Explainable AI (XAI) Local Aplicada à Classificação de Imagens

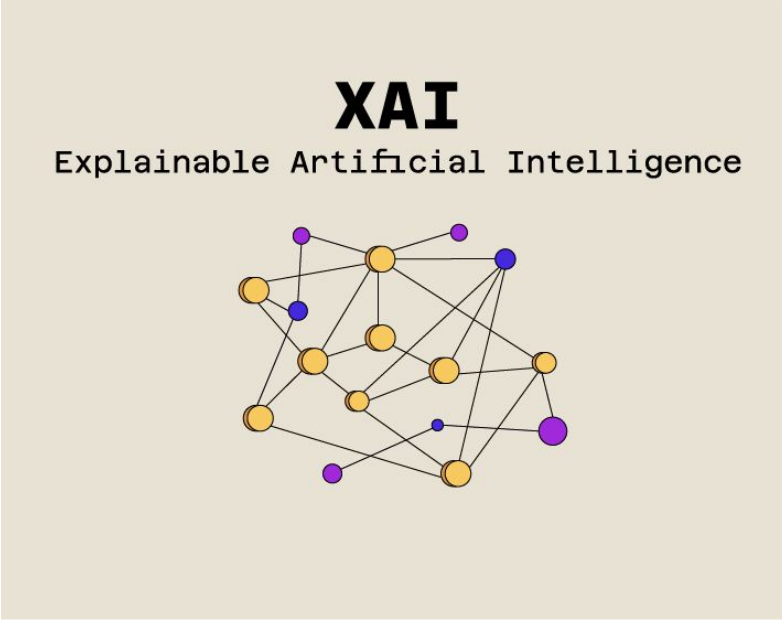
Bruna Zamith Santos
Igor Perez Cunha
Wanusa Pontes

Centro de Informática UFPE
Universidade Federal de Pernambuco

Introdução

1. Introdução

- **Objetivos do Estudo**
  - Aplicar métodos de XAI local para interpretar as previsões de um modelo de classificação de imagens.
  - Avaliar e comparar a relevância de diferentes regiões da imagem na classificação para diferentes datasets.



1. Introdução

- **Por que usar XAI Local?**
  - Explicações Focadas: Oferece interpretações específicas para cada previsão, permitindo identificar o que realmente influenciou a decisão do modelo.
  - Análise de Casos Específicos: Esclarece previsões erradas ou inesperadas, apontando possíveis viéses ou limitações do modelo.
  - Foco na Transparência: Permite o entendimento detalhado de como o modelo enxerga diferentes tipos de dados.

- **Por que XAI Local em Classificação de Imagens?**
  - Modelos Complexos: Redes Neurais Convolucionais (CNNs) frequentemente operam como "caixas-pretas", dificultando a interpretação de suas decisões.
  - Confiabilidade e Responsabilidade: Explicações tornam os modelos mais confiáveis, especialmente em domínios sensíveis, como saúde e segurança.
  - Melhoria de Modelos: Identificar características mal interpretadas permite ajustes nos dados e no treinamento.

1. Introdução

- **Motivação: Validação da Qualidade de Modelos**
  - Paper: *"Why Should I Trust You?": Explaining the Predictions of Any Classifier. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16).*

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

1. Introdução

- **Motivação: Confiabilidade em Setores Críticos**
  - Paper: *"Explainable AI in Medical Imaging: An Overview for Clinical Practitioners – Saliency-Based XAI Approaches." K. Borys, et al. 2023. In European Journal of Radiology.*

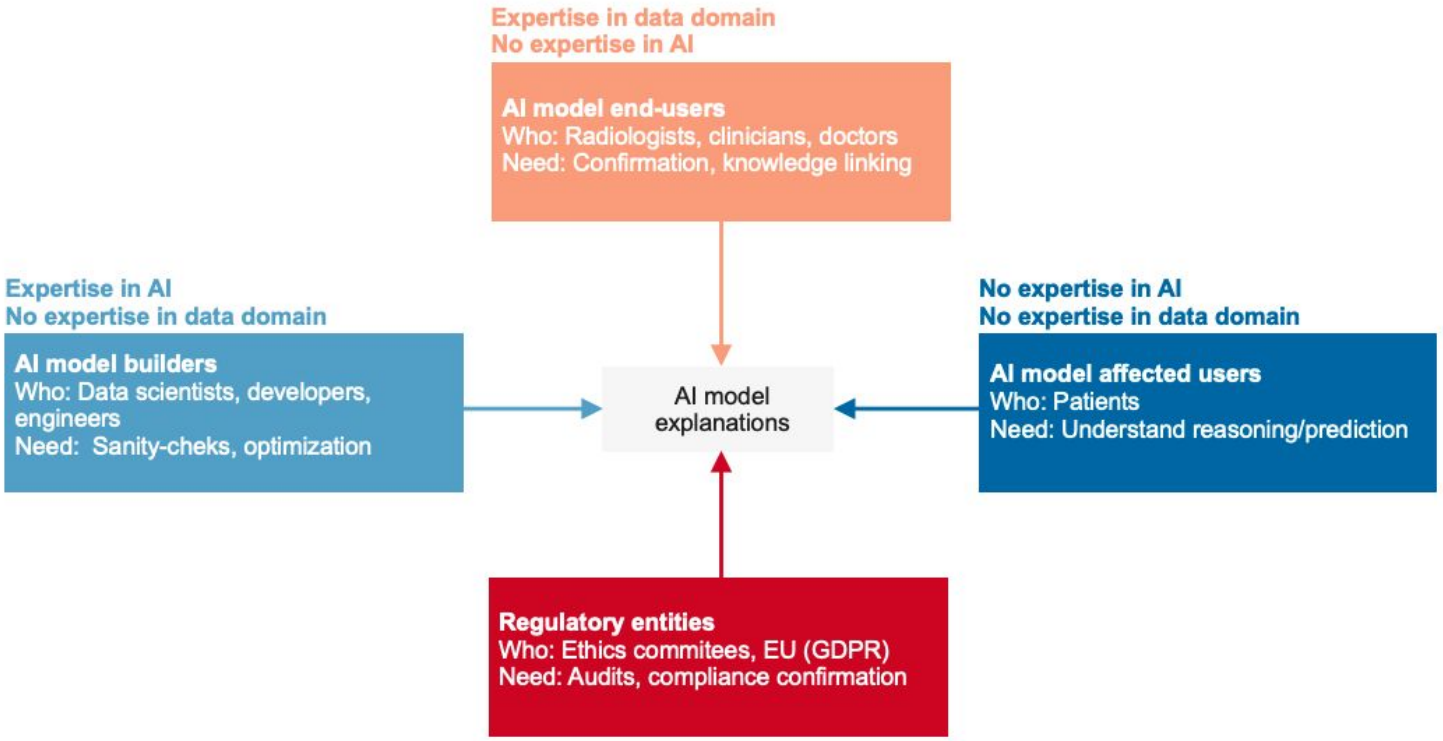
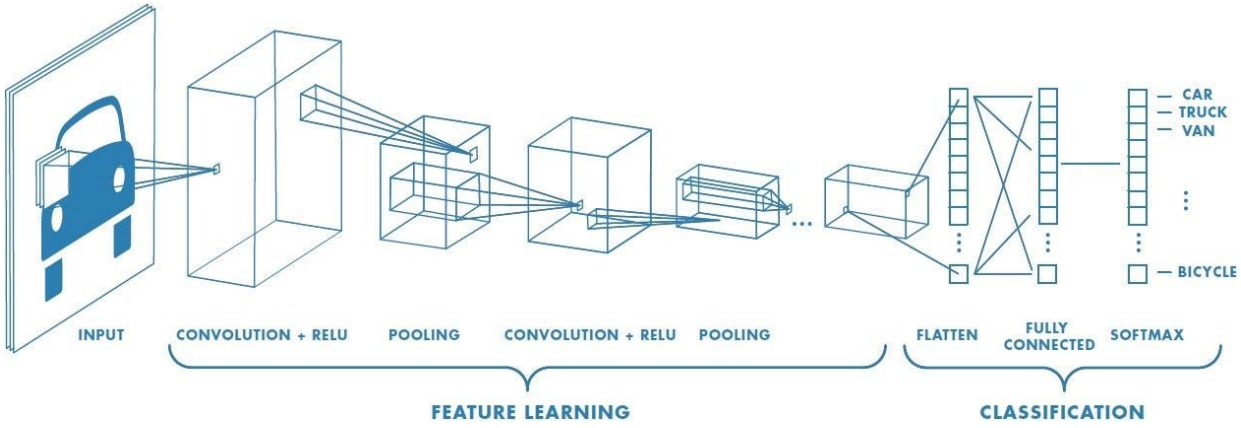


Figure showing different XAI stakeholder groups for AI models deployed in the medical domain, which depending on their domain knowledge and expertise, seek different explanation needs. The overview is partially inspired by [8].

1. Introdução

- **Redes Neurais Convolucionais (CNNs)**
  - Modelos de *deep learning* projetados para processar dados estruturados em grades, como imagens.
  - Compõem-se de camadas convolucionais, que aprendem a extrair características visuais relevantes, como bordas, texturas e padrões.
  - Amplamente utilizadas em tarefas de classificação, segmentação e detecção de objetos.
  - Apesar do excelente desempenho, sofrem do problema de caixa-preta, dificultando a interpretação de suas decisões.



https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

1. Introdução

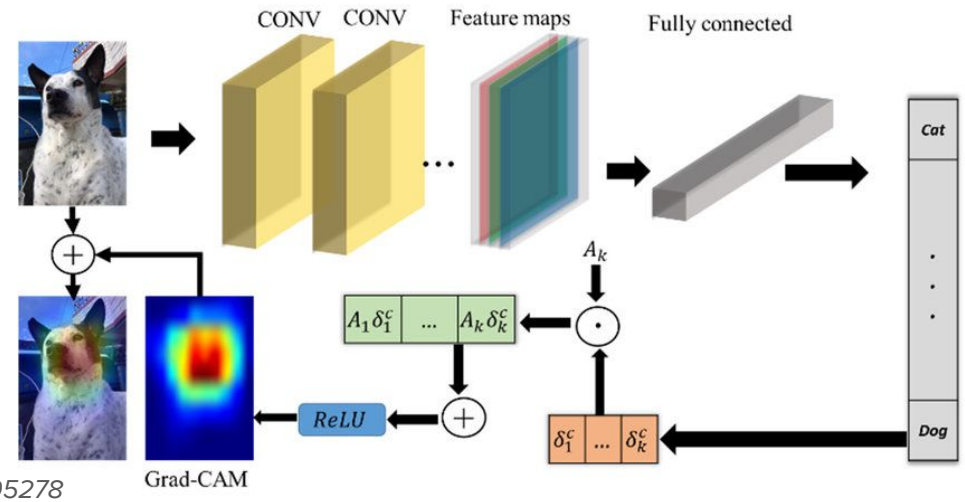
- **Local Interpretable Model-agnostic Explanations (LIME)**
  - Método agnóstico ao modelo que gera explicações locais para previsões individuais.
  - Funciona ajustando um modelo simples (e.g., regressão linear) para aproximar o comportamento do modelo complexo em torno de uma amostra específica.
  - No caso de imagens, divide-as em superpixels e identifica quais regiões têm maior peso na classificação.
  - Vantagens: Rápido, intuitivo e visualmente fácil de interpretar.
  - Limitações: Sensível à segmentação e pode gerar explicações instáveis em diferentes execuções.

1. Introdução

- **SHapley AdditiveexPlanations (SHAP)**
  - Baseado na teoria dos valores de Shapley, provenientes da teoria dos jogos cooperativos.
  - Atribui uma importância a cada característica de forma consistente e equitativa, considerando a contribuição marginal de cada uma.
  - Em imagens, avalia o impacto de cada pixel (ou grupo de pixels) para a previsão do modelo.
  - Vantagens: Explicações consistentes e matematicamente sólidas.
  - Limitações: Computacionalmente caro, especialmente em imagens de alta resolução.

1. Introdução

- **Gradient-weighted Class Activation Mapping (Grad-CAM)**
  - Técnica específica para CNNs que utiliza gradientes para gerar mapas de ativação.
  - Indica quais regiões da imagem ativaram fortemente os neurônios das camadas finais, associadas à decisão do modelo.
  - Permite uma visualização direta e intuitiva das áreas relevantes para cada classe.
  - Vantagens: Aproveita informações da estrutura da CNN e é computacionalmente eficiente.
  - Limitações: Depende da arquitetura da rede e só funciona para modelos baseados em gradientes.



https://www.researchgate.net/figure/Grad-CAM-architecture_fig3_352795278

Experimentos

2. Experimentos

- **CNNs**

"Simples"

- Arquitetura: 2 blocos convolucionais seguidos por 3 camadas fully connected.
- Componentes:
  - Max Pooling: Reduz a dimensionalidade.
  - Batch Normalization: Garante estabilidade no treinamento.
  - ReLU: Introduz não-linearidade.
- Otimizador: SGD com taxa de aprendizado alta.
  - Vantagens: Convergência inicial rápida, melhor exploração do espaço de soluções.
  - Limitações: Possíveis oscilações ou dificuldade nos ajustes finais.

"Robusta"

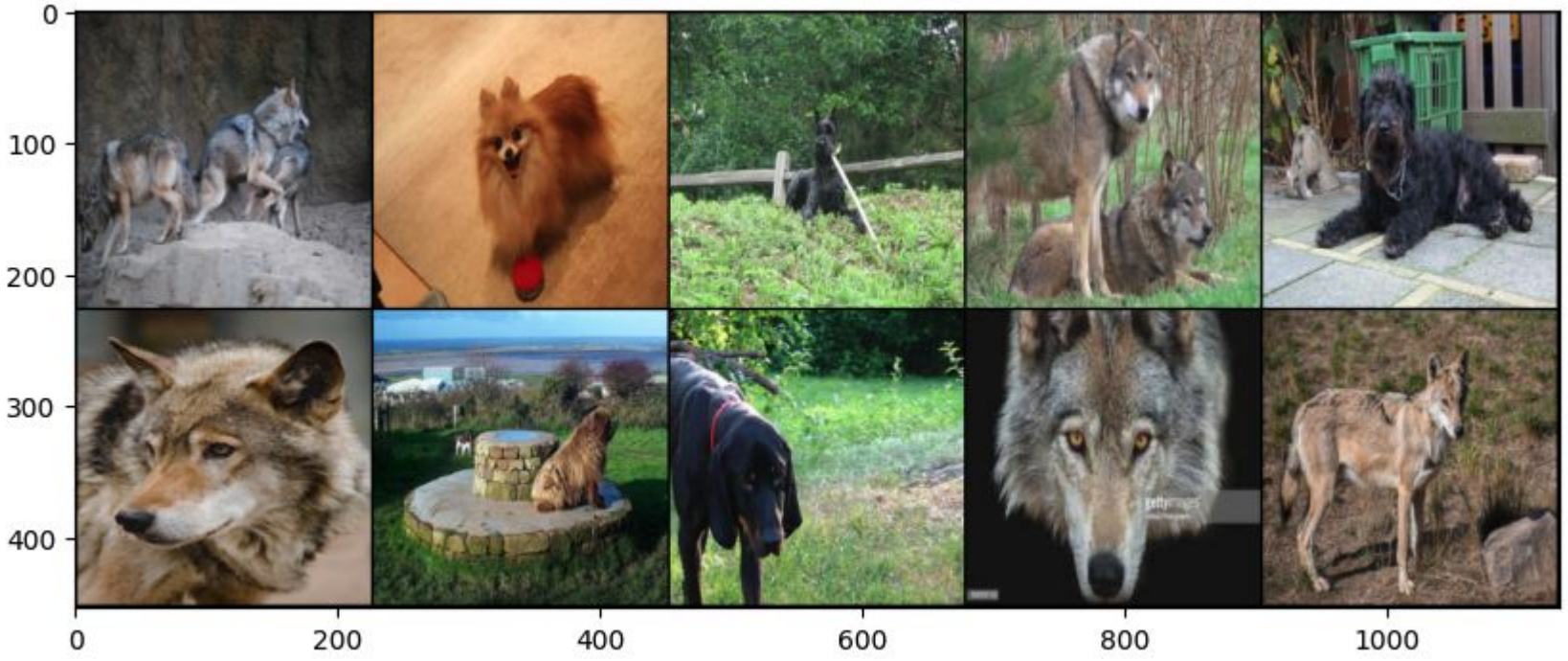
- Arquitetura: 3 blocos convolucionais seguidos por 3 camadas fully connected.
- Componentes:
  - Max Pooling, Batch Normalization, ReLU, ...
  - Dropout: Regularização nas camadas fully connected para reduzir overfitting.
- Otimizador: AdamW com taxa de aprendizado moderada.
  - Vantagens: Convergência estável e eficiente, insensível a ajustes manuais de hiperparâmetros.
  - Limitações: Pode ser menos eficiente em exploração global comparado ao SGD.

Conjunto de Dados 1:

Cachorros vs Lobos

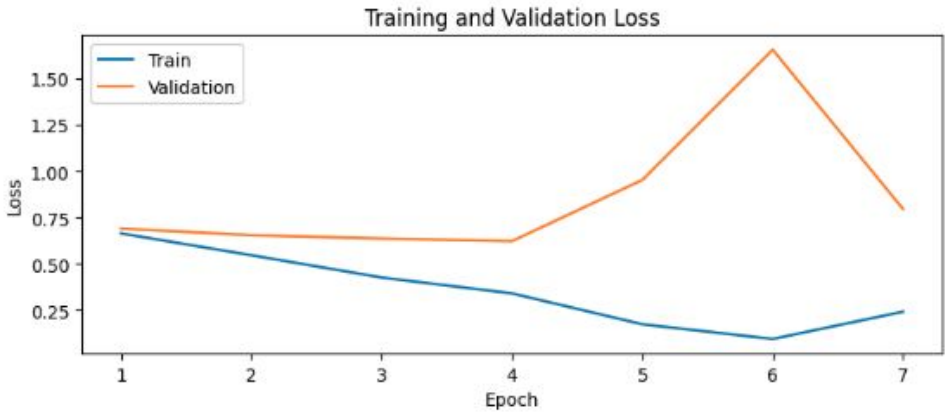
2. Experimentos

- 2000 imagens de cachorros e lobos, coletadas de diversas fontes, incluindo Stanford Dogs Dataset, Imagenet e Google Images [1].
- Classificação binária (2 classes)



2. Experimentos

- **Primeiros testes: CNN "simples"**

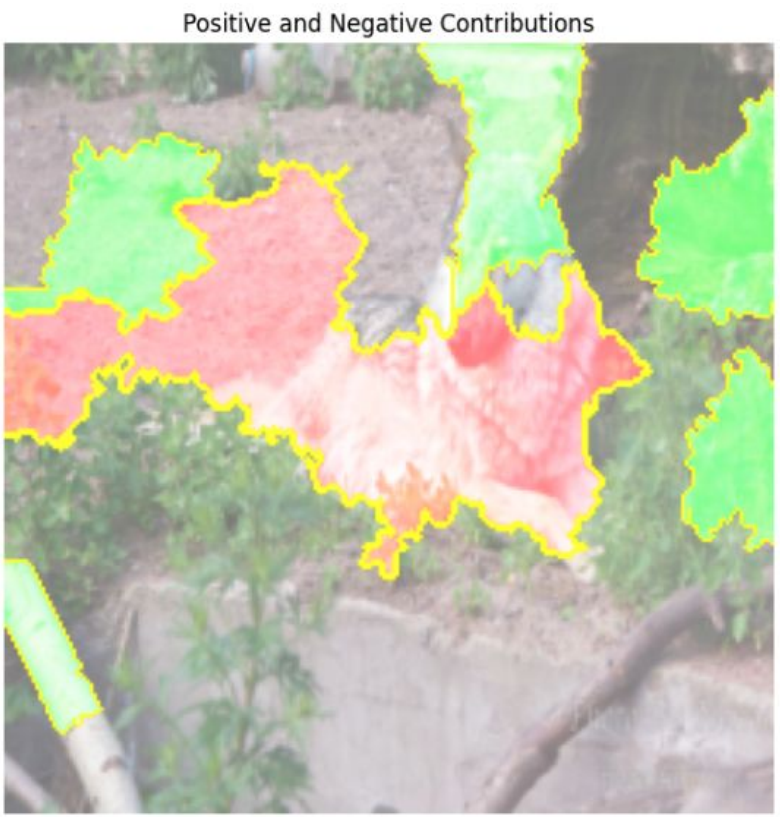


	precision	recall	f1-score	support
dogs	0.57	0.81	0.67	154
wolves	0.72	0.44	0.54	166
accuracy			0.62	320
macro avg	0.64	0.63	0.61	320
weighted avg	0.65	0.62	0.61	320

- Cachorros: Boa detecção com 81% de recall, mas precisão moderada (57%). Muitos falsos positivos.
- Lobos: Alta precisão (72%), mas recall baixo (44%). Muitos falsos negativos.
- Acurácia Geral: Modelo acerta 62% das classificações totais.

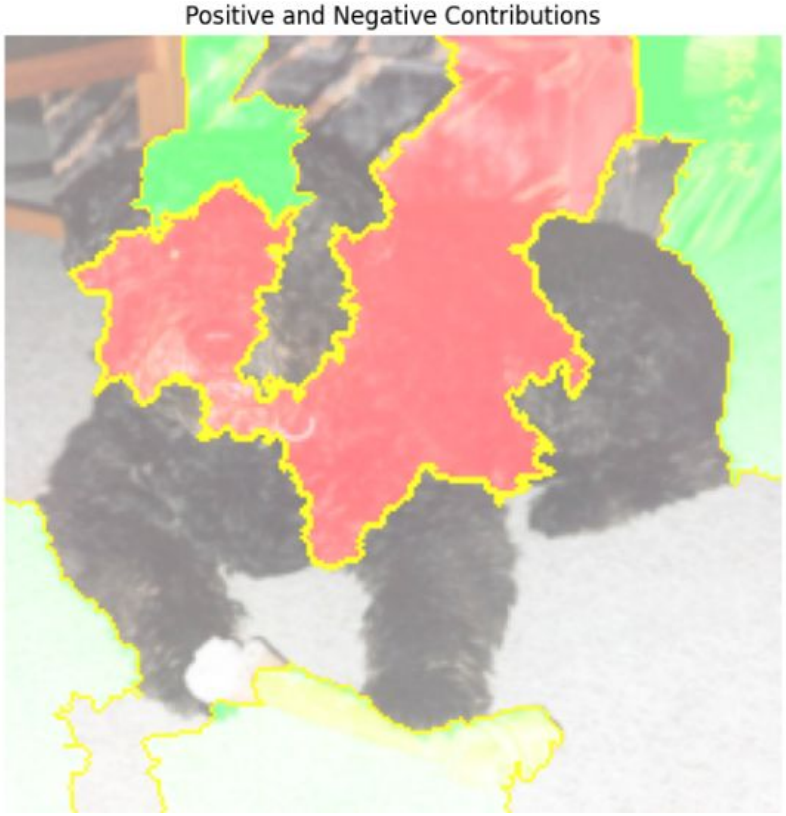
2. Experimentos

- **Primeiros testes: CNN "simples"**
  - LIME



2. Experimentos

- **Primeiros testes: CNN "simples"**
  - LIME

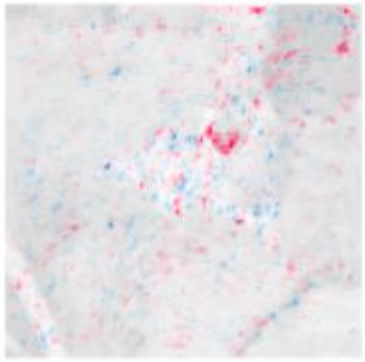


2. Experimentos

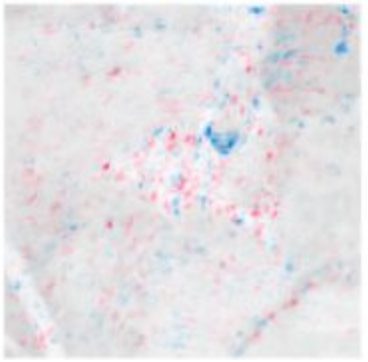
- **Primeiros testes: CNN "simples"**
  - SHAP

wolves

dogs

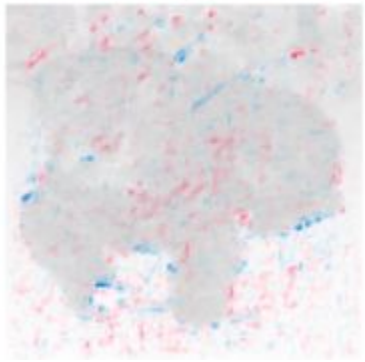


wolves

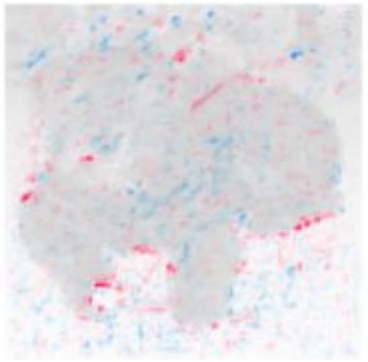


dogs

dogs



wolves

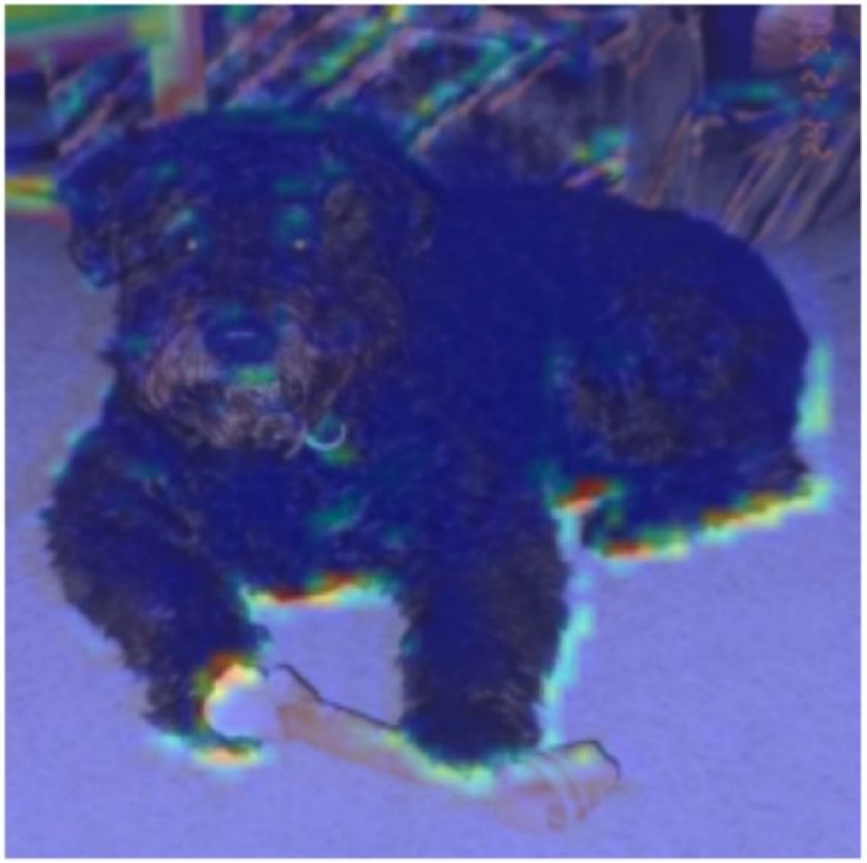


2. Experimentos

- **Primeiros testes: CNN "simples"**
  - Grad-CAM

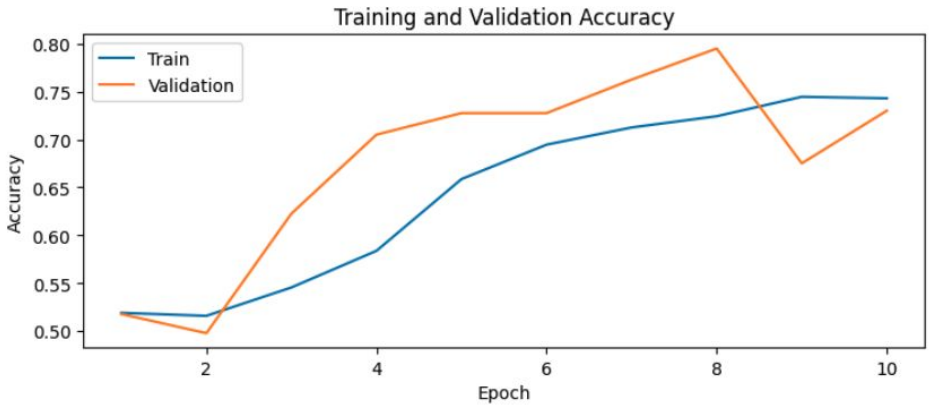
Label: wolves, Predicted: dogs

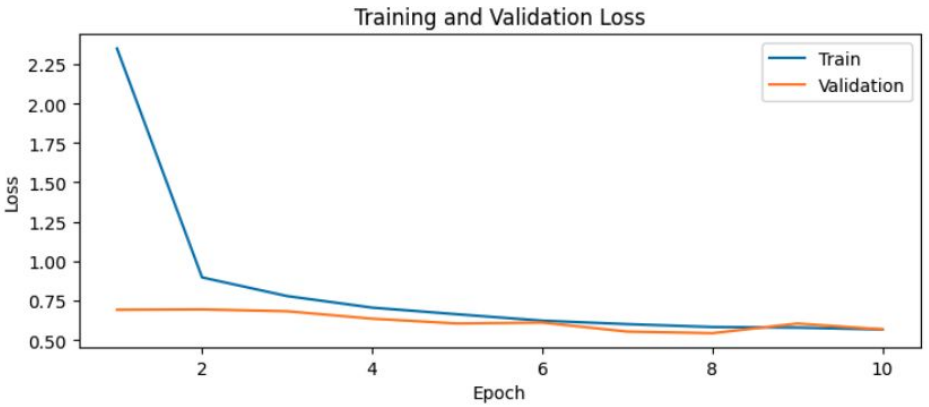
Label: dogs, Predicted: dogs



2. Experimentos

- **Testes finais: CNN "robusta"**



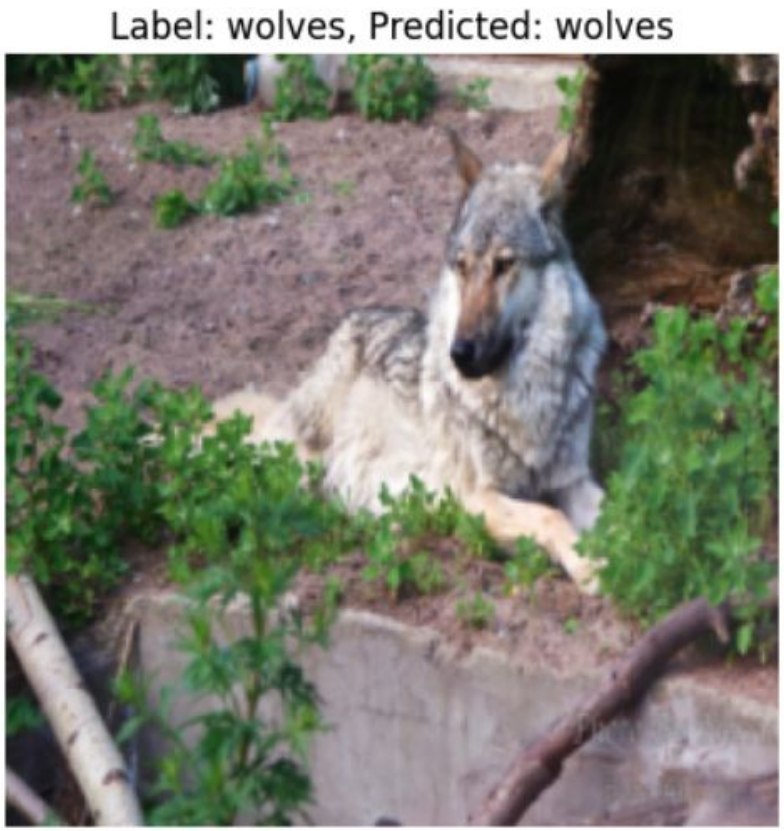


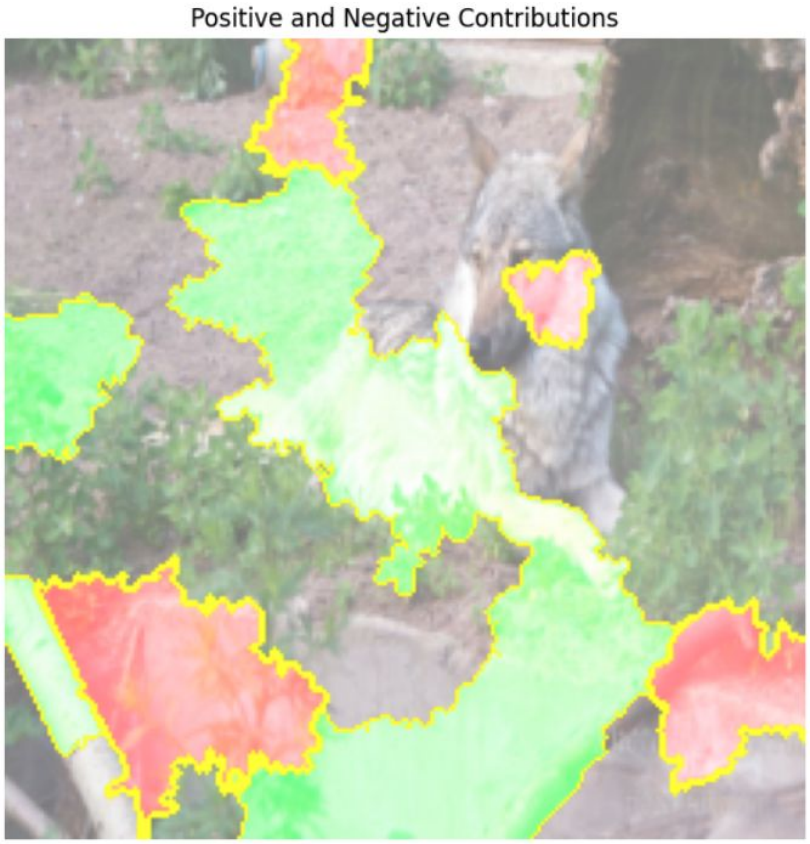
	precision	recall	f1-score	support
dogs	0.74	0.81	0.78	154
wolves	0.81	0.74	0.77	166
accuracy			0.78	320
macro avg	0.78	0.78	0.77	320
weighted avg	0.78	0.78	0.77	320

- Cachorros: Boa detecção com 81% de recall e precisão razoável (74%). Alguns falsos positivos.
- Lobos: Alta precisão (81%) e recall bom (74%). Alguns falsos negativos.
- Acurácia Geral: Modelo acerta 78% das classificações totais.

2. Experimentos

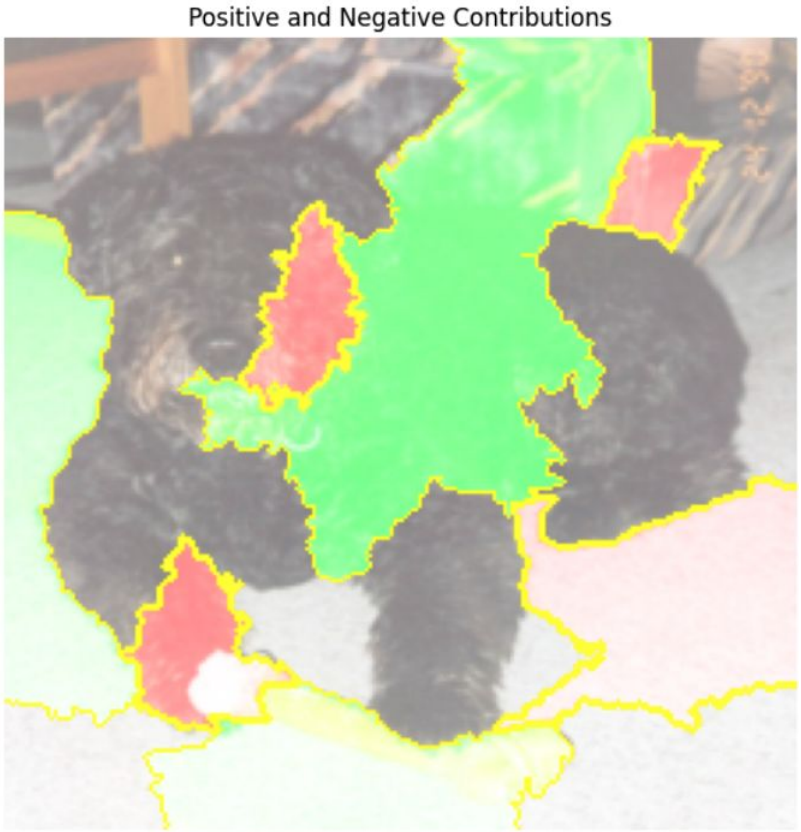
- **Testes finais: CNN "robusta"**
  - LIME





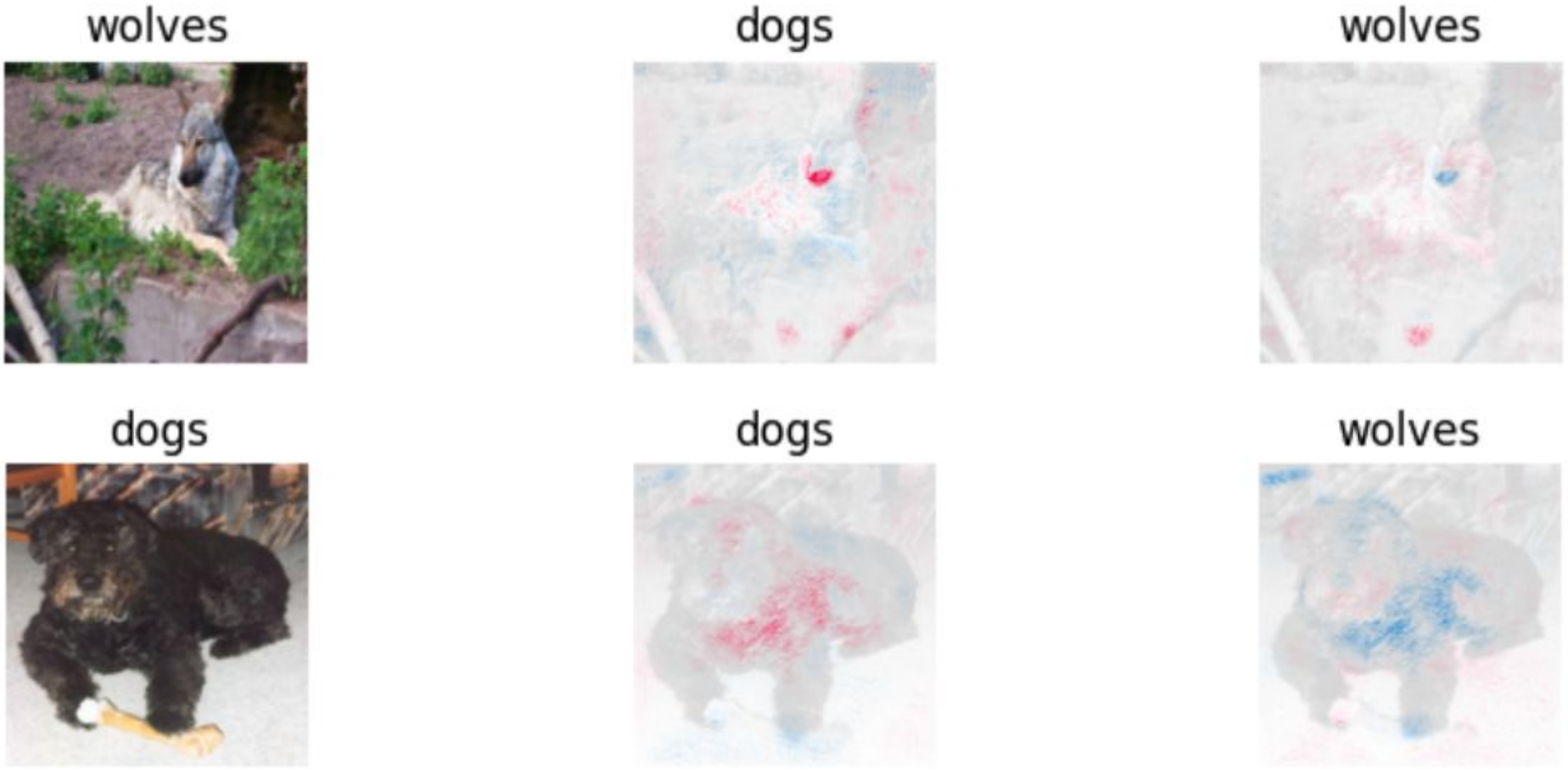
2. Experimentos

- **Testes finais: CNN "robusta"**
  - LIME



2. Experimentos

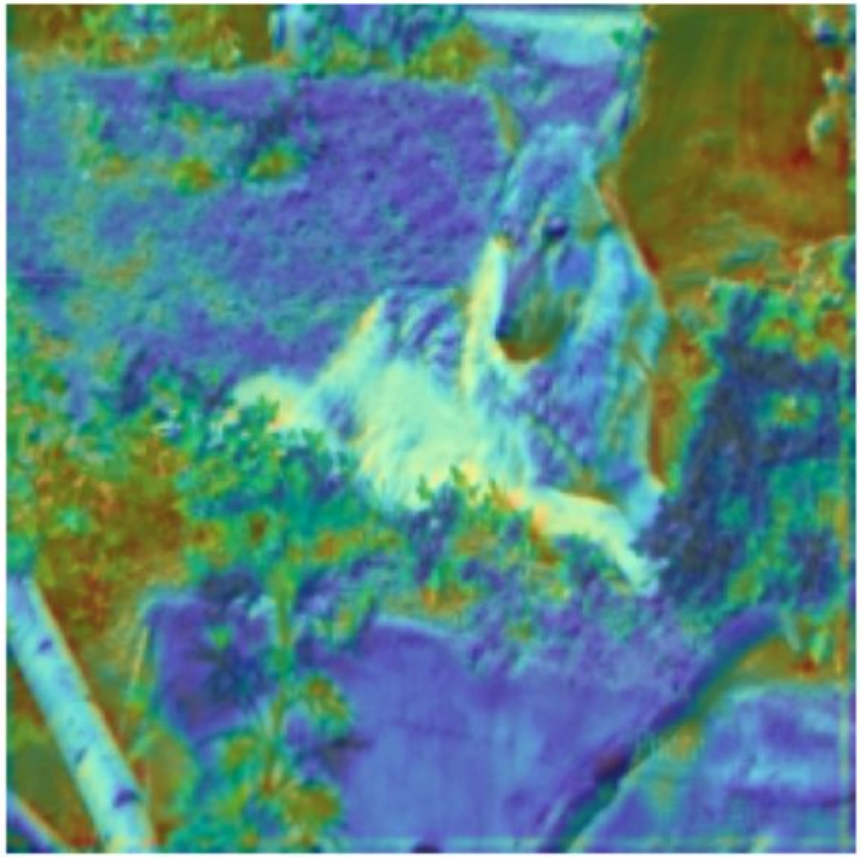
- **Testes finais: CNN "robusta"**
  - SHAP



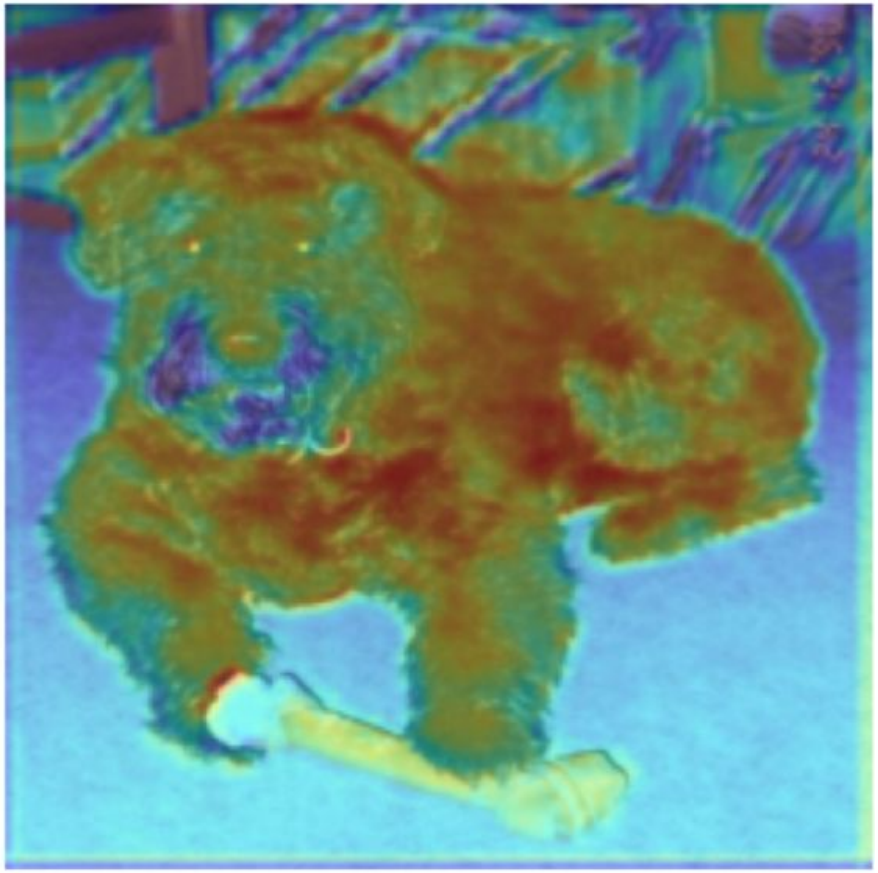
2. Experimentos

- **Testes finais: CNN "robusta"**
  - Grad-CAM

Label: wolves, Predicted: wolves



Label: dogs, Predicted: dogs



2. Experimentos

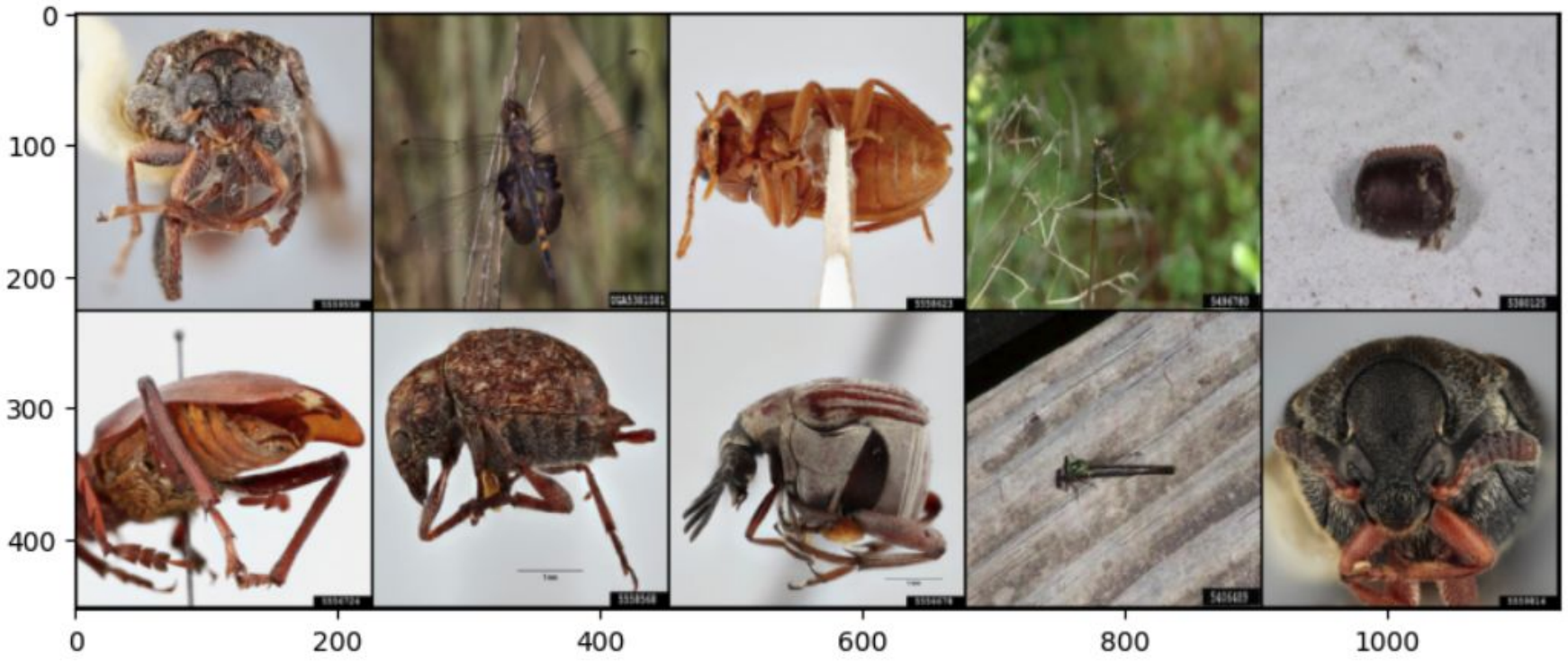
- A CNN "simples" apresentou uma forte tendência a atribuir importância a pixels de elementos como "ossos", "humanos" e "coleiras" ao classificar corretamente a classe "cachorro".
    - Com a migração para a CNN "robusta", esses problemas foram reduzidos, embora alguns novos casos tenham surgido, como a identificação de "grades" de canis recebendo mais relevância que o próprio cachorro.
- O LIME demonstrou uma visualização de fácil interpretação, mas revelou-se bastante instável entre diferentes execuções. Já o SHAP, embora menos intuitivo, destacou diferenças significativas na concentração de pixels relevantes ao migrarmos para a CNN "robusta".
- Por fim, o Grad-CAM apresentou comportamentos bem distintos entre as duas arquiteturas: na CNN "simples", houve foco excessivo em bordas e pequenos pontos; na "robusta", a atenção foi direcionada para regiões maiores e mais relevantes ao contexto da classificação.

Conjunto de Dados 2:

Insetos

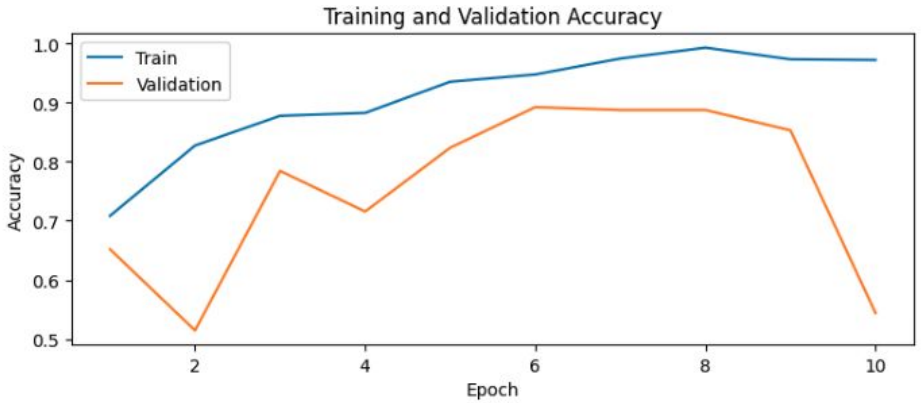
2. Experimentos

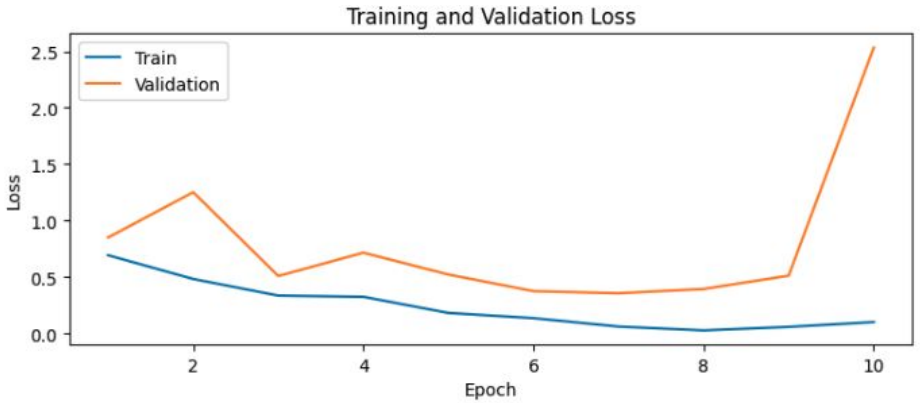
- 1200 imagens de insetos, coletadas do https://www.insectimages.org/.
- Classificação em 3 classes: Besouro, Barata e Libélula



2. Experimentos

- **Primeiros testes: CNN "simples"**



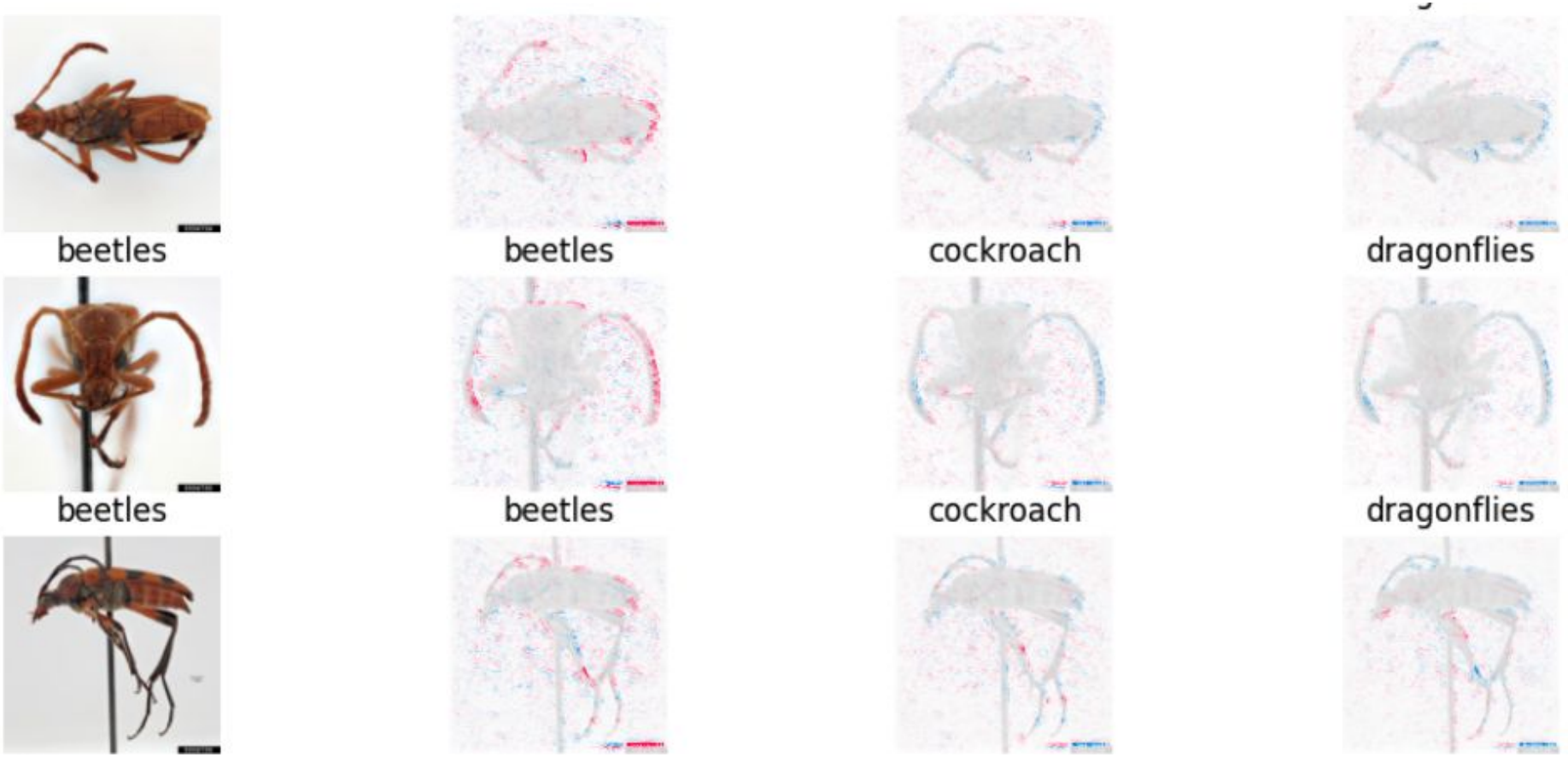


	precision	recall	f1-score	support
beetles	0.90	0.43	0.58	60
cockroach	0.55	1.00	0.71	60
dragonflies	0.95	0.65	0.77	60
accuracy			0.69	180
macro avg	0.80	0.69	0.69	180
weighted avg	0.80	0.69	0.69	180

- Besouros: Alta precisão (90%), mas recall baixo (43%), indicando muitos falsos negativos.
- Barata: Recall perfeito (100%), mas precisão moderada (55%), mostrando muitos falsos positivos.
- Libélulas: Excelente precisão (95%) e recall razoável (65%), com desempenho equilibrado.
- Acurácia Geral: Modelo acerta 69% das classificações totais.

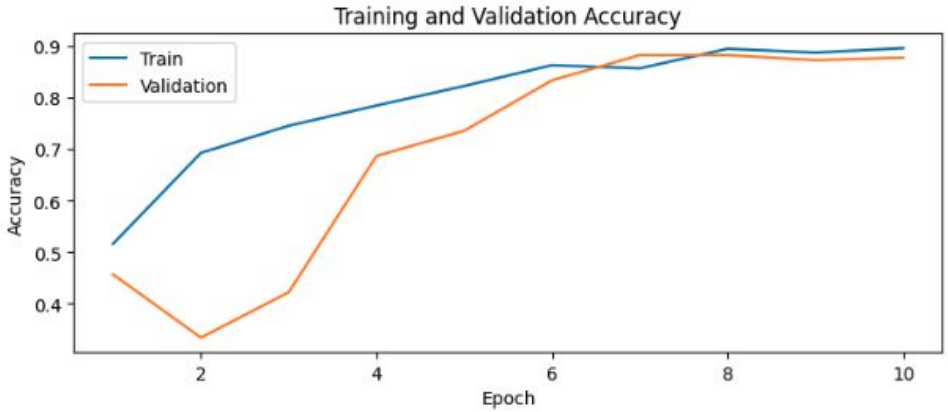
2. Experimentos

- **Primeiros testes: CNN "simples"**
  - SHAP: Dando importância para as *tags* nas imagens (canto inferior direito)
    Potencial *overfitting*



2. Experimentos

- **Testes finais: CNN "robusta" + remoção das *tags* das imagens**



	precision	recall	f1-score	support
beetles	1.00	0.67	0.80	60
cockroach	0.79	0.88	0.83	60
dragonflies	0.79	0.97	0.87	60
accuracy			0.84	180
macro avg	0.86	0.84	0.84	180
weighted avg	0.86	0.84	0.84	180

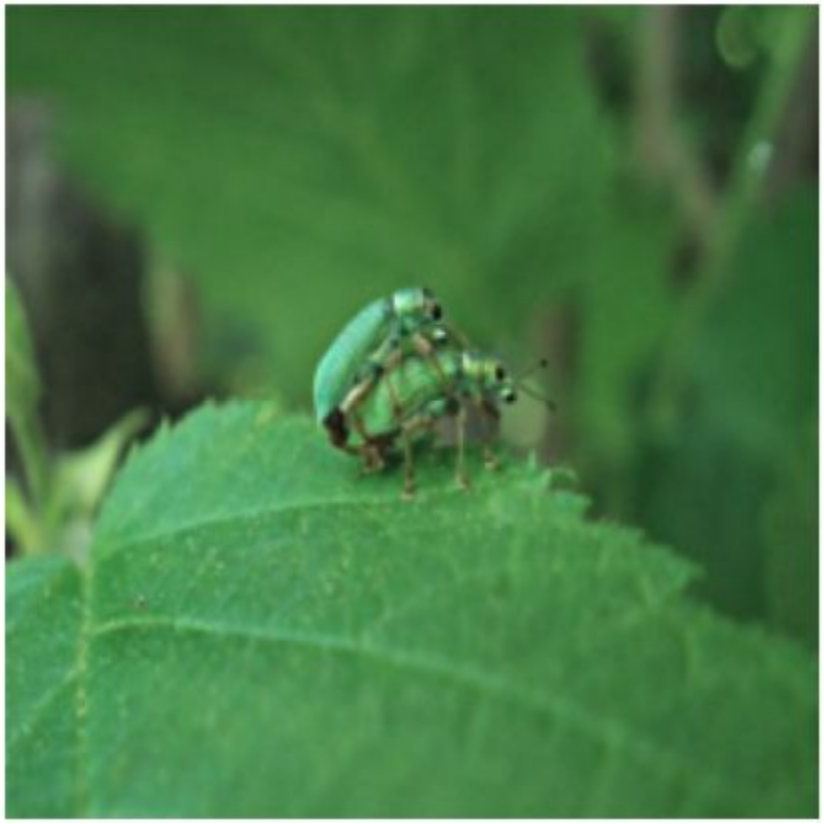
- Besouros: Precisão perfeita (100%), mas recall moderado (67%), indicando muitos falsos negativos.
- Barata: Boa precisão (79%) e recall alto (88%), com desempenho equilibrado.
- Libélulas: Precisão razoável (79%) e recall excelente (97%), com poucos falsos negativos.
- Acurácia Geral: Modelo acerta 84% das classificações totais.

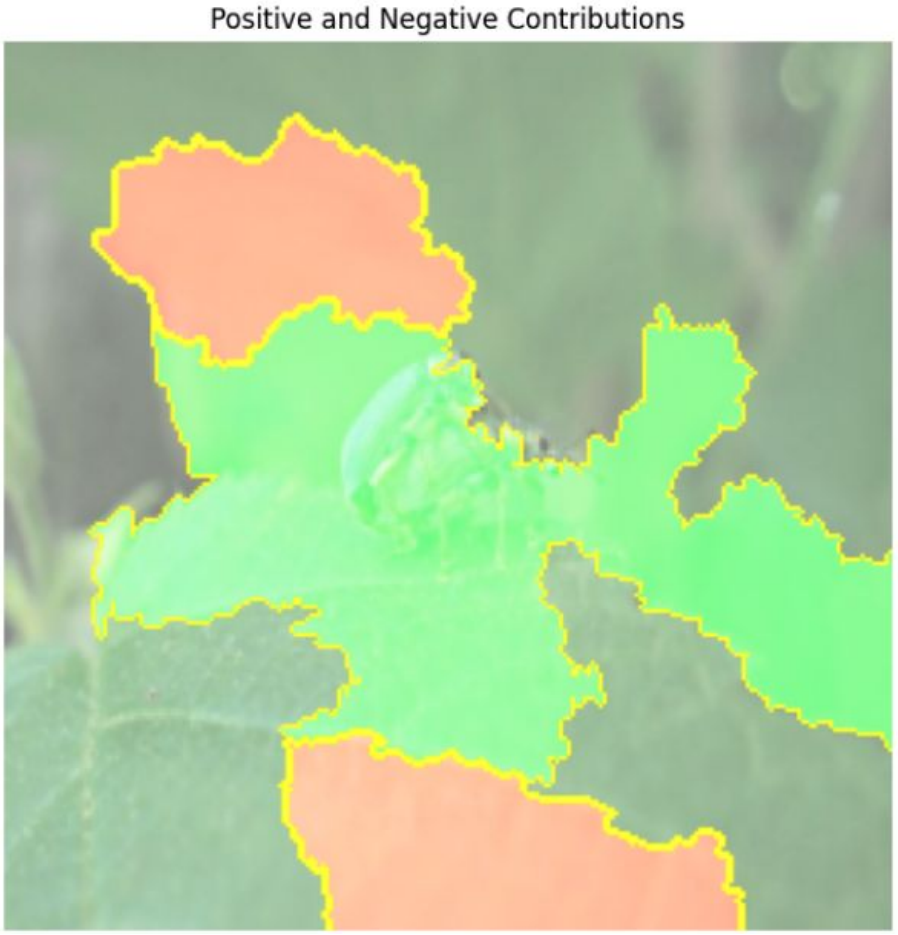
2. Experimentos

- **Testes finais: CNN "robusta" + remoção das *tags* das imagens**
  - LIME

2. Experimentos

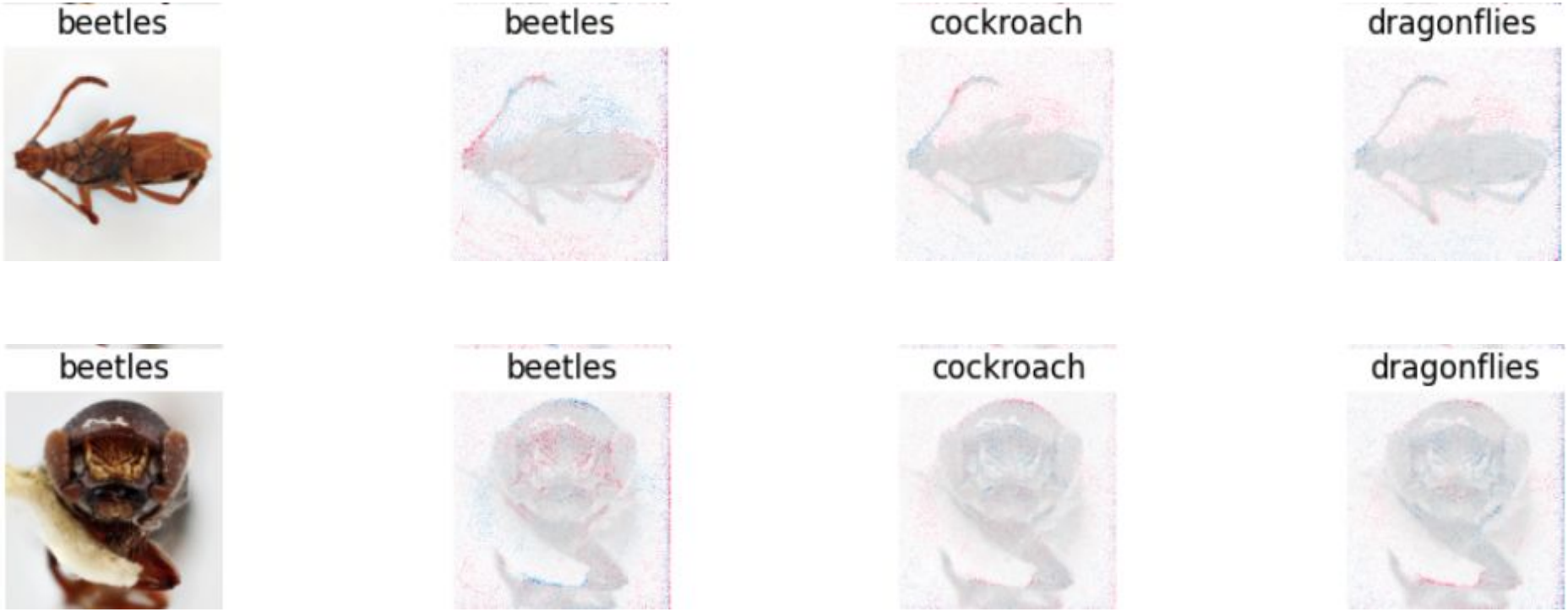
- **Testes finais: CNN "robusta" + remoção das *tags* das imagens**
  - LIME





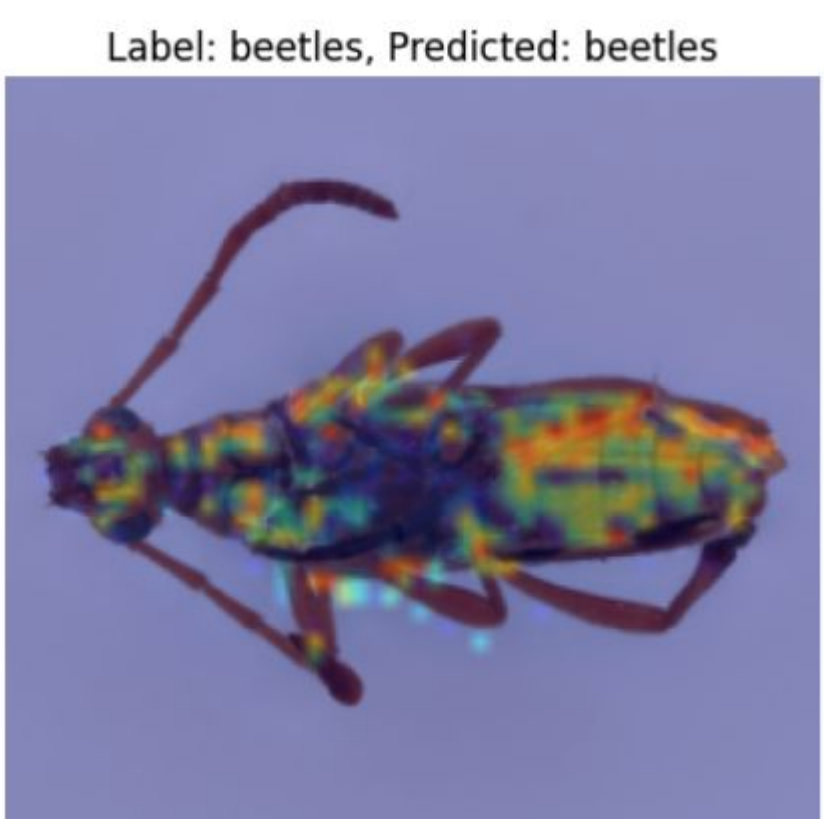
2. Experimentos

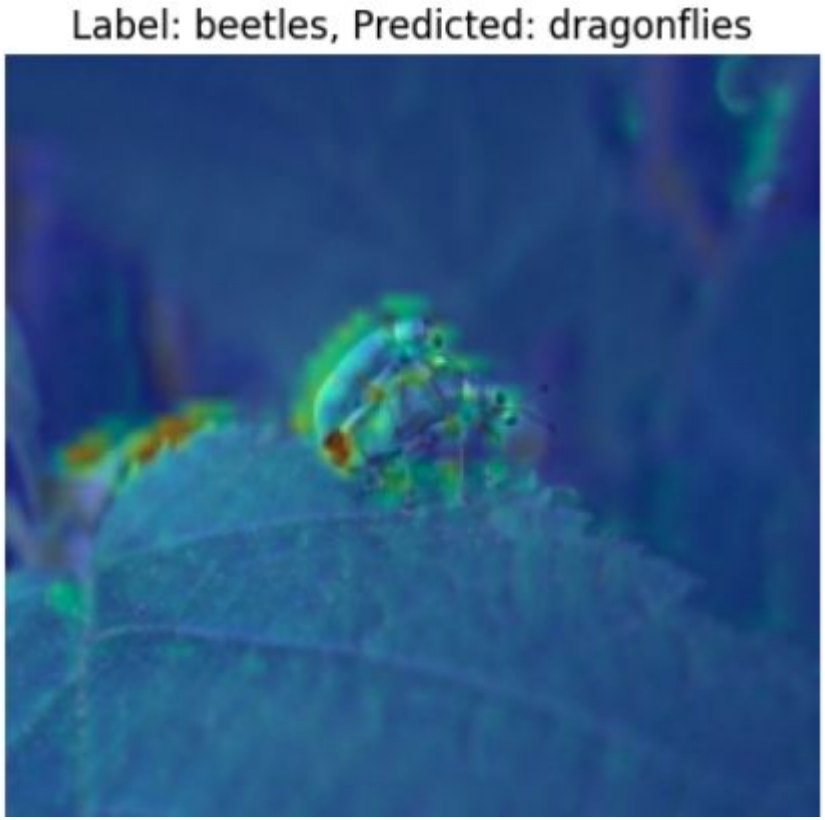
- **Testes finais: CNN "robusta" + remoção das *tags* das imagens**
  - SHAP



2. Experimentos

- **Testes finais: CNN "robusta" + remoção das *tags* das imagens**
  - Grad-CAM





2. Experimentos

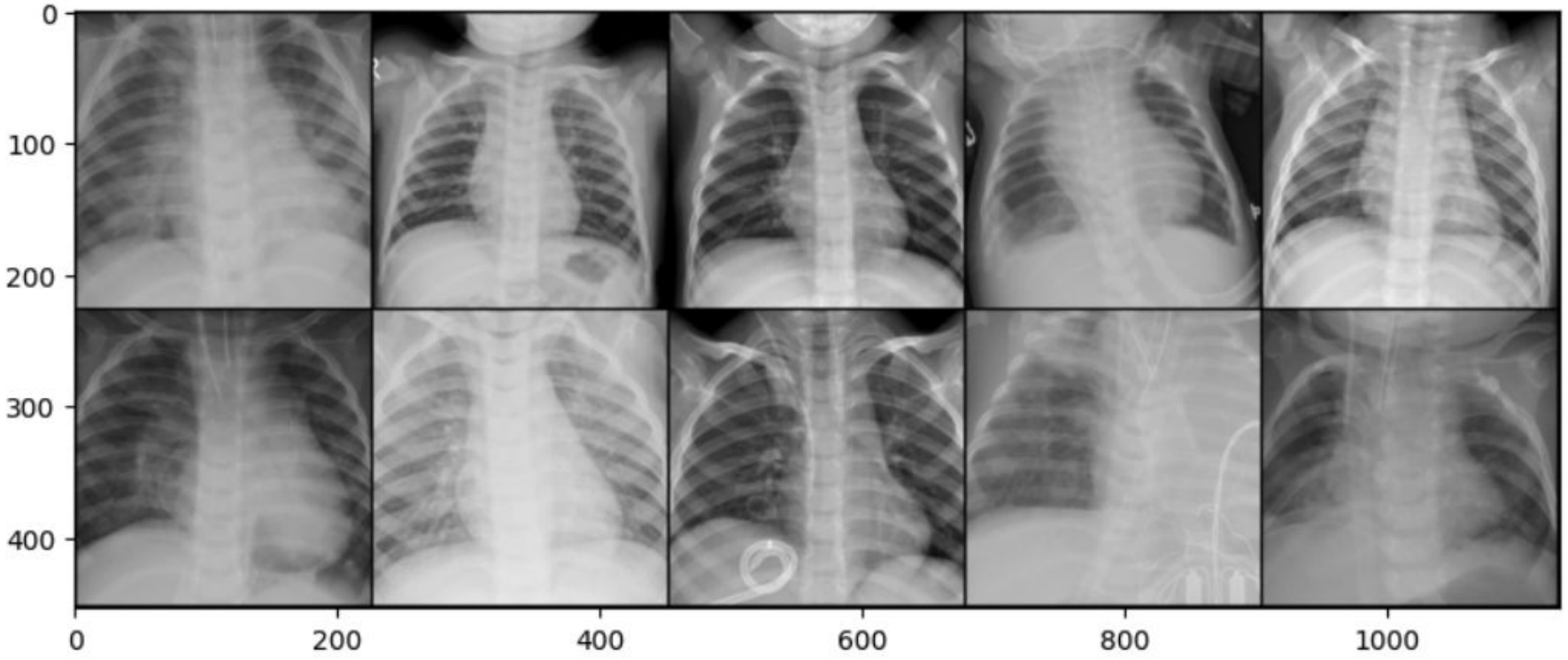
- O SHAP revelou um comportamento "estranho" do modelo, que atribuía grande importância às *tags* das imagens originais, sugerindo um possível *overfitting*.
- Após a remoção das *tags* e a aplicação da CNN "robusta", houve melhorias tanto nas métricas do modelo quanto nos resultados gerados pelos métodos de XAI.
- Ainda assim, o XAI destacou oportunidades de melhoria, como a tendência do modelo em hiper-valorizar imagens com folhas, frequentemente classificando-as como libélulas devido à associação aprendida durante o treino.
- O SHAP foi particularmente útil neste contexto multiclasse, pois permitiu comparar as contribuições de características específicas entre diferentes classes, enriquecendo a análise do comportamento do modelo.

Conjunto de Dados 3:

Raio-X Pulmonar

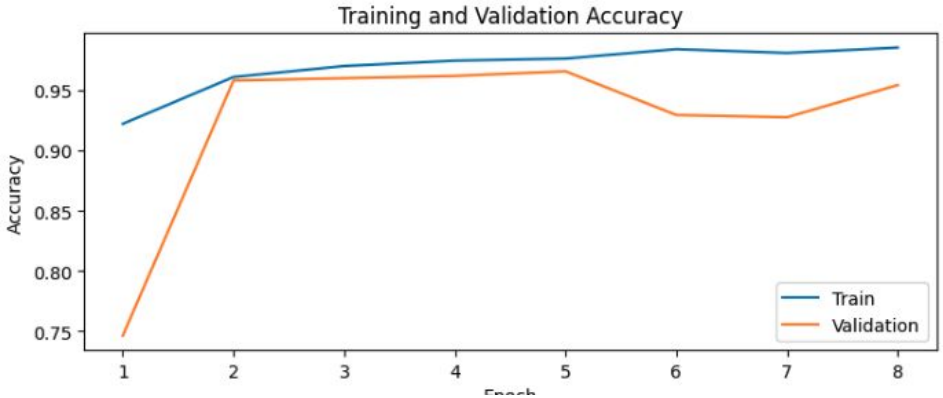
2. Experimentos

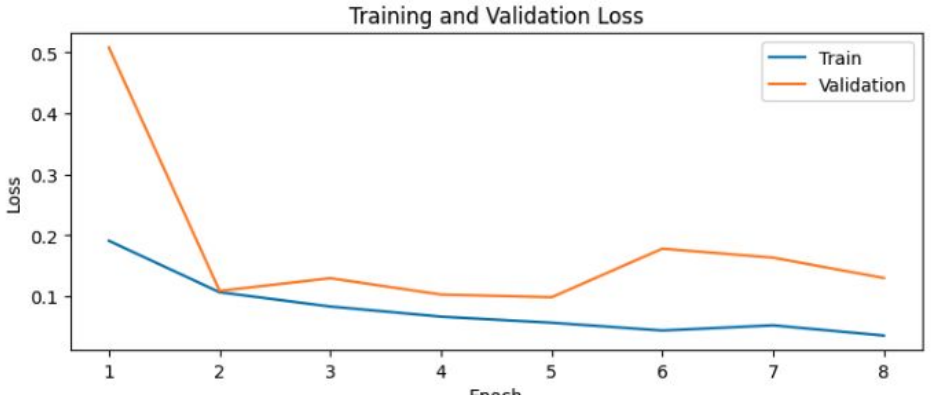
- 5856 imagens de raio-X pulmonar pediátrico, publicadas em *MedMNIST v2 - A Large-Scale Lightweight Benchmark for 2D and 3D Biomedical Image Classification." Junyu Yang, et al. 2023. In Scientific Data, Volume 10, Article 41.*



2. Experimentos

- **Testes finais: CNN "simples"**



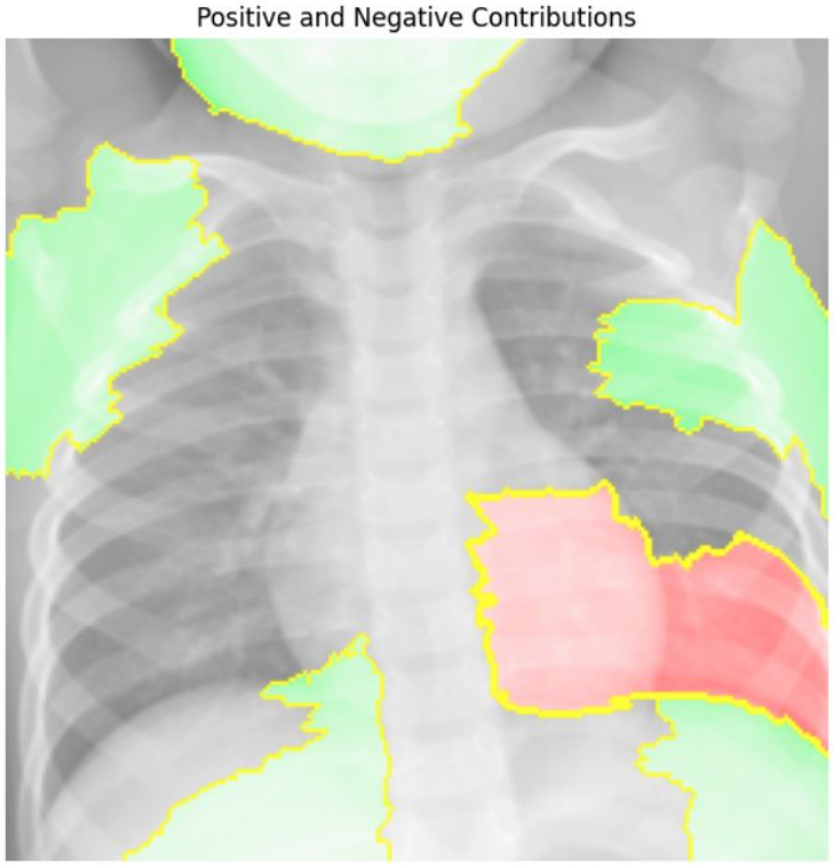


	precision	recall	f1-score	support
normal	0.93	0.77	0.85	234
pneumonia	0.88	0.97	0.92	390
accuracy			0.89	624
macro avg	0.90	0.87	0.88	624
weighted avg	0.90	0.89	0.89	624

- Normal: Alta precisão (93%), mas recall moderado (77%), indicando alguns falsos negativos.
- Pneumonia: Recall excelente (97%) e boa precisão (88%), com poucos falsos positivos.
- Acurácia Geral: Modelo acerta 89% das classificações totais.

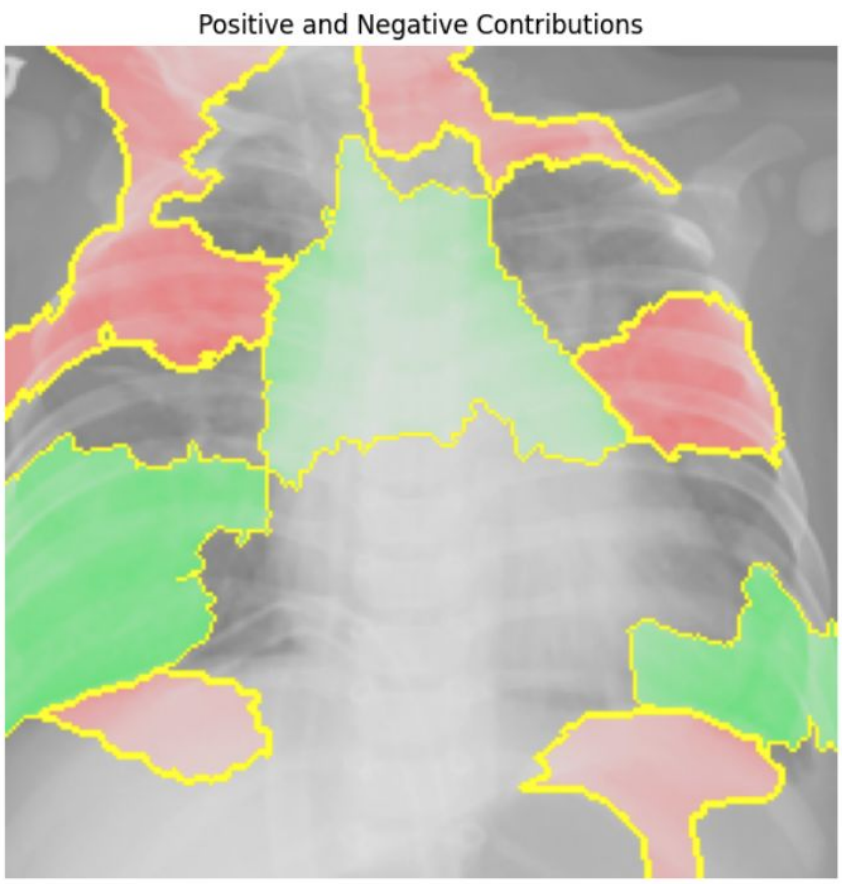
2. Experimentos

- **Testes finais: CNN "simples"**
  - LIME



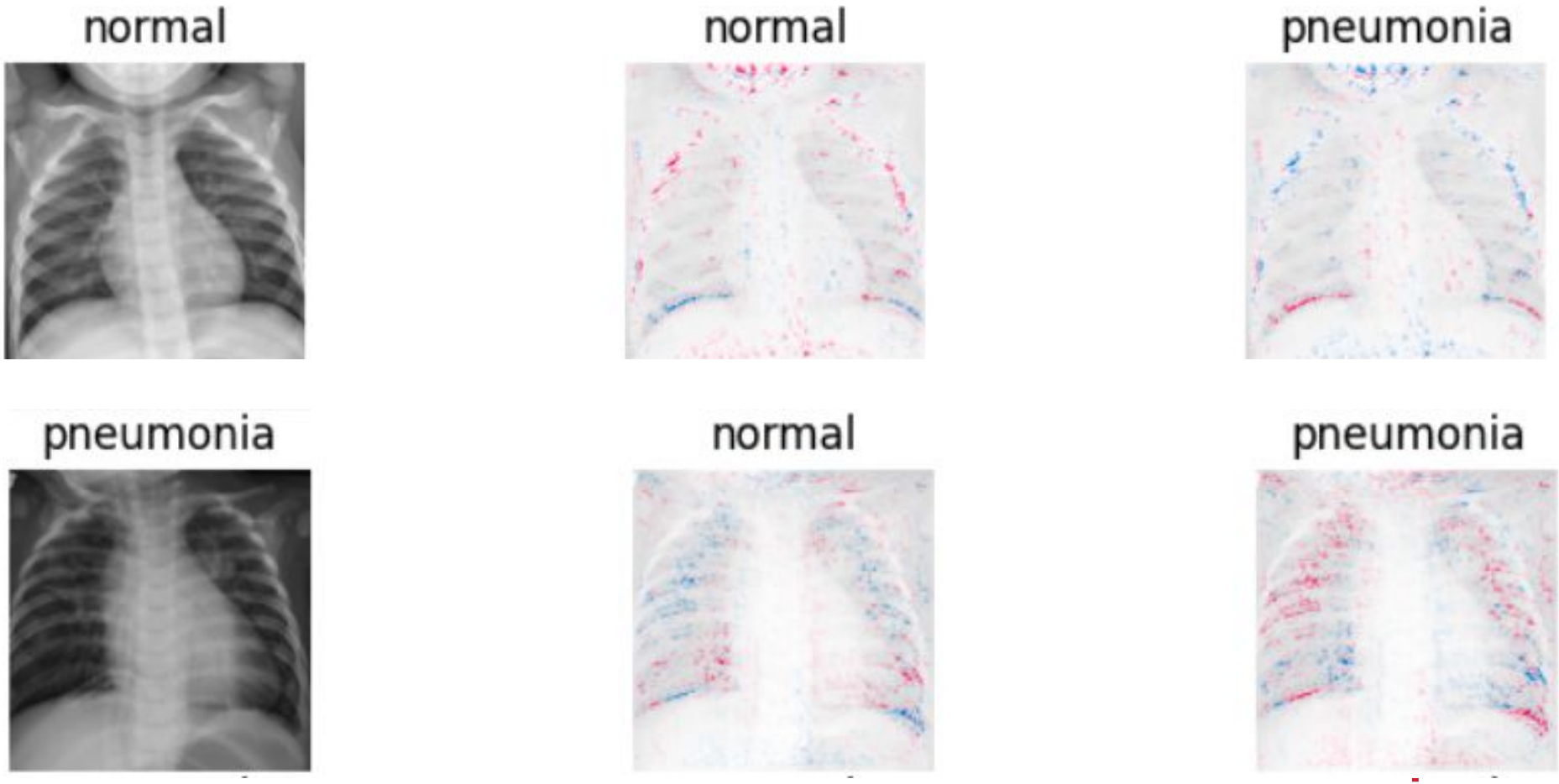
2. Experimentos

- **Testes finais: CNN "simples"**
  - LIME



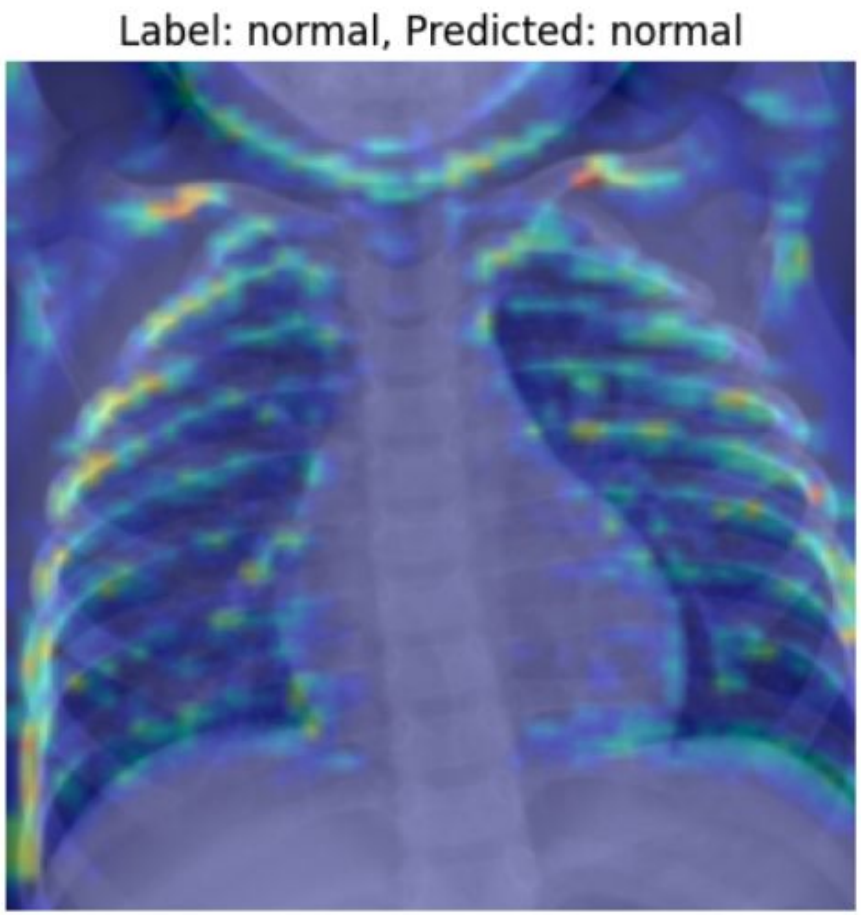
2. Experimentos

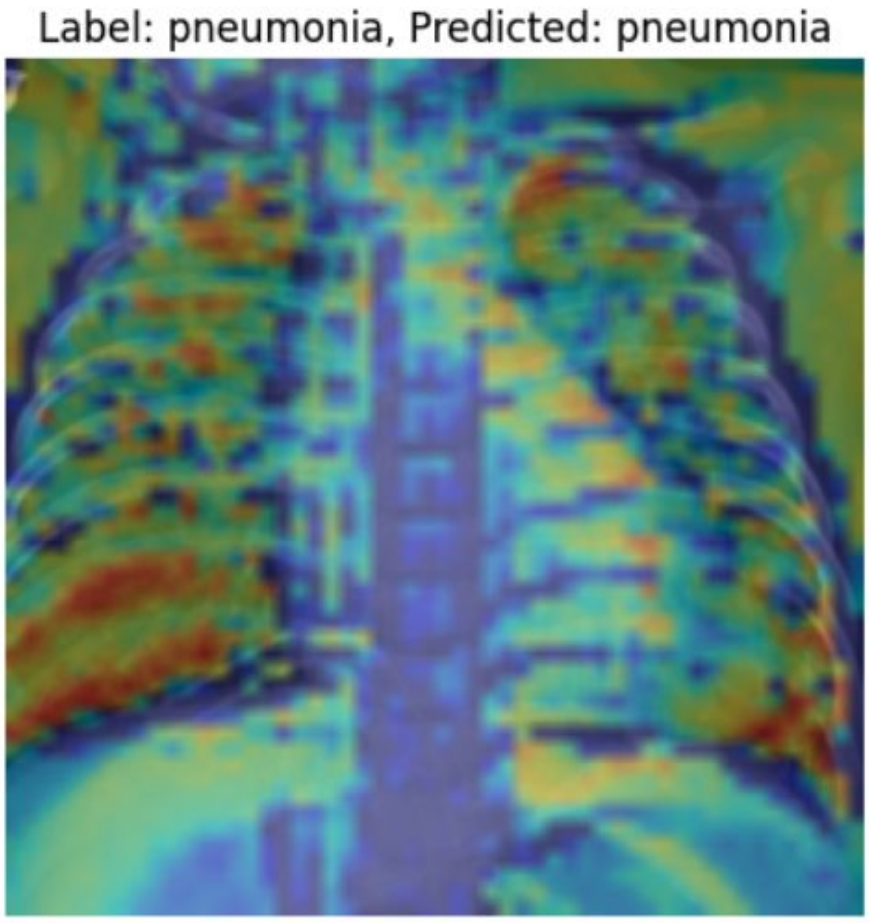
- **Testes finais: CNN "simples"**
  - SHAP



2. Experimentos

- **Testes finais: CNN "robusta"**
  - Grad-CAM





2. Experimentos

- Apesar de ainda haver espaço para melhorias nos falsos negativos, o modelo apresentou uma boa acurácia, mesmo utilizando a CNN "simples".
- Os três métodos de XAI local destacaram diferenças relevantes entre os casos de pneumonia e não pneumonia, com o Grad-CAM se sobressaindo nesse aspecto.
- O LIME se mostrou particularmente útil para especialistas em diagnóstico, pois, além de identificar os pixels que contribuíram positivamente para a classificação (regiões de certeza), também apontou os que poderiam levar a uma classificação oposta (regiões de incerteza). Essas áreas podem ser consideradas pontos de atenção no processo diagnóstico.
- Por outro lado, o SHAP apresentou-se menos intuitivo, dificultando sua aplicação em contextos que exigem maior clareza visual.

Conclusão

3. Conclusão

- Resultados Principais:
  - Modelos mais robustos (CNN "robusta") demonstraram maior acurácia geral e estabilidade nas explicações das previsões.
  - Técnicas de XAI local (LIME, SHAP, Grad-CAM) foram fundamentais para interpretar o comportamento dos modelos, destacando forças e limitações em diferentes cenários.
- Contribuições do Estudo:
  - Proporcionou uma análise detalhada de como diferentes métodos de XAI podem ser aplicados à classificação de imagens em contextos variados, como animais, insetos e imagens médicas.
  - Evidenciou o papel da explicabilidade na confiança e melhoria de modelos de aprendizado profundo.

3. Conclusão

- Implicações:
  - A adoção de técnicas de XAI em domínios críticos, como saúde, pode aumentar a transparência e a confiabilidade de sistemas baseados em IA.
  - O equilíbrio entre desempenho do modelo e interpretabilidade deve ser sempre considerado no design de soluções de IA.
- Próximos Passos:
  - Ampliar os testes para outros datasets e arquiteturas.
  - Explorar otimizações nas técnicas de XAI para maior eficiência computacional.
- Repositório: https://github.com/bzamith/XAI_ImageClassification

AGRADECEMOS A ATENÇÃO!

PERGUNTAS?