Analysis of Visual Explainers Using Clustering Techniques

Lázaro Raimundo de Oliveira[1],
João Carlos Xavier Júnior[1], **Anne M.P. Canuto**[2]

[1]Instituto Metrópole Digital - Universidade Federal do Rio Grande do Norte
Natal, RN - Brasil

[2]Departamento de Informática e Matemática Aplicada - Universidade Federal
do Rio Grande do Norte, Natal, RN - Brasil

lazaro.oliveira.813@ufrn.edu.br,

jcxavier@imd.ufrn, anne.canuto@ufrn.br

Abstract. *Artificial Intelligence (AI) is increasingly integrated into everyday life, heightening the need for methods that explain model decisions — a field known as Explainable Artificial Intelligence (XAI). In visual explainability, numerous techniques and metrics have been proposed; however, there is still no consensus on how to evaluate them in a comparative and reliable manner. This work proposes a novel methodology to analyze the behavioral consistency of explainers using clustering techniques, allowing for the simultaneous evaluation of multiple explainers and metrics. The results indicate that consistency varies according to the dataset and the model employed.*

Resumo. *A Inteligência Artificial (IA) tem se tornado cada vez mais presente no cotidiano, ampliando a demanda por métodos que expliquem as decisões tomadas por seus modelos — o que se convencionou chamar de Explainable Artificial Intelligence (XAI). No campo da explicabilidade visual, diversas técnicas e métricas vêm sendo propostas, no entanto, ainda não há consenso sobre como avaliá-las de forma comparativa e confiável. Este trabalho propõe uma nova metodologia para análise de comportamento de explicadores, com base em técnicas de agrupamento (clustering), permitindo avaliar múltiplos explicadores e métricas simultaneamente. Os resultados indicam que o comportamento entre explicadores varia conforme o conjunto de dados e o modelo de explicação utilizado.*

1. Introdução

A Inteligência Artificial (IA) é, possivelmente, a mais significativa criação humana da era moderna. Diversas definições têm sido propostas para essa área da Ciência da Computação, sendo uma delas a capacidade de sistemas ou dispositivos exibirem comportamentos cognitivos com algum grau de semelhança ao humano [Russell and Norvig 2009]. Essa capacidade tem sido intensamente buscada ao longo dos últimos 50 anos, por meio da integração de soluções oriundas de diversas áreas do conhecimento, como matemática, estatística, computação, engenharia, biologia e linguística. Avanços mais recentes foram impulsionados por técnicas de *deep learning* [Zhang et al. 2023]. Nesse cenário de complexidade crescente, o uso de métricas para

avaliar o desempenho de modelos inteligentes tornou-se não apenas comum, mas necessário.

Se, por um lado, os avanços da IA estão cada vez mais presentes no cotidiano, por exemplo no trabalho, nos estudos e no lazer, por outro, cresce a demanda por explicações sobre as decisões tomadas por seus algoritmos. Esse campo emergente, conhecido como *Explainable AI* (XAI), tem ganhado destaque nos últimos anos, especialmente quando aplicado a modelos de *deep learning*, frequentemente chamados de "caixas-pretas"devido à sua opacidade interpretativa.

A busca por explicabilidade visa, entre outros objetivos: mitigar os riscos associados a decisões errôneas tomadas por algoritmos; auxiliar os indivíduos na tomada de decisões; verificar a aderência dos modelos a valores humanos relevantes; e aprimorar a qualidade de produtos baseados em IA [Arrieta et al. 2020]. Além disso, órgãos governamentais reconhecem que a explicabilidade em IA é um direito da sociedade. A União Europeia, por exemplo, aprovou em 2024 o Regulamento da Inteligência Artificial[1], que inclui como princípio o aumento da confiança pública e a redução da insegurança jurídica. No Brasil, o Projeto de Lei 2338/2023 [2] e a Resolução nº 615/2025 do Conselho Nacional de Justiça [3] determinam que os sistemas de IA devem observar, entre outros princípios, a transparência, a explicabilidade, a inteligibilidade e a auditabilidade.

Em tarefas como a classificação de imagens, por exemplo, compreender as decisões de modelos baseados em *deep learning* exige o uso de técnicas de explicabilidade visual que indicam quais regiões da imagem (pixels) foram relevantes para a predição realizada. No entanto, ainda não há consenso na literatura sobre o que constitui uma "boa explicação", tampouco sobre quais métodos ou métricas devem ser adotados para avaliá-las.

A falta de padronização dificulta aplicações em áreas sensíveis. Na medicina, por exemplo, a explicabilidade pode auxiliar na identificação de áreas afetadas por doenças como o câncer, orientando decisões clínicas importantes. Na engenharia, fissuras estruturais detectadas por IA podem ser mais bem delimitadas com explicações visuais apropriadas. Apesar disso, muitos estudos propõem novos explicadores e métricas sem considerar comparações sistemáticas com métodos existentes [Arras et al. 2021]. Quando o fazem, limitam-se a um conjunto restrito de métricas.

Diante dessas lacunas, este trabalho tem como objetivo principal avaliar em que medida há uma divergência sistemática de comportamento entre diferentes explicadores, por meio de uma nova metodologia baseada em técnicas de Agrupamento (*Clustering*). A proposta consiste na construção de um dataset onde as métricas de explicação formam as colunas (atributos) e as explicações locais (geradas para cada imagem) formam as linhas (instâncias). A partir dessa estrutura, são aplicados algoritmos de agrupamento para identificar padrões nos comportamentos dos explicadores [Mersha et al. 2024]. Espera-se que esses padrões contribuam para a análise da confiabilidade dos métodos de explicação aplicados a modelos de classificação de imagens e, possivelmente, a outras tarefas.

---

[1] https://www.consilium.europa.eu/pt/policies/artificial-intelligence/

[2] https://www25.senado.leg.br/web/atividade/materias/-/materia/157233

[3] https://atos.cnj.jus.br/files/original1555302025031467d4517244566.pdf

Este artigo está organizado da seguinte forma: a Seção 2 apresenta o referencial teórico necessário para a compreensão do estudo; a Seção 3 discute os trabalhos correlatos encontrados na literatura; a Seção 4 descreve os principais aspectos da metodologia experimental; a Seção 5 apresenta e discute os resultados obtidos; e, por fim, a Seção 6 traz as considerações finais, destacando contribuições, limitações e possíveis direções futuras.

2. Referencial Teórico

2.1. Aprendizado de máquina

Grande parte das abordagens contemporâneas em Inteligência Artificial baseia-se em um processo denominado Aprendizado de Máquina (*Machine Learning*). Essa metodologia envolve o uso de algoritmos capazes de identificar padrões em conjuntos de dados com o objetivo de responder, com determinado grau de confiança, a questões relacionadas a essas informações ou realizar previsões sobre novos dados.

O Aprendizado de Máquina pode ser subdividido em diferentes categorias. Neste trabalho, são considerados dois paradigmas tradicionais: o aprendizado supervisionado e o não supervisionado. O primeiro refere-se à aprendizagem a partir de um conjunto de dados rotulados, nos quais existe uma variável-alvo conhecida que se deseja prever com base nas demais variáveis. Já o aprendizado não supervisionado opera sem a presença de rótulos, buscando explorar a estrutura intrínseca dos dados, como agrupamentos ou padrões ocultos [Zhang et al. 2023].

2.2. Agrupamento de Dados

O *Clustering* ou Agrupamento de Dados é uma técnica pertencente ao campo do aprendizado não supervisionado, cujo objetivo é particionar um conjunto de dados em grupos (ou clusters) de forma que os elementos dentro de um mesmo grupo sejam internamente semelhantes, enquanto os diferentes grupos apresentem entre si elevada dissimilaridade [Han et al. 2011].

Os métodos de *clustering* podem ser classificados de acordo com a estratégia de separação dos grupos, sendo as principais categorias: métodos baseados em centralidade, hierarquia, densidade e modelos probabilísticos. Este trabalho utiliza o método k-Means, pertencente à categoria dos algoritmos baseados em centralidade. Tais métodos são, em geral, eficientes para conjuntos de dados com número pequeno a moderado de instâncias e atributos [Han et al. 2011]. O algoritmo k-Means tende a formar grupos cujos elementos estão a uma distância relativamente pequena de um ponto central, o qual representa o centróide do *cluster*.

2.3. Redes Neurais

As Redes Neurais Artificiais (*Artificial Neural Networks* – ANNs) são inspiradas no funcionamento do sistema nervoso biológico, especialmente em virtude da estrutura interconectada de seus nós computacionais, organizados em camadas. As camadas situadas entre a camada de entrada e a de saída são denominadas camadas ocultas, responsáveis por processar representações intermediárias dos dados.

As Redes Neurais Convolucionais (*Convolutional Neural Networks* – CNNs) representam uma evolução dessa arquitetura, incorporando mecanismos de aprendizado

baseados em convoluções que possibilitam a extração automática de características relevantes, especialmente em dados com estrutura espacial, como imagens. As CNNs se destacam por sua capacidade de reduzir significativamente o número de parâmetros do modelo em comparação com redes totalmente conectadas, o que favorece sua escalabilidade e desempenho [O'Shea and Nash 2015].

Essas redes são amplamente empregadas em tarefas de visão computacional, como classificação de imagens, detecção de tumores, reconhecimento facial e identificação de objetos [Witten et al. 2025]. A arquitetura típica de uma CNN é composta por quatro tipos principais de camadas: camadas convolucionais, que extraem padrões locais; camadas de pooling, que realizam a redução da dimensionalidade; funções de ativação, que introduzem não linearidade ao modelo; e camadas totalmente conectadas, que integram as informações extraídas e realizam a predição final.

2.4. Métodos de Explicação

Com o avanço dos modelos de *Deep Learning* e sua crescente adoção em cenários críticos, como medicina, direito e engenharia, surgiu a demanda por sistemas que não apenas apresentem alto desempenho, mas também forneçam justificativas compreensíveis para suas decisões. Esse desafio motivou o surgimento da área conhecida como Inteligência Artificial Explicável (*Explainable Artificial Intelligence* – XAI), cujo objetivo é desenvolver métodos que tornem os modelos de IA mais transparentes, auditáveis e confiáveis [Arrieta et al. 2020].

Dessa forma, a XAI busca construir caminhos confiáveis entre a complexidade dos modelos e a capacidade humana de interpretação. Métodos explicáveis permitem entender quais atributos ou regiões de uma imagem contribuíram para uma determinada predição, promovendo maior confiança, capacidade de auditoria, e também suporte à tomada de decisão humana.

Trabalhos encontrados na literatura em XAI apresentam uma ampla variedade de técnicas, que podem ser classificadas de acordo com diferentes critérios: (i) Nível de escopo (global ou local); (ii) Momento da explicação (ante hoc ou pós hoc); (iii) Tipo de modelo (específico ou agnóstico); (iv) Modalidade dos dados (texto, imagem, tabular, etc.).

Em se tratando de imagens, um método de explicabilidade, um gerador de explicações, um explicador [Ali et al. 2023] — também chamado de modelo de explicabilidade, técnica de explicabilidade ou XAI Method (do inglês) [Bommer et al. 2024] — são termos usados no contexto de explicabilidade visual para se referir a funções que estimam as regiões consideradas relevantes na tomada de decisão de um modelo classificador. Neste trabalho, será adotado o termo explicador, por ser mais simples e direto.

Dessa forma, métodos de explicabilidade (ou explainers) possibilitam a identificação de como um modelo interpreta uma determinada entrada, por meio da análise das ativações em suas camadas internas. A Figura 1 apresenta um exemplo de expectativa ou "visão" de um modelo, por exemplo na imagem original um cachorro de perfil, onde é visto apenas um olho. Ao gerar a imagem de explicabilidade do modelo, percebe-se o delinear de 2 círculos na posição dos olhos, sugerindo que o modelo espera uma imagem de um cachorro com dois olhos (área marcada com quadrado na imagem). Neste trabalho foram considerados dois explicadores para compor uma análise de con-

sistência de comportamento, sendo eles: (DeepLift [Shrikumar et al. 2017] e GradCAM [Selvaraju et al. 2020]).

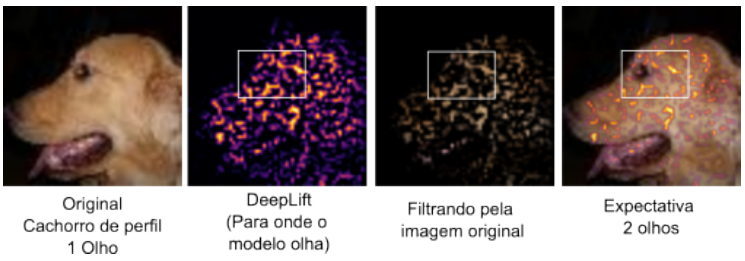

Figura 1. Análise de visão do modelo conforme Explicador DeepLift

2.5. Métricas de Explicabilidade

As métricas de explicabilidade, conforme discutido por [Bommer et al. 2024] e [Hedström et al. 2023], podem ser classificadas de acordo com a principal característica que avaliam:

- **Fidelidade** — Avalia o quanto as características destacadas pelos explicadores afetam de fato as decisões do modelo de classificação. Quanto maior, melhor (↑).
- **Robustez** — Mede o grau de variação das explicações diante de pequenas perturbações na entrada. Explicadores mais estáveis sofrem menos alterações. Quanto menor, melhor (↓).
- **Complexidade** — Avalia o quão concisas são as explicações geradas. Explicações mais simples tendem a ser mais interpretáveis. Quanto menor, melhor (↓).
- **Localização** — Verifica o quanto da região apontada pelos explicadores coincide com uma área de referência (ground truth). Quanto maior a sobreposição, melhor (↑).
- **Randomização** — Testa a resistência do explicador à degradação das explicações quando os dados são embaralhados ou aleatorizados. Quanto maior, melhor (↑).
- **Axioma** — Mede o quanto as explicações obedecem a propriedades formais (axiomas) desejáveis. Quanto maior o atendimento aos axiomas, melhor (↑).

Grupo	Métricas
Fidelidade	Faithfulness Correlation, Faithfulness Estimate, Pixel-Flipping, Region Segmentation, Selectivity, SensitivityN, IROF, ROAD, Infidelity e Sufficiency;
Robustez	Max-Sensitivity, Avg-Sensitivity, Consistency, Relative Input Stability, Relative Output Stability e Relative Representation Stability;
Complexidade	Complexity e Sparseness;
Localização	Pointing Game, Top-K Intersection, Relevance Mass Accuracy, Relevance Rank Accuracy, Attribution Localisation e AUC;
Randomização	MPRT, Smooth MPRT, Efficient MPRT e Random Logit.

Tabela 1. Lista de Métricas de Explicabilidade

Essas métricas possuem escalas e direções distintas. Para viabilizar comparações adequadas, é necessário padronizar os valores e, quando aplicável, inverter o sentido de

algumas métricas, de modo que todas apontem na mesma direção (quanto maior, melhor). As métricas adotadas neste trabalho estão listadas na Tabela 1, conforme categorização apresentada em [Hedström et al. 2023].

É importante ressaltar que as métricas axiomáticas não foram utilizadas nesta análise devido ao alto custo computacional de processamento. Logo, foi utilizado um total de 28 métricas.

3. Trabalhos Relacionados

Foi realizada uma revisão sistemática da literatura com o objetivo de identificar trabalhos recentes e suas principais contribuições para a área de *Explainable Artificial Intelligence* (XAI). O foco da pesquisa foi direcionado às bibliotecas e *toolkits* que disponibilizassem implementações de explicadores e métricas de avaliação. A expressão de busca "*XAI Visual Explain Metric Toolkit*" foi aplicada no Google Scholar, sendo considerados os 100 primeiros resultados com PDF disponível. Todos os trabalhos encontrados foram analisados com base nas questões descritas na Tabela 2.

Índice	Questão
Q1	O artigo é de revisão, benchmarking ou outro tipo?
Q2	Em qual base ou catálogo o artigo está indexado?
Q3	Se for revisão, quantos artigos foram inicialmente considerados e quantos permaneceram após a triagem?
Q4	Qual o período coberto pela revisão?
Q5	O artigo propõe uma taxonomia ou nomenclatura do tema?
Q6	Quantos e quais toolkits ou bibliotecas são mencionados?
Q7	O artigo menciona explicadores?
Q8	O artigo menciona métricas de explicabilidade?
Q9	Quais *datasets* foram utilizados?

Tabela 2. Questões de revisão utilizadas na análise dos trabalhos

A análise das questões Q1 a Q5 revelou uma diversidade de trabalhos com revisões recentes e abrangentes, de tal forma que foram encontrados 11 artigos de revisão e 12 de benchmarking (Q1). Os artigos analisados pertencem a catálogos como Springer Link, arXiv, IEEE Xplore, ScienceDirect, ACM Digital Library, MDPI, AMETSOC, Wiley Online Library, NeurIPS, The CVF, JMLR, entre outros (Q2).

O trabalho de maior escopo identificou inicialmente 289 artigos, mantendo um subconjunto após triagem (Q3). O período abrangido pelas revisões vai de 2016 a 2024 (Q4). Quase metade dos artigos incluíram alguma forma de nomenclatura ou taxonomia (Q5). O toolkit mais citado foi o Quantus[4] [Hedström et al. 2023], seguido pelo AIX360[5] [Arya et al. 2019] (Q6). A maioria dos trabalhos (87%) faz referência a explicadores (Q7), enquanto 54% discutem métricas de explicabilidade (Q8). Os *datasets* mais frequentemente citados foram o *ImageNet* e o *CIFAR-10*. Além desses, foram identificados conjuntos específicos para áreas como saúde, ou artificiais, criados para demonstrar propriedades específicas (Q9).

[4]https://quantus.readthedocs.io/en/latest/

[5]https://github.com/Trusted-AI/AIX360

O estudo de [Bommer et al. 2024] investiga o uso de métricas na escolha de explicadores para modelos climáticos, utilizando a biblioteca Quantus. Uma das metodologias utilizadas é o *Spider Plot*, também comum em análises de agrupamento .Por outro lado, em [Miró-Nicolau et al. 2023], os autores compararam diversos explicadores com foco na métrica de fidelidade, identificando agrupamentos de desempenho.

Muitos estudos fazem comparações de explicadores sem considerar a influência do dataset, tratando os resultados como avaliação do explicador. Em contraste, [Hedström et al. 2023] propõe que a explicação é uma função dos parâmetros modelo e do dataset. Logo, como é possível notar, não houve consenso nas abordagens dos diversos autores no que diz respeito a combinação de explicadores e métricas.

3.1. Discussão

O toolkit Quantus, desenvolvido com suporte às bibliotecas *TensorFlow* e *PyTorch*, foi o mais citado nos artigos analisados. Ele oferece 29 explicadores e 33 métricas de avaliação, com suporte a CPU e GPU, além de uma documentação rica e atualizada[6]. Esses fatores motivaram sua escolha para a implementação deste trabalho.

A análise de diferentes arquiteturas de redes neurais pode contribuir para uma compreensão mais ampla do comportamento dos explicadores. Diversos trabalhos mencionam arquiteturas como *ResNet* [He et al. 2015], *EfficientNet* [Tan and Le 2020] e *Swin Transformer* [Liu et al. 2021]. Neste trabalho, optou-se pelo uso do modelo *EfficientNet-B1*, por oferecer boa eficiência sem requerer hardware computacional oneroso [Tan and Le 2020].

Em se tratando de base de imagens, foi escolhido o *ImageNet* que é um dataset com mais de 3 milhões de imagens de aproximadamente 400 x 350 pixels, organizadas em mais de 5000 classes [Deng et al. 2009]. Para fins de teste e avaliação, é comum o uso de versões reduzidas, como o *ImageNet Tiny* [7], que possui 100 mil imagens coloridas de 64x64 pixels, classificadas em 200 classes. Essa versão inclui dados de segmentação, o que a torna adequada para avaliação de métricas de explicabilidade, além de ser mais leve do ponto de vista computacional.

Por último, foram encontrados poucos artigos que utilizam técnicas de clusterização de explicadores com base em métricas. Contudo, em [Hedström et al. 2023] os autores empregam métricas típicas de agrupamento, como IAC (*Intra-Agreement Consistency*) e IEC (*Inter-Agreement Consistency*), além de gráficos de radar e de barras, para comparar o desempenho de grupos de explicadores. Baseado em tudo que foi discutido aqui, serão utilizados 2 explicadores, 28 métricas, algoritmos de agrupamento e várias formas de visualização para representar o desempenho dos explicadores analisados neste trabalho. A principal distinção entre este trabalho e o estudo de [Hedström et al. 2023] reside no enfoque metodológico adotado. Enquanto o trabalho citado menciona algumas técnicas típicas de clusterização, sugerindo a possibilidade de uma análise baseada em agrupamento, tais algoritmos não foram, de fato, aplicados. Em contrapartida, o presente estudo adota explicitamente uma abordagem fundamentada na clusterização, empregando diretamente algoritmos de agrupamento como base de sua metodologia.

[6] `https://github.com/understandable-machine-intelligence-lab/Quantus`
[7] `https://image-net.org/`

4. Metodologia Experimental

Neste trabalho, utilizou-se o dataset *Tiny ImageNet*, que contém 100.000 imagens coloridas no formato 64x64 pixels, distribuídas em 200 classes, com 500 imagens por classe. O conjunto foi dividido em treinamento (80%) e validação (20%).

Como o objetivo é investigar a explicabilidade do funcionamento de modelos de classificação de imagens, especialmente redes neurais convolucionais (CNNs), optou-se pelo modelo *EfficientNet-B1*. Essa arquitetura é composta por 186 camadas, totalizando aproximadamente 7,9 milhões de parâmetros. O modelo foi inicializado com pesos pré-treinados(ImageNet-1k,240x240) e submetido a ajuste fino (*fine-tuning*) na ImageNet tiny, alcançando uma acurácia de 69,42% em 20 épocas. Segundo [Luo et al. 2021], com otimizações adequadas, o *EfficientNet-B1* pode atingir para a ImageNet-1k uma acurácia de até 79,1%(Top-1) chegando a 84,39% com o modelo EfficientNet-B7.

4.1. Construção da base de dados para o agrupamento

Para a construção da base de dados foram selecionados aleatoriamente 2408 imagens do conjunto de validação utilizado para fazer ajuste fino no modelo EfficientNet-B1. Os explicadores GradCAM(explicador tradicional) e DeepLift(proposto como melhor que o GradCam e outros) foram utilizados para interpretar a classificação de cada uma dessas imagens pelo modelo e para cada explicador uma instância foi definida, totalizando 4.816 instâncias. Cada instância contêm 30 atributos, que são: explicador usado, o resultado da classificação pelo modelo(C/E), e 28 colunas com os valores de cada métrica.

A base resultante continha apenas 4,7% de valores nulos nas métricas Relative Input Stability, Relative Output Stability, Relative Representation Stability que pela natureza da métrica esses valores foram preenchidos com 0. Os valores de cada variável ou métrica foram normalizados para 0 intervalo 0 e 1 a fim de ficarem na mesma escala.

4.2. Metodologia

O pipeline de processamento e execução seguiu as etapas ilustradas na Figura 2. A sua execução foi realizada em um computador com CPU AMD Ryzen 5 5600, 16 GB de RAM, SSD e GPU NVIDIA RTX 3060 com 12 GB de memória dedicada. A execução total levou aproximadamente uma semana.

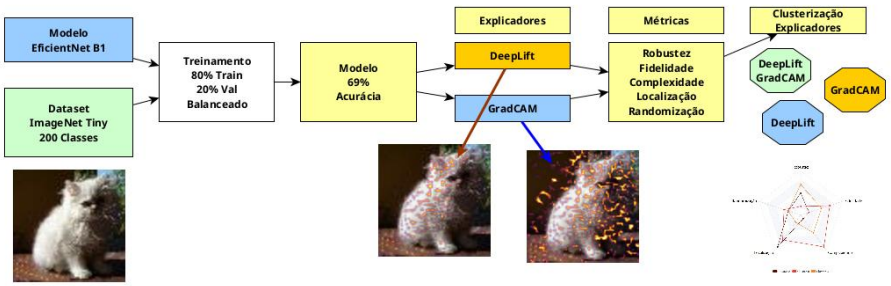

Figura 2. Fluxograma do pipeline experimental

Em relação à análise de agrupamento (*Cluster Analysis*), inicialmente, foram utilizadas duas técnicas: k-Means e DBSCAN [Aggarwal and Reddy 2018].

A validação dos resultados foi realizada com o uso de dois índices amplamente adotados na literatura: o índice de Silhouette(S) [Rousseeuw 1987] e o índice de Davies-Bouldin(DB)[Davies and Bouldin 1979]. No compativo da técnica de clusterização optou-se pelo K-Means(S:0,3927;DB:1,0813) ao invés do DBSCAN(S:0,3919;DB:1,7529) por apresentar resultado melhor em relação ao DB.

5. Resultados Experimentais

A quantidade ideal de grupos (*clusters*) para o método k-Means foi determinada por meio da análise dos índices de validação Silhouette e Davies-Bouldin (DB). Como mostra a Figura 3, ambos os índices indicaram o valor $k = 2$ como o mais adequado. Para complementar essa escolha, foi realizada uma análise de componentes principais (PCA) com visualização tridimensional, que também sugeriu a existência de dois agrupamentos bem definidos. A validade dessa visualização foi reforçada pela verificação de que os três primeiros componentes principais explicavam 68,03% da variabilidade total dos dados (Figura 3C).

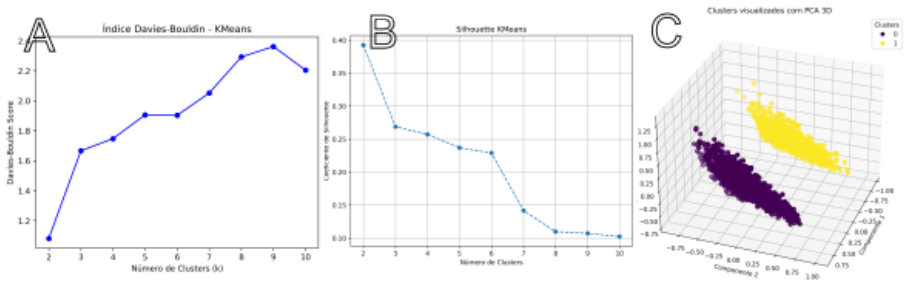

Figura 3. Análise de número de clusters via Silhouette (A), Davies-Bouldin (B) e PCA 3D (C)

Dado o número de clusters sugerido, é natural associar os dois grupos à presença dos dois explicadores utilizados (*GradCAM* e *DeepLift*). No entanto, como todas as imagens foram avaliadas por ambos os explicadores, não há, a priori, um fator determinante que vincule diretamente um explicador a um cluster específico. Ainda assim, a análise da composição dos clusters revelou uma predominância parcial de cada explicador em grupos distintos.

Na Figura 4, observa-se que, para os casos em que a classificação foi correta (acerto = SIM), o cluster 0 contém 1.477 instâncias associadas ao *GradCAM*, contra 397 instâncias associadas ao *DeepLift*. Já no cluster 1 ocorre o padrão inverso, com predominância do *DeepLift*. Esse comportamento pode ser melhor visualizado por meio das imagens ilustrativas: no exemplo (A1), uma mesma imagem aparece no cluster 1 com o explicador *DeepLift*, enquanto em (A2), a mesma imagem aparece no cluster 0 com *GradCAM*. Já em (B), a mesma imagem está presente no mesmo cluster para ambos os explicadores.

Para se fazer uma análise do comportamento das métricas, as 28 métricas foram agrupadas em 5 grupos, conforme descrição da Tabela 1. As características gerais de cada cluster podem ser observadas na Figura 5. Nesta figura, as métricas de cada cluster foram agrupadas por grupo (Tabela 1) usando os valores médios de todas as métricas deste grupos, apresentadas em um gráfico de radar. É possível perceber nesta figura uma diferença

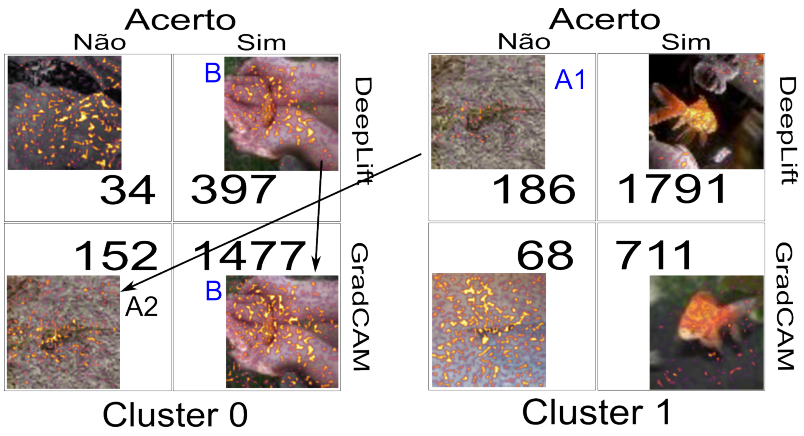

Figura 4. Composição dos clusters: distribuição dos explicadores (A1, A2, B)

marcante no grupo de métricas de Localização: cluster 0 apresenta valor médio de 0,294, enquanto o cluster 1 alcança 0,567. Os demais grupos de métricas exibem variações menores, visivelmente menos perceptíveis no gráfico. Esses resultados estão alinhados com a formulação proposta por [Hedström et al. 2023]. Este pode ser um indicativo que as métricas de localização estão sendo mais determinantes para a separação entre os grupos, quanto comparadas aos outros grupos de métricas.

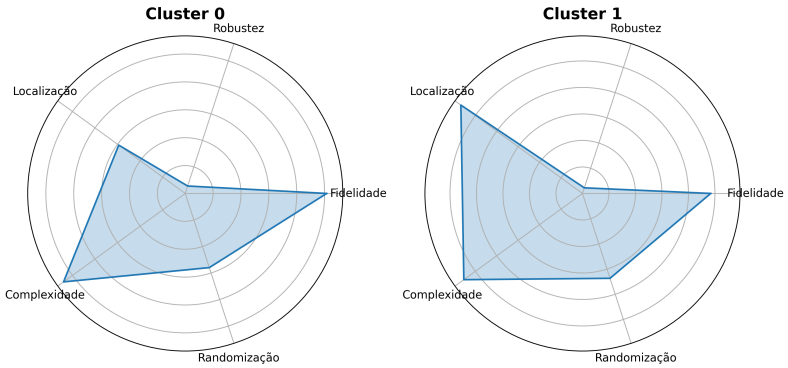

Figura 5. Análise dos clusters: médias por grupo de métricas

Considerando a média geral das métricas por tipo, o cluster 1 apresentou desempenho padronizado de 0,392 — um valor 19% superior ao do cluster 0 (0,329). Nesse grupo, o explicador *DeepLift* apareceu quase 2,5 vezes mais que o *GradCAM*, o que sugere um desempenho superior desse explicador para o modelo e conjunto de dados utilizados.

Por fim, a fim de identificar quais métricas mais influenciaram a separação entre os grupos, foram comparadas as médias das métricas entre os dois grupos. Observou-se que as maiores variações estavam concentradas nas métricas pertencentes à categoria Localização. As quatro métricas que apresentaram as maiores diferenças foram, em ordem decrescente: Pointing Game, Attribution Localisation, Top-K Intersection e Relevance Rank Accuracy. Esses resultados indicam que os explicadores pertencentes ao cluster 1 demonstram maior assertividade em identificar a região da imagem que se espera conter a explicação. Estes resultados apenas corroboram com os resultados obtidos pelo gráfico de radar da Figura 5.

6. Considerações Finais

Este trabalho propôs e implementou um pipeline experimental para avaliação sistemática de métricas de explicabilidade, com suporte à inclusão de diferentes explicadores, modelos de redes neurais e conjuntos de dados. A abordagem adotada permitiu agrupar explicadores com base em métricas quantitativas, revelando padrões de semelhança e distinção a partir da distribuição nos clusters.

Os resultados indicaram que, para o modelo Efficientnet-b1 e base ImageNet-Tiny, mesmo sem um vínculo direto entre explicador e agrupamento, formaram-se clusters com predominância distinta entre os explicadores analisados. Essa estrutura sugere que, para determinados conjuntos de dados e modelos, certos explicadores tendem a oferecer melhor desempenho sob critérios objetivos, como as métricas de localização. Além disso, a análise evidenciou que o explicador *DeepLift* apresentou desempenho superior ao *Grad-CAM* no cenário avaliado, com predominância nos grupos de métricas de maiores valores atribuídos às instâncias avaliadas. Tal resultado reforça a eficácia do DeepLift na geração de explicações mais alinhadas às regiões relevantes da imagem, conforme os critérios considerados.

Como trabalho futuro, propõe-se a expansão do estudo com a inclusão de novos explicadores, arquiteturas de modelos (como Vision Transformers) e bases de dados de diferentes domínios, como medicina e clima, também uma investigação sobre os valores baixos de robustez. Também se vislumbra a possibilidade de desenvolver explicadores híbridos ou estratégias de Ensemble, combinando múltiplos explicadores para potencializar a robustez e a qualidade das explicações geradas, adaptando-se melhor às particularidades de cada tarefa ou domínio.

Referências

Aggarwal, C. and Reddy, C. (2018). *Data Clustering: Algorithms and Applications*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. CRC Press.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., and Herrera, F. (2023). Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99:101805.

Arras, L., Osman, A., and Samek, W. (2021). Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81.

Arrieta, A. B., Díaz-Rodríguez, N., Ser, J. D., and et al (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Arya, V., Bellamy, R. K. E., Chen, P.-Y., and et al (2019). One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques.

Bommer, P. L., Kretschmer, M., Hedström, A., Bareeva, D., and Höhne, M. M.-C. (2024). 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):e230074.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

Hedström, A., Weber, L., Krakowczyk, D., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., and Höhne, M. M. M. (2023). Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows.

Luo, Y., Wong, Y., Kankanhalli, M., and Zhao, Q. (2021). Direction concentration learning: Enhancing congruency in machine learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):1928–1946.

Mersha, M., Lam, K., Wood, J., AlShami, A. K., and Kalita, J. (2024). Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *Neurocomputing*, 599:128111.

Miró-Nicolau, M., i Capó, A. J., and Moyà-Alcover, G. (2023). Assessing fidelity in xai post-hoc techniques: A comparative study with ground truth explanations datasets.

O'Shea, K. and Nash, R. (2015). An introduction to convolutional neural networks.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Russell, S. and Norvig, P. (2009). *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, USA, 3rd edition.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2):336–359. arXiv:1610.02391 [cs].

Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. (2017). Not just a black box: Learning important features through propagating activation differences.

Tan, M. and Le, Q. V. (2020). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv:1905.11946 [cs].

Witten, I., Frank, E., Hall, M., Pal, C., and Foulds, J. (2025). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Zhang, A., Lipton, Z. C., Li, M., and Smola, A. J. (2023). *Dive into Deep Learning*. Cambridge University Press. `https://D2L.ai`.