IA EXPLICÁVEL EM REDES NEURAIS CONVOLUCIONAIS: COMPARANDO LIME, GRAD-CAM E SHAP

Edmar Junyor Bevilaqua
Instituto de Informática
Universidade Federal do Rio Grande do Sul
Porto Alegre, Rio Grande do Sul - Brasil
edmar.bevilaqua@inf.ufrgs.br

Felipe Ferro Callil Nascimento
Instituto de Informática
Universidade Federal do Rio Grande do Sul
Porto Alegre, Rio Grande do Sul - Brasil
felipe.ferro@inf.ufrgs.br

Lucas Leonardo Fazioni
Instituto de Informática
Universidade Federal do Rio Grande do Sul
Porto Alegre, Rio Grande do Sul - Brasil
lucas.fazioni@inf.ufrgs.br

RESUMO

O progresso em modelos de *deep learning*, particularmente Redes Neurais Convolucionais (CNNs), tem sido acompanhado por desafios crescentes quanto à interpretabilidade de suas decisões. Este estudo apresenta uma análise comparativa dos métodos LIME, Grad-CAM e SHAP para explicar predições em modelos de classificação de imagens (ResNet-50, VGG16, InceptionV3, EfficientNet-B0 e ViT-Base/16) utilizando o conjunto de dados Imagewoof. Seguindo os princípios de IA Explicável (XAI) estabelecidos , avaliamos sistematicamente a fidelidade, eficiência computacional e utilidade prática das explicações geradas. Nossos resultados demonstram que o Grad-CAM se destaca em eficiência computacional (0.34s para EfficientNet-B0) e interpretabilidade visual, enquanto o LIME, apesar de fornecer explicações granulares , exibe variações em suas explicações devido à sensibilidade a perturbações locais. O método SHAP, fundamentado na teoria dos jogos cooperativos, oferece uma base matemática mais rigorosa, utilizando mais recursos computacionais que o Grad-CAM. Experimentos comparativos como curvas de deleção progressiva, seguindo a metodologia proposta na literatura, validam que a remoção de características identificadas como importantes por esses métodos reduz significativamente a confiança do modelo. Este estudo contribui para a seleção de uma das abordagens de explicação em aplicações práticas em CNNs, destacando a variabilidade entre rigor teórico, eficiência computacional e adaptabilidade a aspectos críticos para sistemas de IA. Os resultados encontrados reforçam a necessidade de abordagens adaptativas que considerem tanto requisitos técnicos quanto contextos de aplicação específicos, particularmente em domínios sensíveis como diagnóstico médico, financeiro, como também sistemas autônomos.

***K*eywords** IA Explicável · Redes Neurais Convolucionais · Interpretabilidade · LIME · Grad-CAM · SHAP

ABSTRACT

Progress in deep learning models, particularly Convolutional Neural Networks (CNNs), has been accompanied by growing challenges regarding the interpretability of their decisions. This study presents a comparative analysis of LIME, Grad-CAM, and SHAP methods for explaining predictions in image classification models (ResNet-50, VGG16, InceptionV3, EfficientNet-B0, and ViT-Base/16) using the Imagewoof dataset. Following the principles of Explainable AI (XAI) established by Barredo, we systematically evaluate the fidelity, computational efficiency, and practical utility of the

generated explanations. Our results demonstrate that Grad-CAM excels in computational efficiency (0.34s for EfficientNet-B0) and visual interpretability, while LIME, despite providing granular explanations, the variations in its explanations due to sensitivity to local perturbations. The SHAP method, grounded in cooperative game theory , offers a more rigorous mathematical foundation while utilizing more computational resources than Grad-CAM. Comparative experiments such as progressive deletion curves, following the methodology proposed in the literature, validate that removing features identified as important by these methods significantly reduces model confidence. This study contributes to the selection of explanation approaches for practical applications in CNNs, highlighting the variability between theoretical rigor, computational efficiency, and adaptability to critical aspects of AI systems. The results found reinforce the need for adaptive approaches that consider both technical requirements and specific application contexts, particularly in sensitive domains such as medical diagnosis, financial, as well as autonomous systems.

***K*eywords** Explainable AI · Convolutional Neural Networks · Interpretability · LIME · Grad-CAM · SHAP

1 Introdução

A Inteligência Artificial (IA), e em particular o campo do Aprendizado de Máquina (*Machine Learning*), transcendeu o ambiente acadêmico para se tornar uma força transformadora em diversos setores da sociedade, desde o diagnóstico médico e a análise de risco financeiro até os sistemas de recomendação e veículos autônomos [1]. A capacidade desses sistemas de identificar padrões complexos em grandes volumes de dados impulsionou avanços significativos e abriu novas fronteiras para a inovação tecnológica.

Grande parte do avanço recente nesta área é atribuída a modelos de alta complexidade, notadamente as redes neurais profundas (*deep neural networks*), que demonstram uma capacidade preditiva sem precedentes. Contudo, essa performance superior frequentemente ocorre em detrimento da clareza, resultando em modelos denominados "caixa-preta"(*black-box*). A opacidade inerente a esses sistemas representa uma barreira crítica para sua adoção em domínios de alto risco, gerando desafios significativos relacionados à confiança, depuração de erros, e à verificação de justiça e equidade, uma vez que podem perpetuar ou até mesmo amplificar vieses sociais existentes [2, 3].

Em resposta a esses desafios, emergiu um esforço de pesquisa concentrado no desenvolvimento de uma IA mais transparente e confiável, frequentemente consolidado sob a rubrica de Inteligência Artificial Explicável (XAI - *Explainable Artificial Intelligence*) [4]. A crescente produção acadêmica e industrial, no entanto, levou a um uso por vezes inconsistente de terminologias fundamentais.

1.1 Interpretabilidade

A interpretabilidade refere-se ao grau em que um ser humano pode, de forma consistente, compreender a lógica causal subjacente às decisões ou predições de um modelo [5]. Este conceito está intrinsecamente associado a modelos que são, por sua natureza, transparentes ou de "caixa-branca"(*white-box*). Em tais arquiteturas, o mecanismo de mapeamento entre as variáveis de entrada (X) e a saída (Y) é diretamente inspecionável e inteligível.

Exemplos canônicos incluem modelos de regressão linear, nos quais os coeficientes revelam explicitamente o peso de cada característica, e árvores de decisão, cujo fluxo decisório pode ser percorrido e compreendido sem a necessidade de artifícios adicionais [6]. Nesses casos, o próprio modelo constitui a sua melhor e mais fiel explicação.

1.2 Explicabilidade

A explicabilidade, em contraste, emerge da necessidade de elucidar o comportamento de modelos opacos ou de "caixa-preta"(*black-box*), cuja complexidade intrínseca (e.g., redes neurais profundas, *gradient boosting machines*) impede a compreensão direta de seu funcionamento interno. A explicabilidade é, portanto, definida como a aplicação de um conjunto de métodos post-hoc — ou seja, posteriores ao treinamento — para gerar justificativas e razões humanamente compreensíveis para as predições de um modelo [7].

O foco da explicabilidade reside frequentemente na geração de explicações locais, que buscam justificar por que uma predição específica foi feita para uma determinada instância, em vez de descrever o comportamento global do modelo. Técnicas como LIME (*Local Interpretable Model-agnostic Explanations*), SHAP (*SHapley Additive exPlanations*) e Grad-CAM (*Gradient-Weighted Class Activation Mapping*) são exemplos proeminentes, que produzem aproximações localmente fiéis do modelo complexo para fins de diagnóstico e compreensão [8, 9, 10].

1.3 Transparência

A transparência é o conceito mais abrangente dos três, representando um atributo holístico do sistema de IA. Ela se refere à medida em que as operações de um modelo, desde a origem e o tratamento dos dados de treinamento até a lógica de sua arquitetura e o seu processo de inferência, podem ser compreendidas por um stakeholder (seja ele um desenvolvedor, usuário ou auditor), muitas vezes em resposta a regulações que preveem um "direito à explicação"[11].

Um sistema transparente é aquele que permite a um observador humano entender como e por que ele se comporta de determinada maneira, em um nível que habilite o controle, a confiança e a responsabilização (accountability) [12]. A transparência, portanto, não é uma propriedade exclusiva do modelo, mas de todo o pipeline de IA. Ela pode ser alcançada tanto pela adoção de modelos inerentemente interpretáveis (1.1) quanto pela suplementação de modelos opacos com robustas técnicas de explicabilidade (1.2).

No campo da Inteligência Artificial (IA), a necessidade de desenvolver sistemas confiáveis e auditáveis tornou imperativa a clara distinção entre os conceitos de interpretabilidade, explicabilidade e transparência. Embora frequentemente utilizados de forma intercambiável, estes termos descrevem facetas distintas do esforço para tornar os modelos de aprendizado de máquina compreensíveis para o ser humano.

1.4 Objetivos

Visando demonstrar potenciais e limitações de modelos de explicabilidade, este trabalho propõe uma avaliação comparativa entre LIME, Grad-CAM e SHAP, aplicados a 4 arquiteturas de rede neural convolucional (CNNs) pré-treinadas: VGG16, ResNet-50, InceptionV3, EfficientNet-B0, com o objetivo de avaliar o desempenho, custo computacional e poder de explicabilidade desses modelos, especificamente em imagens de cães, extraídas de um subconjunto do ImageNet.

2 Métodos

Para a realização deste trabalho, utilizou-se a linguagem de programação Python incorporada ao ambiente de desenvolvimento colaborativo Google Colaboratory [1]. Para facilitar a replicação deste trabalho, fixou-se a semente aleatória (*seed*) utilizada, além disso, também foi disponibilizado um repositório publico na plataforma GitHub de um dos autores[2].

2.1 Conjunto de Dados

O conjunto de dados utilizado para o desenvolvimento e a avaliação dos modelos de classificação de imagens neste estudo foi o ImageWoof [3]. Este dataset é um subconjunto do ImageNet [13], uma base de dados famosa para a pesquisa em visão computacional, compondo um dos principais benchmarks para novas arquiteturas.

O ImageWoof foi especificamente curado e disponibilizado pela equipe do *fast.ai* [4] com o objetivo de fornecer um benchmark que combina dois atributos desejáveis para a pesquisa aplicada: (1) ser computacionalmente mais acessível que o ImageNet completo, permitindo o treinamento e a iteração rápida de modelos complexos e (2) representa uma tarefa de classificação de grão fino (*fine-grained classification*) notavelmente desafiadora. A principal dificuldade intrínseca deste dataset advém da seleção de 10 classes de cães que possuem alta similaridade visual, que desafia e força os modelos a aprender a discriminar características sutis em vez de depender de atributos de alto nível.

As 10 classes presentes no ImageWoof são: Terrier australiano, Border terrier, Samoieda, Beagle, Shih-Tzu, Foxhound inglês, Rhodesian ridgeback, Dingo, Golden retriever e Sheepdog inglês. Mais detalhes a respeito das classes podem ser encontrados na Tabela **??**. A versão do dataset empregada neste trabalho contém um total de 12.954 imagens, com a maior dimensão sendo comprimida a 320 pixels, mantendo a proporção largura x altura. Além disso, o conjunto de dados passou por uma divisão 70/30, ou seja, 9.025 para o conjunto de treinamento e 3.929 para o conjunto de validação. Todas as imagens foram pré-processadas para uma resolução de 224x224 pixels, mantendo o padrão de entrada para as redes neurais convolucionais avaliadas. A escolha do ImageWoof, portanto, justifica-se por sua capacidade de avaliar robustamente a performance dos modelos dentro de um escopo computacionalmente viável.

[1] `https://colab.google/`

[2] `https://github.com/edmar-bevilaqua/xai_project_cmp627`

[3] `https://github.com/fastai/imagenette`

[4] `https://www.fast.ai/`

Tabela 1: Mapeamento dos Identificadores de Classe do WordNet para os Nomes das Raças de Cães no Subconjunto ImageWoof Utilizado.

ID da Classe (WordNet)	Nome da Classe (Raça)
n02086240	Shih-Tzu
n02087394	Rhodesian ridgeback (leão-da-rodésia)
n02088364	Beagle
n02089973	English foxhound (Foxhound inglês)
n02093754	Border terrier
n02096294	Australian terrier (Terrier australiano)
n02099601	Golden retriever
n02105641	Old English sheepdog (Sheepdog inglês)
n02111889	Samoyed (Samoieda)
n02115641	Dingo

2.2 Redes Neurais Convolucionais

As Redes Neurais Convolucionais (em inglês, *Convolutional Neural Networks* - CNNs) representam uma classe especializada de redes neurais artificiais, projetadas primordialmente para a análise de dados com topologia de grade, como as imagens [14]. Inspiradas no funcionamento do córtex visual humano, as CNNs se tornaram destaque em tarefas de visão computacional, incluindo classificação de imagens, detecção de objetos e segmentação semântica, devido à sua capacidade de aprender automaticamente hierarquias de características espaciais a partir dos dados brutos [15]. A respeito das redes neurais convolucionais utilizadas neste trabalho, a Tabela 2 traz um resumo das principais características de cada uma.

A VGG-16 [16] foi escolhida por seu marco histórico e por apresentar a abordagem de empilhar, de forma homogênea, camadas convolucionais de filtros pequenos (3x3) para alcançar grande profundidades, ao custo de um número elevado de parâmetros. A ResNet-50 [17] foi escolhida por introduzir o de conexões residuais (*skip connections*), mitigando o problema de *fading gradients*. A InceptionV3 [18] foi escolhida por apresentar um paradigma ligeiramente diferente, focando em redes mais largas do que profundas, usando módulos *inception*, daí seu nome. A escolha da EfficientNet-B0 [19] ocorreu por um fator cada vez mais em alta, a eficiência computacional e portabilidade de IAs para sistemas com recursos cada vez mais limitados, portanto esta é uma rede que busca otimizar o balanço entre acurácia, quantidade de parâmetros e operações de ponto flutuante (FLOPs).

Tabela 2: Comparativo das Arquiteturas de Redes Neurais Convolucionais Utilizadas

Arquitetura	Ano	Parâmetros (Aprox.)	Top-1 Acc. (ImageNet)
VGG16	2014	138 Milhões	71.3%
ResNet-50	2015	25.6 Milhões	76.0%
InceptionV3	2015	23.8 Milhões	78.8%
EfficientNet-B0	2019	5.3 Milhões	77.1%

A escolha destas arquiteturas de CNNs ocorreu por dois principais fatores: o primeiro foi a evolução arquitetural destas redes, selecionando modelos que representam marcos cronológicos e mudanças de paradigma no design de CNNs. O segundo critério foi a eficiência computacional, incluindo arquiteturas leves como a EfficientNet-B0, que foi projetada para operar em ambientes com recursos computacionais restritos.

2.3 Modelos de Explicabilidade

2.3.1 LIME (Local Interpretable Model-agnostic Explanations)

A ideia central do LIME [9] é explicar um modelo complexo de "caixa-preta", aproximando seu comportamento na vizinhança local de uma única previsão com um modelo "substituto"mais simples e inerentemente interpretável, como uma regressão linear ou uma árvore de decisão. A explicação para a previsão do modelo complexo é, então, a explicação do modelo simples. Essa abordagem é poderosa por ser agnóstica ao modelo, logo não precisa de acesso ao funcionamento interno da caixa-preta, apenas à sua função de previsão. Em vez de examinar o funcionamento interno do modelo, o LIME analisa a relação entre as entradas e saídas do modelo para gerar explicações.

O LIME [9] aproxima um modelo complexo f por um modelo interpretável g (e.g., regressão linear) em torno de uma instância x, minimizando:

$$g \in Gargmin\, \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{1}$$

com:

- $\mathcal{L}$: Perda (e.g., erro quadrático) entre f e g
- π_x: Função de proximidade que pondera amostras próximas a x
- $\Omega(g)$: Termo de regularização da complexidade de g

Para classificadores, a perda é tipicamente:

$$\mathcal{L} = \sum_{z \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z)\right)^2 \tag{2}$$

2.3.2 SHAP (SHapley Additive exPlanations)

A ideia central do SHAP [8] fundamenta a explicabilidade na estrutura matemática da teoria dos jogos cooperativos. Ele utiliza os valores de Shapley [20], que fornecem uma maneira única e justa de distribuir o "pagamento"(a previsão do modelo) entre os "jogadores"(as características de entrada). A contribuição de cada característica é sua contribuição marginal média em todas as combinações (coalizões) possíveis de características. Isso garante que a distribuição da previsão entre as características seja justa e equitativa.

Baseado nos valores de Shapley [20], a contribuição ϕ_i do recurso i para $f(x)$ é:

$$\phi_i(f, x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left(f_x(S \cup \{i\}) - f_x(S)\right) \tag{3}$$

onde:

- N: Conjunto de todos os recursos
- S: Subconjunto de recursos sem i
- $f_x(S)$: Valor esperado condicional aos recursos em S

Para modelos lineares, simplifica-se para:

$$\phi_i = \beta_i(x_i - \mathbb{E}[x_i]) \tag{4}$$

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

O Grad-CAM [10] é uma técnica específica do modelo projetada para produzir "explicações visuais"para Redes Neurais Convolucionais (CNNs). Ele gera um mapa de localização grosseiro, ou então um mapa de saliência (heatmap), destacando as regiões importantes em uma imagem de entrada que a CNN usou para fazer uma previsão de classe específica. Ele funciona utilizando a informação do gradiente que flui para a camada convolucional final da rede para produzir um mapa de localização que destaca as regiões importantes na imagem para prever o conceito. Isso torna os modelos baseados em CNN mais transparentes, sem exigir alterações na arquitetura ou retreinamento [5].

O Grad-CAM [10] calcula pesos de importância α_k^c para cada canal k da última camada convolucional, ponderados pelos gradientes da classe c:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{5}$$

onde:

- A_{ij}^k é a ativação do canal k na posição (i, j)

- y^c é a saída da classe c antes da função *softmax*
- Z é o número de elementos no mapa de ativações

O mapa de calor $L^c_{Grad-CAM}$ é obtido pela combinação linear:

$$L^c_{Grad-CAM} = ReLU\left(\sum_k \alpha^c_k A^k\right) \tag{6}$$

2.4 Pré-processamento

Para garantir a consistência dos dados de entrada (*inputs*) e a compatibilidade com as arquiteturas de Redes Neurais Convolucionais (CNNs) avaliadas, um pipeline de pré-processamento padronizado foi aplicado a todas as imagens do conjunto de dados ImageWoof antes de serem inseridas nos modelos.

O processo foi: começamos com o redimensionamento da menor aresta de cada imagem para 256 pixels, mantendo sua proporção original. Após isso, é realizado um recorte central para extrair uma região de 224x224 pixels. Esta dimensão é o padrão de entrada para a vasta maioria das arquiteturas de CNN proeminentes, incluindo VGG16, ResNet-50, InceptionV3 e EfficientNet-B0, sendo uma herança dos modelos que competiram no desafio ImageNet [21].

Após o redimensionamento, as imagens, representadas em formato PIL, são convertidas para tensores do PyTorch [22]. Esta operação realiza duas funções essenciais: converte os valores dos pixels do intervalo [0, 255], para um formato de ponto flutuante no intervalo [0.0, 1.0], e após isso, reordena as dimensões da imagem de HxWxC (Altura x Largura x Canais) para CxHxW (Canais x Altura x Largura), o formato esperado pelas camadas convolucionais avaliadas no framework.

Por fim, é realizado a normalização do tensor, em que cada canal do tensor da imagem é normalizado subtraindo-se sua respectiva média e dividindo-se pelo seu desvio padrão. Os valores utilizados para esta operação foram: média = [0.485, 0.456, 0.406] e desvio padrão = [0.229, 0.224, 0.225] para os canais R, G e B, respectivamente. A escolha destes valores não é arbitrária, uma vez que correspondem à média e ao desvio padrão do conjunto de dados ImageNet [13].

2.5 Curvas de Deleção

Para avaliar quantitativamente a fidelidade (*faithfulness*) das explicações geradas pelos métodos LIME, Grad-CAM e SHAP, foi utilizada a métrica da Curva de Deleção (*Deletion Curve*). Esta métrica é uma técnica padrão na literatura de XAI para verificar se as regiões de uma imagem destacadas como importantes por um método de explicação são, de fato, as mais influentes na decisão do modelo [4]. A hipótese é que a remoção progressiva dos pixels mais importantes, conforme indicado pelo modelo explicativo, deve levar a uma queda mais acentuada na probabilidade da classe predita do que a remoção de pixels em uma ordem aleatória ou menos importante [23].

Para os cálculos da curva de deleção neste trabalho, utilizou-se uma seleção aleatória (com *seed* fixada) de 30 imagens do conjunto de validação. Após a seleção, as imagens passaram pelo pipeline de pré-processamento disposto na seção 2.4. Após calculado a curva para todas as arquiteturas, modelos e imagens, foi-se então calculado a AUC (*Area Under the Curve*) média e o desvio padrão.

3 Trabalhos Relacionados

Com a crescente demanda por modelos de inteligência artificial explicáveis (*xAI*), diversos trabalhos vêm sendo produzidos nesta área, seja propondo novas formas e métodos de abordar a explicabilidade, aprimorando métodos já existentes ou realizando pesquisas bibliográficas extensivas para mapear e organizar o conhecimento existente.

Na vertente de pesquisa de novos modelos, pesquisadores têm se dedicado a superar algumas limitações dos métodos seminais. Por exemplo, técnicas de visualização baseadas em gradientes, como o Grad-CAM, foram estendidas em propostas como o Grad-CAM++ [24], que oferece melhores localizações de objetos múltiplos, e o Score-CAM [25], que remove a dependência dos gradientes para evitar problemas de saturação e produzir mapas de calor mais fiéis à decisão do modelo. Da mesma forma, a instabilidade das explicações do LIME motivou os seus criadores no desenvolvimento de alternativas como o Anchors [26], que busca gerar regras com escopo de aplicação mais claro e preciso.

Quanto ao SHAP, embora seu framework forneça garantias teóricas robustas, sua aplicação direta, especialmente em domínios de alta dimensão como o de imagens, ainda apresenta desafios significativos de custo computacional. A versão mais comum para CNNs, o KernelSHAP [8], depende de aproximações baseadas em perturbações de "superpixels",

o que pode impactar a fidelidade e a granularidade das explicações. Consequentemente, uma vertente expressiva da pesquisa recente tem se dedicado a especializar, otimizar e estender o SHAP para superar essas limitações. Uma linha de pesquisa foca em tornar o cálculo dos valores de Shapley mais eficiente. Propostas como o FastSHAP [27] buscam estimar os valores de Shapley de forma muito mais rápida, aproximando o modelo condicional de valor esperado com um modelo de aprendizagem de máquina, tornando a explicabilidade em tempo real mais viável. Outros trabalhos, como o Improved KernelSHAP [28], refinam o algoritmo original para melhorar a convergência e a estabilidade das estimativas. Também há uma linha que busca aprimorar a fidelidade das explicações SHAP para imagens. O PixelSHAP [29], por exemplo, foi desenvolvido para calcular os valores de Shapley em nível de pixel de forma eficiente, oferecendo uma granularidade superior à abordagem de superpixels. Abordagens híbridas também surgiram, como o Shap-CAM [30], que combina a fundamentação teórica do SHAP com a capacidade de localização do Grad-CAM, utilizando os valores de Shapley para ponderar os mapas de ativação de forma mais fiel. Métodos inspirados em princípios similares, como o BONES [31], exploram novas formas de agregar contribuições de características, demonstrando a contínua inovação em torno da teoria dos jogos cooperativos para XAI.

Na vertente de revisões sistemáticas, a rápida expansão do campo motivou a publicação de trabalhos que buscam organizar o cenário da XAI. Alguns trabalhos oferecem taxonomias detalhadas que classificam os métodos de explicação segundo critérios como escopo (global ou local), a família da técnica e o tipo de modelo ao qual se aplicam [7, 32, 33, 34]. Outras revisões focam na sistematização não apenas dos métodos, mas também das métricas utilizadas para avaliar quantitativamente a qualidade, fidelidade e robustez das explicações geradas [35]. Essa confluência de inovação metodológica e de sistematização teórica evidencia a maturidade e a relevância crescente da explicabilidade como um pilar para o desenvolvimento de uma IA confiável [36, 37].

4 Resultados

Nesta seção, apresentamos os resultados obtidos a partir da aplicação dos métodos de IA Explicável (XAI), como: LIME, Grad-CAM e SHAP em diferentes arquiteturas de redes neurais convolucionais pré-treinadas. O pipeline desenvolvida para este estudo consistiu em quatro etapas principais: (1) pré-processamento das imagens do conjunto de dados ImageWoof, (2) inferência dos modelos de classificação, (3) aplicação dos métodos de explicabilidade, e (4) avaliação quantitativa e qualitativa das explicações geradas.

4.1 Desempenho Comparativo dos Métodos XAI

A Tabela 3 apresenta uma comparação abrangente do tempo de execução e da confiança das predições para cada combinação de modelo e método de explicabilidade. Os resultados demonstram variações significativas no custo computacional entre os métodos, com o Grad-CAM destacando-se como o mais eficiente em todas as arquiteturas testadas. Vale ressaltar que, o tempo e custo de processamento, podem sofrer alterações direta se alterados a máquina e o processador utilizado.

Tabela 3: Desempenho comparativo de métodos XAI em arquiteturas CNN

Modelo	Confiança (%)	Grad-CAM (s)	LIME (s)	GradientSHAP (s)
ResNet-50	50.5 %	2.19	202.93	2.89
VGG16	97.5 %	2.16	617.63	6.29
InceptionV3	100.0 %	0.79	165.19	1.91
EfficientNet-B0	90.9 %	0.34	59.73	0.83

O Grad-CAM mostrou-se consistentemente mais rápido, com tempo de execução variando entre 0.34s (EfficientNet-B0) e 2.19s (ResNet-50), corroborando sua eficiência computacional conforme destacado por Selvaraju[10]. Em contraste, o LIME apresentou os maiores tempos de execução, especialmente para arquiteturas mais complexas como VGG16 (617.63s), devido à sua abordagem baseada em perturbações locais [9].

4.2 Avaliação da Fidelidade das Explicações

Para avaliar a fidelidade das explicações geradas, empregamos a metodologia de curvas de deleção progressiva. Esta abordagem remove gradualmente as características identificadas como mais importantes por cada método XAI e monitora o impacto na confiança da predição do modelo. A Figura 1 mostra um exemplo específico para o EfficientNet-B0, enquanto as Figuras 2a–2d apresentam as curvas médias para todas as arquiteturas avaliadas.

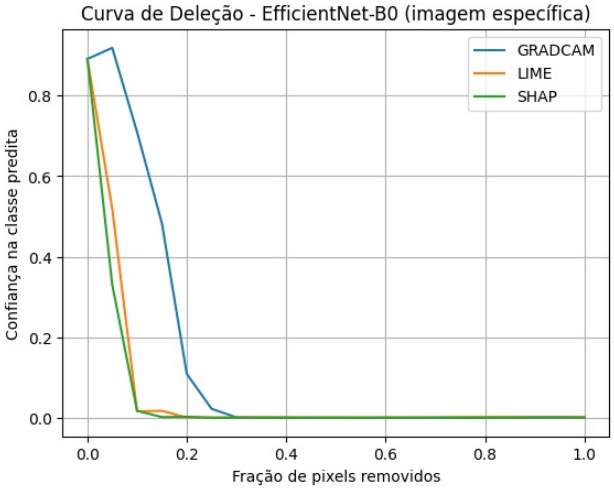
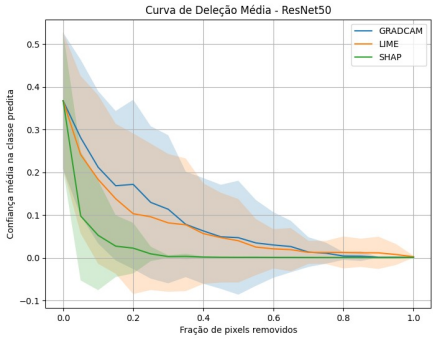
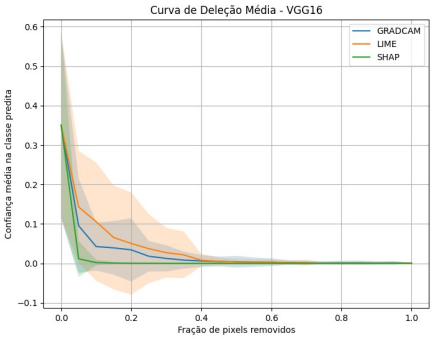


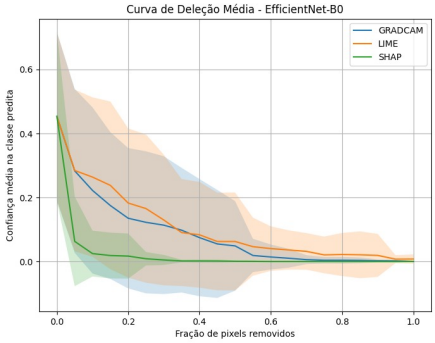
Figura 1: Curva de deleção para uma imagem específica no EfficientNet-B0, mostrando o impacto na confiança do modelo à medida que os pixels mais importantes (segundo cada método) são removidos.



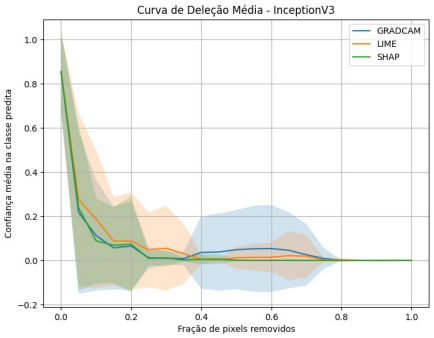
(a) ResNet-50



(b) VGG16



(c) EfficientNet-B0



(d) InceptionV3

Figura 2: Curvas de deleção média para diferentes arquiteturas de CNN, mostrando o comportamento comparativo dos métodos XAI.

4.2.1 Análise por Arquitetura

- **ResNet-50 (Figura 2a)**: Observa-se que o SHAP apresenta o declínio mais acentuado na confiança do modelo à medida que os pixels mais importantes são removidos, indicando alta fidelidade na identificação das regiões críticas. O Grad-CAM mostra um comportamento intermediário, enquanto o LIME apresenta variações mais suaves.

- **VGG16 (Figura 2b)**: Nesta arquitetura mais profunda, o Grad-CAM demonstra melhor desempenho na identificação de regiões importantes, com queda mais consistente na confiança. O SHAP mantém boa performance, mas com maior variabilidade entre instâncias.

- **EfficientNet-B0 (Figura 2c)**: Para esta arquitetura eficiente, os três métodos apresentam comportamentos similares, com o LIME mostrando ligeira vantagem na identificação de características discriminativas, especialmente na faixa de 40-60% de pixels removidos.

- **InceptionV3 (Figura 2d)**: A arquitetura Inception mostra a maior diferença entre métodos, com o SHAP destacando-se claramente na identificação de regiões críticas, seguido pelo Grad-CAM. O LIME apresenta menor consistência nesta arquitetura.

(a) Grad-CAM (b) LIME (c) SHAP

Figura 3: Exemplo visual da remoção de 30% dos pixels mais importantes identificados por cada método (EfficientNet-B0).

As áreas removidas demarcadas em preto, correspondem às regiões consideradas mais relevantes por cada técnica. Logo, podemos identificar as adaptações de cada método. Na Tabela 4 complementa esta análise com os valores quantitativos de AUC (Área Sob a Curva) para cada combinação de modelo e método, confirmando as tendências observadas visualmente nas curvas de deleção.

Tabela 4: Resultados de AUC Média e Desvio Padrão por Modelo e Método de Explicabilidade

Modelo	**Método**	**AUC (média)**	**AUC (std)**
EfficientNet-B0	SHAP	0.019	0.023
EfficientNet-B0	Grad-CAM	0.081	0.097
EfficientNet-B0	LIME	0.103	0.100
InceptionV3	SHAP	0.046	0.045
InceptionV3	Grad-CAM	0.061	0.075
InceptionV3	LIME	0.065	0.056
ResNet50	SHAP	0.021	0.023
ResNet50	LIME	0.063	0.071
ResNet50	Grad-CAM	0.081	0.075
VGG16	SHAP	0.010	0.006
VGG16	Grad-CAM	0.023	0.022
VGG16	LIME	0.034	0.038

4.3 Análise Qualitativa das Explicações

A análise qualitativa revelou diferenças marcantes na natureza das explicações geradas:

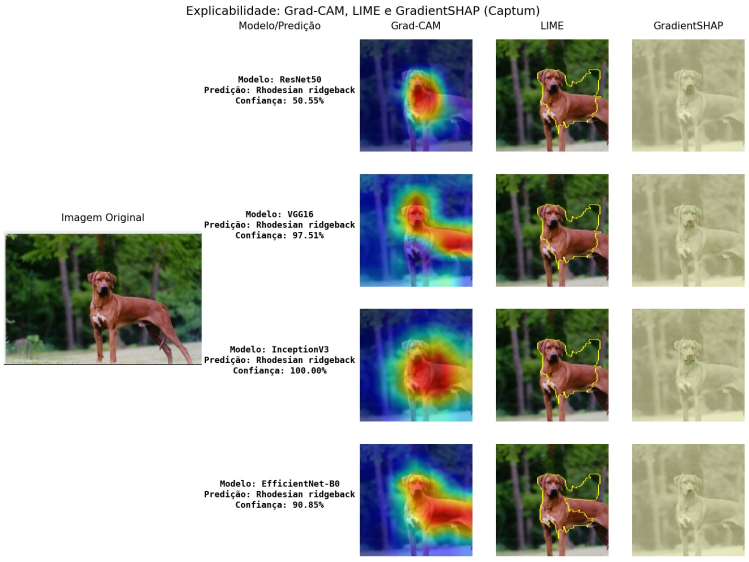


Figura 4: Resultados obtidos após avaliação de predição dos modelos, e comparação de método XAI

- **Grad-CAM**: Produziu mapas de calor que destacam regiões ativadas nas camadas convolucionais finais, proporcionando uma visão intuitiva das áreas da imagem que mais contribuíram para a decisão do modelo [10]. Este método mostrou-se particularmente eficaz para modelos como InceptionV3, EfficientNet-B0 e VGG16.
- **LIME**: Gerou explicações baseadas em superpixels, identificando tanto características positivas quanto negativas para a classificação [9]. No entanto, observamos certa instabilidade nas explicações devido à natureza estocástica do método.
- **SHAP**: Ofereceu explicações mais matematicamente fundamentadas, atribuindo valores de importância específicos para cada pixel ou região da imagem [8]. Apesar do maior custo computacional em comparação com o Grad-CAM, o SHAP proporcionou insights mais detalhados sobre as interações entre características.Assim como o LIME, o modelo de SHAP apresenta certa instabilidade nas explicações devido à sua natureza estocástica.

4.4 Limitações e Desafios

Nossos experimentos também revelaram limitações importantes:

- O LIME apresentou alta variabilidade em suas explicações quando executado com diferentes sementes aleatórias, um problema já documentado por M. Ribeiro[9].
- O SHAP, embora teoricamente robusto, mostrou-se computacionalmente exigente, especialmente para imagens de alta resolução, conforme discutido por Scott M. Lundberg[8].
- O Grad-CAM, apesar de eficiente, fornece explicações menos granulares em comparação com os outros métodos, limitando-se a destacar regiões amplas da imagem [10].

5 Conclusão

Diante do exposto, para a tarefa de classificação de imagens de forma detalhada (*fine-grained*), reafirmou-se que não existe uma solução única. Ao avaliar sistematicamente a eficiência computacional, a fidelidade e a natureza das explicações geradas, demonstramos que, em alguns casos, um método pode ser mais eficaz que outro. A escolha do método de XAI mais apropriado depende intrinsecamente de um equilíbrio teórico, do custo computacional e dos requisitos específicos da aplicação.

Para nosso caso o **Grad-CAM** se destaca como a abordagem computacionalmente mais eficiente, gerando explicações visuais intuitivas em um tempo muito menor do que o exigido pelos outros métodos. Essa característica o torna ideal para cenários que demandam análise rápida ou para integração em sistemas com recursos computacionais limitados. Porém, a granularidade limitada do método restringe a análise a uma identificação mais geral das áreas de interesse, sem detalhar os atributos específicos que influenciaram a decisão. O **SHAP**, por sua vez, confirmou sua robustez teórica,

baseada na teoria dos jogos. As explicações geradas por ele foram mais fiéis, como validado pelas curvas de deleção progressiva, e capazes de atribuir com maior precisão a contribuição de cada característica para a decisão do modelo. Essa profundidade analítica, no entanto, implica um custo computacional significativamente maior, o que pode ser um fator limitante para aplicações em tempo real ou em larga escala. Já o **LIME**, embora útil por sua capacidade de identificar características positivas e negativas para uma predição de forma independente do modelo, apresentou instabilidade notável em suas explicações. A sensibilidade do método a distúrbios locais e à aleatoriedade do processo de amostragem exige cautela em sua aplicação, especialmente em domínios críticos, onde a consistência das explicações é fundamental.

Referências

[1] Michael I Jordan. Machine learning: trends, perspectives and challenges. In *Proceedings of ACM Turing Celebration Conference-China*, pages 8–8, 2018.

[2] Cathy O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2017.

[3] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[4] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.

[5] Molnar Christoph. Interpretable machine learning: A guide for making black box models explainable. 2020.

[6] Trevor Hastie. The elements of statistical learning: data mining, inference, and prediction, 2009.

[7] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 6 2020.

[8] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[9] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?"explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 13-17-August-2016, pages 1135–1144. Association for Computing Machinery, 8 2016.

[10] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. 10 2016.

[11] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

[12] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[14] Ian Goodfellow. Deep learning, 2016.

[15] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[19] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[20] Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. 12 2019.

[23] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

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

[25] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020.

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[27] Neil Jethani, Mukund Sudarshan, Ian Covert, Su-In Lee, and Rajesh Ranganath. Fastshap: Real-time shapley value estimation. 7 2021.

[28] Ian Covert and Su-In Lee. Improving kernelshap: Practical shapley value estimation via linear regression. 12 2020.

[29] Roni Goldshmidt. Attention, please! pixelshap reveals what vision-language models actually focus on. *arXiv preprint arXiv:2503.06670*, 2025.

[30] Quan Zheng, Ziwei Wang, Jie Zhou, and Jiwen Lu. Shap-cam: Visual explanations for convolutional neural networks based on shapley value. 8 2022.

[31] Davide Napolitano and Luca Cagliero. Bones: a benchmark for neural estimation of shapley values. 7 2024.

[32] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[33] Philine Lou Bommer, Marlene Kretschmer, Anna Hedström, Dilyara Bareeva, and Marina M.-C. Höhne. Finding the right xai method—a guide for the evaluation and ranking of explainable ai methods in climate science. *Artificial Intelligence for the Earth Systems*, 3, 3 2024.

[34] Mohamed Karim Belaid, Eyke Hüllermeier, Maximilian Rabus, and Ralf Krestel. Do we need another explainable ai method? toward unifying post-hoc xai evaluation methods into an interactive and multi-dimensional benchmark. 6 2022.

[35] Felix Tempel, Daniel Groos, Espen Alexander F. Ihlen, Lars Adde, and Inga Strümke. Choose your explanation: A comparison of shap and gradcam in human activity recognition. 12 2024.

[36] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

[37] Waddah Saeed and Christian Omlin. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. 11 2021.