do limiar de visualização o torna uma alternativa promissora em sistemas computacionais voltados à inspeção visual automatizada.

O potencial impacto dessa pesquisa na prática da engenharia civil é significativo, sobretudo em cenários nos quais a inspeção periódica de estruturas é logisticamente inviável. Pontes, viadutos e outras obras de infraestrutura frequentemente carecem de manutenção preventiva sistemática, seja por limitações operacionais, orçamentárias ou pela escassez de profissionais especializados. A incorporação de modelos de inteligência artificial interpretáveis, como o proposto neste trabalho, possibilita a triagem automatizada de manifestações patológicas, promovendo a detecção precoce de danos estruturais. Tais sistemas, quando integrados a dispositivos embarcados, como drones ou sensores visuais remotos, podem representar uma solução eficaz para aumentar a segurança, reduzir custos e ampliar a cobertura das inspeções técnicas em regiões remotas ou de difícil acesso.

Contudo, apesar dos bons resultados gerais, o modelo apresentou limitações em situações específicas relacionadas à presença de outliers e superfícies de concreto mais complexas e heterogêneas. Tal constatação indica a necessidade de aprimoramentos futuros que contemplem técnicas avançadas de segmentação e experimentações com diferentes arquiteturas de redes

neurais convolucionais (CNNs).

Até onde se tem conhecimento, não foram encontrados estudos anteriores que integrem as metodologias Grad-CAM e segmentação K-means para análise de imagens de fissuras em concreto. Acredita-se, portanto, que esta pesquisa constitua uma contribuição pioneira na literatura. Além disso, o compartilhamento aberto do código e dos detalhes de implementação, disponíveis em repositório público, facilita a replicabilidade e o desenvolvimento de pesquisas subsequentes (FRAGOSO, 2025).

Para trabalhos futuros, sugere-se aprimorar a robustez do modelo frente às situações adversas mencionadas, incorporando técnicas mais sofisticadas de segmentação e explorando arquiteturas CNN alternativas. Ademais, ampliar a base de dados utilizada, abrangendo maior diversidade de tipos de fissuras e condições ambientais, será crucial para melhorar a generalização do modelo. Tais avanços contribuirão para tornar o sistema uma ferramenta ainda mais efetiva e confiável na automação de inspeções estruturais em aplicações reais.

Adicionalmente, recomenda-se a realização de estratégias de clusterização das imagens antes do treinamento, visando mitigar potenciais vieses nos dados e promover uma distribuição mais equilibrada entre as características presentes. Recomenda-se, ainda, investigar o comportamento da explicabilidade com a execução de treinamentos com ajustes de peso (weight adjustments), mesmo que isso implique eventuais reduções na acurácia do modelo, considerando que interpretações mais confiáveis podem ser prioritárias em aplicações críticas.

Além dessas direções, ressalta-se a importância de investigar formas mais rigorosas de avaliação da explicabilidade. A métrica baseada na sobreposição espacial entre mapas de ativação e regiões segmentadas, embora funcional, não capta aspectos semânticos das decisões do modelo. Nesse sentido, propõe-se a adoção futura de abordagens psicométricas de confiabilidade inter-método e métricas mais refinadas, conforme discutido na literatura recente.

Também se reconhece que a comparação entre métodos explicáveis pode ter sido impactada pela diferença de natureza entre eles (e.g., model-specific vs. model-agnostic), o que justifica a necessidade de protocolos de avaliação mais justos. Por fim, recomenda-se a extensão do conjunto de métodos XAI avaliados, incorporando técnicas como Integrated Gradients, RISE ou B-cos, e o refinamento do fine-tuning das redes utilizadas para garantir uma adaptação mais profunda ao domínio específico da engenharia civil.

Essas limitações não comprometem as contribuições centrais deste trabalho, mas indicam oportunidades claras de continuidade científica e tecnológica no campo da Inteligência Artificial Explicável aplicada à engenharia estrutural.

REFERÊNCIAS

ACHANTA, R.; SHAJI, A.; SMITH, K.; LUCCHI, A.; FUA, P.; SÜSSTRUNK, S. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 34, n. 11, p. 2274–2282, 2012.

ADEBAYO, J.; GILMER, J.; MUELLY, M.; GOODFELLOW, I.; HARDT, M.; KIM, B. Sanity checks for saliency maps. *Advances in neural information processing systems*, v. 31, 2018.

AKHTAR, N.; AHMAD, M. V. A modified fuzzy c means clustering using neutrosophic logic. In: IEEE. *2015 Fifth International Conference on Communication Systems and Network Technologies*. [S.l.], 2015. p. 1124–1128.

ARIAS-DUART, A.; PARÉS, F.; GARCIA-GASULLA, D.; GIMENEZ-ABALOS, V. Focus! rating xai methods and finding biases. In: IEEE. *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. [S.l.], 2022. p. 1–8.

AYODELE, T. O. Types of machine learning algorithms. *New advances in machine learning*, v. 3, n. 19-48, p. 5–1, 2010.

BACH, S.; BINDER, A.; MONTAVON, G.; KLAUSCHEN, F.; MÜLLER, K.-R.; SAMEK, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, Public Library of Science San Francisco, CA USA, v. 10, n. 7, p. e0130140, 2015.

BARCELLOS, W.; GONZAGA, A. Periocular authentication in smartphones applying ulbp descriptor on cnn feature maps. In: *Anais do XVII Workshop de Visão Computacional*. Porto Alegre, RS, Brasil: SBC, 2021. p. 57–63. ISSN 0000-0000.

BRYAN-KINNS, N. Reflections on explainable ai for the arts (xaixarts). *Interactions*, ACM New York, NY, USA, v. 31, n. 1, p. 43–47, 2024.

BURGER, W.; BURGE, M. J. *Digital image processing: An algorithmic introduction*. [S.l.]: Springer Nature, 2022.

CAMACHO, D. M. et al. Next-generation machine learning for biological networks. *Cell*, Elsevier, v. 173, n. 7, p. 1581–1592, 2018.

CHATTOPADHAY, A.; SARKAR, A.; HOWLADER, P.; BALASUBRAMANIAN, V. N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: IEEE. *2018 IEEE winter conference on applications of computer vision (WACV)*. [S.l.], 2018. p. 839–847.

COTRIM, G. G.; SHIMIZU, A. H.; ROSA, L. P. C. Identificação das características patológicas em estruturas de concreto. 2023.

DAI, L.; LIU, X.; LIU, Y.; YANG, C.; WEI, L.; LIN, Y.; CHEN, R. Enhancing two-view correspondence learning by local-global self-attention. *Neurocomputing*, Elsevier, v. 459, p. 176–187, 2021.

DENG, J.; DONG, W.; SOCHER, R.; LI, L.-J.; LI, K.; FEI-FEI, L. Imagenet: A large-scale hierarchical image database. In: IEEE. *2009 IEEE conference on computer vision and pattern recognition*. [S.l.], 2009. p. 248–255.

DHORE, V.; BHAT, A.; NERLEKAR, V.; CHAVHAN, K.; UMARE, A. Enhancing explainable ai: A hybrid approach combining gradcam and lrp for cnn interpretability. *arXiv preprint arXiv:2405.12175*, 2024.

DONG, W.; LIANG, Z.; WANG, L.; TIAN, G.; LONG, Q. Unsupervised domain adaptation segmentation algorithm with cross-domain data augmentation and category contrast. *Neurocomputing*, v. 623, p. 129393, 2025. ISSN 0925-2312.

DUCHI, J.; HAZAN, E.; SINGER, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, v. 12, n. 7, 2011.

ELGENDY, M. *Deep learning for vision systems*. [S.l.]: Manning, 2020.

FEITOSA, J. d. C. Análise da explicabilidade dos métodos de inteligência artificial explicável aplicados a segmentação de imagens de peixes pacu. Universidade Estadual Paulista (Unesp), 2024.

FELLOUS, J.-M. et al. Explainable artificial intelligence for neuroscience: Behavioral neurostimulation. *Frontiers in Neuroscience*, v. 13, 2019.

FERNANDES, A.; JUNIOR, G.; DINIZ, J.; SILVA, A.; MATOS, C. Efficientdeeplab for automated trachea segmentation on medical images. In: *Anais da XII Brazilian Conference on Intelligent Systems*. Porto Alegre, RS, Brasil: SBC, 2023. p. 154–166. ISSN 2643-6264.

FILHO, O. M.; NETO, H. V. *Processamento digital de imagens*. [S.l.]: Brasport, 1999.

FOREST, F.; PORTA, H.; TUIA, D.; FINK, O. From classification to segmentation with explainable ai: A study on crack detection and growth monitoring. *Automation in Construction*, Elsevier, v. 165, p. 105497, 2024.

FRAGOSO, G. A. M. *XAI Crack Detection Repository*. 2025. Https://github.com/bielarnaud/Xai$_{F}or_{C}rack_{D}etection_{P}aper$.

FRESZ, B.; LÖRCHER, L.; HUBER, M. Classification metrics for image explanations: towards building reliable xai-evaluations. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. [S.l.: s.n.], 2024. p. 1–19.

GEORGE, A. S. Recent techniques in machine learning: A review. 2024.

GHASSEMI, M.; OAKDEN-RAYNER, L.; BEAM, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, Elsevier, v. 3, n. 11, p. e745–e750, 2021.

GILPIN, L. H.; BAU, D.; YUAN, B. Z.; BAJWA, A.; SPECTER, M.; KAGAL, L. Explaining explanations: An overview of interpretability of machine learning. In: IEEE. *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. [S.l.], 2018. p. 80–89.

GIPIŠKIS, R.; TSAI, C.-W.; KURASOVA, O. Explainable ai (xai) in image segmentation in medicine, industry, and beyond: A survey. *ICT Express*, Elsevier, 2024.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A.; BENGIO, Y. *Deep learning*. [S.l.]: MIT press Cambridge, 2016. v. 1.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 770–778.

HOLZINGER, A.; CARRINGTON, A.; MÜLLER, H. Measuring the quality of explanations: the system causability scale (scs) comparing human and machine explanations. *KI-Künstliche Intelligenz*, Springer, v. 34, n. 2, p. 193–198, 2020.

HOOKER, S.; ERHAN, D.; KINDERMANS, P.-J.; KIM, B. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, v. 32, 2019.

HORTA, V. A.; TIDDI, I.; LITTLE, S.; MILEO, A. Extracting knowledge from deep neural networks through graph analysis. *Future Generation Computer Systems*, v. 120, p. 109–118, 2021. ISSN 0167-739X.

JUNIOR, W. M. P.; SILVA, S. F. d.; SILVA, A. R.; REZIO, L. H. F.; SILVA, M. P. d.; GUIMARÃES, N. R. d. S.; CANUTO, S. D. C. Cracks detection in images of concrete structures using deep neural networks. *Matéria (Rio de Janeiro)*, SciELO Brasil, v. 29, p. e20240354, 2024.

KAUFMAN, D. *A inteligência artificial irá suplantar a inteligência humana?* [S.l.]: Estação das Letras e Cores EDI, 2019.

KAVITHA, S.; BASKARAN, K.; DHANAPRIYA, B. Explainable ai for detecting fissures on concrete surfaces using transfer learning. In: IEEE. *2023 International Conference on Inventive Computation Technologies (ICICT)*. [S.l.], 2023. p. 376–384.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, v. 25, 2012.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *nature*, Nature Publishing Group UK London, v. 521, n. 7553, p. 436–444, 2015.

LINARDATOS, P.; PAPASTEFANOPOULOS, V.; KOTSIANTIS, S. Explainable ai: A review of machine learning interpretability methods. *Entropy*, MDPI, v. 23, n. 1, p. 18, 2020.

LUDERMIR, T. B. Inteligência artificial e aprendizado de máquina: estado atual e tendências. *Estudos Avançados*, v. 35, n. 101, p. 85–94, 2021.

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

LUNDBERG, S. M.; NAIR, B.; VAVILALA, M. S.; HORIBE, M.; EISSES, M. J.; ADAMS, T.; LISTON, D. E.; LOW, D. K.-W.; NEWMAN, S.-F.; KIM, J. et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, Nature Publishing Group, v. 2, n. 10, p. 749, 2018.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, v. 1, n. 14, p. 281–297, 1967.

MAHESH, B. et al. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, v. 9, n. 1, p. 381–386, 2020.

MARQUES, J.; GONçALVES, C.; FILHO, A. C.; VERAS, R.; SILVA, R. V. e. Automated segmentation of computed tomography images for covid-19 patient evaluation. In: *Anais da XXXIV Brazilian Conference on Intelligent Systems*. Porto Alegre, RS, Brasil: SBC, 2024. p. 125–140. ISSN 2643-6264.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, Springer, v. 5, p. 115–133, 1943.

MINAEE, S.; LUO, P.; LIN, Z.; BOWYER, K. Going deeper into face detection: A survey. *arXiv preprint arXiv:2103.14983*, 2021.

MONTAVON, G.; LAPUSCHKIN, S.; BINDER, A.; SAMEK, W.; MÜLLER, K.-R. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, Elsevier, v. 65, p. 211–222, 2017.

MONTAVON, G.; SAMEK, W.; MÜLLER, K.-R. Methods for interpreting and understanding deep neural networks. *Digital signal processing*, Elsevier, v. 73, p. 1–15, 2018.

ORYNBAY, L.; BEKMANOVA, G.; YERGESH, B.; OMARBEKOVA, A.; SAIRANBEKOVA, A.; SHARIPBAY, A. The role of cognitive computing in nlp. *Frontiers in Computer Science*, Frontiers Media SA, v. 6, p. 1486581, 2025.

OSISANWO, F.; AKINSOLA, J.; AWODELE, O.; HINMIKAIYE, J.; OLAKANMI, O.; AKINJOBI, J. et al. Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, v. 48, n. 3, p. 128–138, 2017.

OTTONI, A. L. C.; AL. et. *Métodos para recomendação de hiperparâmetros de aprendizado de máquina na classificação de imagens da construção civil*. Tese (Doutorado) — Nome da Universidade, Cidade, País, 2022.

ÖZGENEL, Ç. F. Concrete crack images for classification. *Mendeley Data*, v. 2, p. 2019, 2019.

PEDRINI, H.; SCHWARTZ, W. R. *Análise de imagens digitais: princípios, algoritmos e aplicações*. [S.l.]: Cengage Learning, 2008.

PEREIRA, W. F. Identificação automatizada de manifestações patológicas por meio de segmentação semântica: fissuras, rachaduras, trincas e fendas. Universidade Federal do Maranhão, 2023.

PEREIRA, W. M.; SILVA, S. F. d.; SILVA, A. R. e.; REZIO, L. H. F.; SILVA, M. P. d.; GUIMARÃES, N. R. d. S.; CANUTO, S. D. C. Avaliação da presença de fissuras em imagens de estruturas de concreto através do uso de redes neurais profundas. *Matéria (Rio de Janeiro)*, SciELO Brasil, v. 29, n. 4, p. e20240354, 2024.

PETSIUK, V.; DAS, A.; SAENKO, K. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

POWERS, D. M. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*, 2020.

QUEIROZ, J. E. R. de; GOMES, H. M. Introdução ao processamento digital de imagens. *Rita*, v. 13, n. 2, p. 11–42, 2006.

QURESHI, M. N.; AHAMAD, M. V. An improved method for image segmentation using k-means clustering with neutrosophic logic. *Procedia Computer Science*, v. 132, p. 534–540, 2018. ISSN 1877-0509. International Conference on Computational Intelligence and Data Science.

REAUNGAMORNRAT, S.; SARI, H.; CATANA, C.; KAMEN, A. Multimodal image synthesis based on disentanglement representations of anatomical and modality specific features, learned using uncooperative relativistic gan. *Medical Image Analysis*, v. 80, p. 102514, 2022. ISSN 1361-8415. Disponível em: <https://www.sciencedirect.com/science/article/pii/S136184152200161X>.

REYES, M.; MEIER, R.; PEREIRA, S.; SILVA, C. A.; DAHLWEID, F.-M.; TENGG-KOBLIGK, H. v.; SUMMERS, R. M.; WIEST, R. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiology: artificial intelligence*, Radiological Society of North America, v. 2, n. 3, p. e190043, 2020.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should i trust you?": Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 1135–1144. ISBN 9781450342322. Disponível em: <https://doi.org/10.1145/2939672.2939778>.

ROZENDO, G. B.; ROBERTO, G. F.; NASCIMENTO, M. Z. do; NEVES, L. A.; LUMINI, A. Weeds classification with deep learning: An investigation using cnn, vision transformers, pyramid vision transformers, and ensemble strategy. In: SPRINGER. *Iberoamerican Congress on Pattern Recognition*. [S.l.], 2023. p. 229–243.

RUDIN, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, v. 1, p. 206–215, 2019.

RUSS, J. C. *The image processing handbook*. [S.l.]: CRC press, 2006.

RUSSELL, S. J.; NORVIG, P. *Inteligência artificial*. [S.l.]: Elsevier, 2004.

SAMEK, W.; MÜLLER, K.-R. Towards explainable artificial intelligence. In: *Explainable AI: interpreting, explaining and visualizing deep learning*. [S.l.]: Springer, 2019. p. 5–22.

SAMEK, W.; WIEGAND, T.; MÜLLER, K.-R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.

SASAKI, Y. et al. The truth of the f-measure. *Teach tutor mater*, v. 1, n. 5, p. 1–5, 2007.

SELVARAJU, R. R.; COGSWELL, M.; DAS, A.; VEDANTAM, R.; PARIKH, D.; BATRA, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2017. p. 618–626.

SELVARAJU, R. R.; DAS, A.; VEDANTAM, R.; COGSWELL, M.; PARIKH, D.; BATRA, D. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016.

SHAH, S. S.; SHEPPARD, J. W. Evaluating explanations of convolutional neural network image classifications. In: IEEE. *2020 International joint conference on neural networks (IJCNN)*. [S.l.], 2020. p. 1–8.

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.; VEDALDI, A.; ZISSERMAN, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

SPRINGENBERG, J. T.; DOSOVITSKIY, A.; BROX, T.; RIEDMILLER, M. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

SRIVASTAVA, N.; HINTON, G.; KRIZHEVSKY, A.; SUTSKEVER, I.; SALAKHUTDINOV, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, JMLR. org, v. 15, n. 1, p. 1929–1958, 2014.

TANWAR, S. *Artificial Intelligence and Machine Learning-Principles and Applications*. [S.l.]: Academic Guru Publishing House, 2024.

TURING, A. M. *Computing machinery and intelligence*. [S.l.]: Springer, 2009.

YAMASHITA, R.; NISHIO, M.; DO, R. K. G.; TOGASHI, K. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, Springer, v. 9, p. 611–629, 2018.

ZEILER, M. D.; FERGUS, R. Visualizing and understanding convolutional networks. In: SPRINGER. *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. [S.l.], 2014. p. 818–833.

ZHANG, J.; BARGAL, S. A.; LIN, Z.; BRANDT, J.; SHEN, X.; SCLAROFF, S. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, Springer, v. 126, n. 10, p. 1084–1102, 2018.