
Universidade Federal de Pernambuco
Centro de Informática

Graduação em Engenharia da Computação

# Estudo e aplicação de técnicas baseadas em *Deep Learning* para sistemas de suporte à decisão em diagnósticos de radiografias do tórax.

Gabriel Souza Marques

Trabalho de Graduação

Recife
22 de setembro de 2023

Universidade Federal de Pernambuco
Centro de Informática

Gabriel Souza Marques

# Estudo e aplicação de técnicas baseadas em *Deep Learning* para sistemas de suporte à decisão em diagnósticos de radiografias do tórax.

*Trabalho apresentado ao Programa de Graduação em Engenharia da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Engenharia da Computação.*

Orientador: *Fernando Maciano de Paula Neto*

Recife
22 de setembro de 2023

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Marques, Gabriel Souza.
Estudo e aplicação de técnicas baseadas em Deep Learning para sistemas de suporte à decisão em diagnósticos de radiografias do tórax. / Gabriel Souza Marques. - Recife, 2023.
30 p. : il., tab.

Orientador(a): Fernando Maciano Neto
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Informática, Engenharia da Computação - Bacharelado, 2023.

1. Sistemas de suporte à decisão. 2. Aprendizado de máquina. 3. Aprendizado profundo. 4. Redes neurais convolucionais. 5. Interpretabilidade. I. Neto, Fernando Maciano. (Orientação). II. Título.

000 CDD (22.ed.)

# Resumo

Os Sistemas de Apoio à Decisão Clínica (*Clinical Decision Support System* - CDSS) na área da saúde preenchem a lacuna entre as diretrizes clínicas e as decisões reais de tratamento utilizando dados de pacientes provenientes de diversas fontes. Eles fornecem recomendações de tratamento personalizadas e baseadas em diversos tipos de evidências, como imagens radiográficas. CDSS que utilizam técnicas de aprendizado de máquina (*Machine Learning* - ML) aprimoraram a análise de radiografias, ultrapassando potencialmente os radiologistas humanos em precisão. Essas técnicas tem o poder de otimizar o fluxo de trabalho, atuando como uma segunda opinião e auxiliando no rastreio de doenças em grande escala. No entanto, a integração com ML em ambientes clínicos apresenta desafios de interpretabilidade. Técnicas de interpretabilidade, como mapas de calor, ajudam os prestadores de cuidados de saúde a compreender as decisões do modelo de ML, levando a diagnósticos mais precisos e a um melhor atendimento ao paciente.

Este trabalho tem como objetivo, a partir do treinamento de uma rede neural convolucional para classificação de imagens de radiografias do tórax, a comparação de duas técnicas de interpretabilidade empregadas na geração de mapas de calor - Mapeamento de Ativação de Classe (CAM) e Mapeamento de Ativação de Classe Ponderado por Gradiente (Grad-CAM). O ponto focal da investigação reside em discernir diferenças entre CAM e Grad-CAM, ambas técnicas de interpretabilidade promissoras para desvendar os processos de tomada de decisão de redes neurais profundas.

Os resultados fundamentam a hipótese de que o Grad-CAM, supera a precisão de localização do CAM. Além disso, surge uma correlação convincente entre o grau de certeza do modelo e a precisão alcançada pelas técnicas de interpretabilidade na localização de patologias nas radiografias de tórax. Esta conexão ressalta a sinergia potencial entre a certeza diagnóstica e a eficácia das técnicas de interpretabilidade na identificação de regiões anatômicas relevantes.

Além disso, o trabalho traduz suas descobertas em aplicações práticas, desenvolvendo uma aplicação de software dedicada. Esta aplicação é elaborada para aproveitar os *insights* e técnicas derivados do estudo, tornando efetivamente os métodos acessíveis e utilizáveis em cenários do mundo real.

**Palavras-chave:** Sistemas de Suporte à Decisão, Radiografias, ChestX-ray14, Redes Neurais Convolucionais, CheXnet, Técnica de Interpretabilidade de CNN, CAM, Grad-Cam

# Abstract

Clinical Decision Support Systems (CDSS) in healthcare bridge the gap between clinical guidelines and actual treatment decisions using patient data from multiple sources. They provide personalized treatment recommendations based on different types of evidence, such as radiographic images. CDSS utilizing machine learning (ML) techniques have improved radiograph analysis, potentially surpassing human radiologists in accuracy. These techniques have the power to optimize workflow, acting as a second opinion and helping to screen for diseases on a large scale. However, integration with ML in clinical settings presents interpretability challenges. Interpretability techniques such as heatmaps help healthcare providers understand ML model decisions, leading to more accurate diagnoses and better patient care.

This work aims, based on the training of a convolutional neural network for classifying chest x-ray images, to compare two interpretability techniques used in the generation of heat maps - Class Activation Mapping (CAM) and Gradient-Weighted Class Activation (Grad-CAM). The focal point of the investigation lies in discerning differences between CAM and Grad-CAM, both promising interpretability techniques for unraveling the decision-making processes of deep neural networks.

The results support the hypothesis that Grad-CAM,outperforms the localization accuracy of CAM. Furthermore, a convincing correlation emerges between the degree of certainty of the model and the precision achieved by interpretability techniques in localizing pathologies on chest radiographs. This connection highlights the potential synergy between diagnostic certainty and the effectiveness of interpretability techniques in identifying relevant anatomical regions.

Furthermore, the work translates its findings into practical applications by developing a dedicated software application. This application is designed to leverage the insights and techniques derived from the study, effectively making the methods accessible and usable in real-world scenarios.

**Keywords:** Decision Support Systems, Radiographs, ChestX-ray14, Convolutional Neural Networks, CheXnet, CNN Interpretability Technique, CAM, Grad-Cam

# Sumário

# Lista de Figuras

# Lista de Tabelas

Capítulo 1

# Introdução

Os Sistemas de Apoio à Decisão Clínica (*Clinical Decision Support System* - CDSS) servem como ferramentas no âmbito dos cuidados de saúde modernos, servindo para preencher a lacuna entre as diretrizes clínicas e as decisões reais de tratamento tomadas pelos profissionais de saúde. Esses sistemas aproveitam de dados de pacientes provenientes de diversas origens, abrangendo prontuários eletrônicos de saúde, avaliações diagnósticas e outros repositórios médicos. Através de uma síntese desses dados, o CDSS facilita a criação de avaliações personalizadas dos pacientes, capacitando assim os prestadores de cuidados de saúde com recomendações de tratamento baseadas em evidências que enriquecem o seu processo de tomada de decisão [5]. Estas recomendações têm o potencial de influenciar positivamente os resultados dos pacientes e contribuir para o avanço da medicina de precisão.

O cenário da aquisição de dados médicos apresenta um terreno fértil para a operação do CDSS. Em particular, a integração de dados de testes de diagnóstico, incluindo, entre outros, imagens de raios X, aumenta significativamente a capacidade de diagnóstico desses sistemas. A adoção do CDSS no domínio da interpretação de imagens médicas atraiu atenção e inovação substanciais. Esses sistemas têm a capacidade de analisar padrões e anomalias em imagens médicas, ajudando assim os profissionais de saúde a identificar rapidamente possíveis patologias e a tomar decisões clínicas bem informadas [12]. A radiografia de tórax, o exame de imagem mais comum globalmente, tem grande importância dentro do contexto devido à sua ampla utilização no rastreamento, diagnóstico e orientação no tratamento de várias doenças críticas. A sua natureza não invasiva e a capacidade de revelar informações vitais sobre as estruturas torácicas, incluindo o coração, os pulmões e a anatomia circundante, sublinham o seu papel central na tomada de decisões clínicas em diversas especialidades médicas [11].

A evolução da tecnologia médica inaugurou uma nova era de capacidades de diagnóstico, com a inteligência artificial (IA) e a aprendizagem automática relacionando-se com a radiologia. A aplicação dessas tecnologias avançadas à radiografia de tórax abriu caminhos promissores para aumentar a precisão e a eficiência do diagnóstico. Algoritmos de aprendizado de máquina treinados em vastos conjuntos de dados demonstraram o potencial para analisar autonomamente radiografias de tórax, reconhecendo padrões sutis e anomalias que podem escapar até mesmo aos radiologistas experientes. O potencial transformador do suporte de diagnóstico alimentado por IA é ainda mais acentuado pela sua capacidade de facilitar a interpretação rápida e consistente de imagens, mitigando potencialmente a variabilidade inerente às avaliações humanas [4].

Conseguir uma interpretação automatizada das radiografias de tórax que corresponda

ou supere a experiência dos radiologistas traz uma série de vantagens de longo alcance. A otimização do fluxo de trabalho torna-se uma realidade alcançável, pois os algoritmos de IA podem agilizar a análise de imagens, permitindo que os profissionais de saúde aloquem mais tempo para a interação com o paciente e o gerenciamento de casos complexos. Além disso, a implementação de sistemas de apoio à decisão clínica baseados em IA pode servir como uma segunda opinião valiosa, corroborando os resultados diagnósticos e enriquecendo o processo de pensamento diagnóstico dos profissionais médicos. Essa abordagem colaborativa entre a IA e os médicos aumenta a probabilidade de diagnósticos precisos, melhorando os resultados dos pacientes e a eficácia do tratamento [7]. Além disso, o impacto vai além do atendimento individual ao paciente. A interpretação automatizada de radiografias de tórax com tecnologia de IA abre caminho para iniciativas de triagem em grande escala, especialmente em regiões com recursos limitados, onde o acesso a cuidados de saúde especializados é limitado. Ao auxiliar na detecção precoce de doenças como tuberculose, câncer de pulmão e patologias cardiovasculares, a interpretação da radiografia de tórax baseada em IA tem o potencial de reduzir significativamente a progressão da doença e reduzir as taxas de mortalidade em escala global [8].

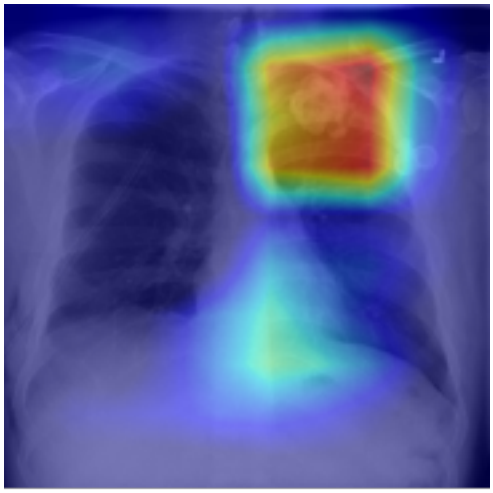

**Figura 1.1** Mapa de calor em radiografia do tórax.

Nos últimos anos, tem havido um aumento notável na integração de técnicas de IA em ambientes clínicos, com ênfase específica no avanço do CDSS [3]. Essa tendência generalizada estende seu alcance ao domínio da análise automatizada de radiografias de tórax, onde modelos de aprendizado de máquina são aproveitados para reforçar a precisão das previsões diagnósticas. No entanto, a assimilação da aprendizagem automática em contextos clínicos introduz um desafio, centrado na interpretabilidade desses sistemas. Embora os esforços convencionais de aprendizado de máquina priorizem a precisão preditiva, as aplicações clínicas exigem uma ênfase de mesmo grau no fornecimento aos profissionais de saúde de elucidações sobre a lógica por trás das decisões do modelo. Para isso, as técnicas de interpretabilidade emergem como ferramentas essenciais destinadas a aumentar a transparência dos vereditos dos modelos baseados em IA. Dentre essas es-

tratégias, uma abordagem notável envolve a geração de mapas de calor (Figura 1.1) que acentuam regiões de interesse presentes nas imagens de raios X. Esses mapas de calor se materializam como resultado da implantação de técnicas de localização não supervisionadas, que por sua vez revelam os pixeis da imagem que exercem a influência mais profunda no processo de tomada de decisão do modelo. Ao identificar estas zonas cruciais, podem ser discernidas anomalias indicativas de patologias, permitindo assim que os prestadores de cuidados de saúde canalizem a sua atenção com precisão. Consequentemente, gera-se o potencial para diagnósticos mais meticulosos, alinhando-se com os objetivos globais de reforçar o atendimento ao paciente e a eficácia diagnóstica [25].

Capítulo 2

# Referencial Teórico

## 2.1 Aprendizado de máquina

O aprendizado de máquina é um subconjunto da IA que se concentra no desenvolvimento de algoritmos e modelos estatísticos, permitindo que computadores melhorem seu desempenho em uma tarefa específica por meio de experiência e dados, sem serem explicitamente programados. Ganhou ampla atenção e significado em várias áreas, revolucionando as indústrias e permitindo aplicações inovadoras [9]. Os próximos parágrafos falam sobre os conceitos fundamentais envolvendo o aprendizado de máquina.

O primeiro conceito é a aprendizagem supervisionada. Aprendizagem supervisionada é um tipo de aprendizado de máquina em que algoritmos aprendem a partir de dados de treinamento rotulados para fazer previsões ou decisões sem intervenção humana. As tarefas comuns incluem classificação e regressão. Em contraste existe a aprendizagem não supervisionada. Na aprendizagem não supervisionada os algoritmos aprendem a partir de dados não rotulados para descobrir padrões, estruturas ou *insights* ocultos nos dados. Agrupamento e redução de dimensionalidade são aplicações típicas [9].

Outro conceito importante dentro do contexto de aprendizado de máquina é a aprendizagem por reforço. Na aprendizagem por reforço concentra-se no treinamento de agentes para tomar sequências de decisões em um ambiente para maximizar uma recompensa cumulativa. Este paradigma é crucial para aplicações como robótica e jogos [9].

Por último o aprendizado profundo ou *Deep learning*. O *Deep learning* é um subcampo do aprendizado de máquina que usa redes neurais com múltiplas camadas (redes neurais profundas) para modelar padrões complexos em dados. Redes Neurais Convolucionais (CNNs) para análise de imagens e Redes Neurais Recorrentes (RNNs) para dados sequenciais são arquiteturas populares de aprendizado profundo [9].

## 2.2 Redes Neurais Convolucionais

Redes Neurais Convolucionais (*Convolutional Neural Network* - CNN) representam um avanço fundamental no campo de aprendizado de máquina, particularmente em tarefas de visão computacional. Elas permitiram avanços na classificação de imagens, detecção de objetos, reconhecimento facial, imagens médicas e muitas outras aplicações [20]. São uma classe especializada de redes neurais profundas projetadas para processar dados semelhantes a grades, como imagens e vídeos [9]. As CNNs ganharam imensa popularidade devido à sua capacidade de aprender e extrair automaticamente recursos hierárquicos de

dados brutos de entrada. Seu sucesso em diversas tarefas relacionadas a imagens pode ser atribuído à sua capacidade de capturar hierarquias e padrões espaciais de forma eficiente [20]. Diversas arquiteturas foram propostas em diversos estudos como a ImageNet [15] ou DenseNet [10], porém uma CNN normalmente possui três camadas: uma camada convolucional, uma camada de *pooling* e uma camada densa.

### 2.2.1 Camada de Convolução

A camada convolucional é o alicerce central da CNN, suportando o peso da carga de trabalho computacional da rede. A função principal desta camada envolve a execução de uma operação de produto escalar entre duas matrizes: o filtro ou *kernel*, que contém os parâmetros ajustáveis, enquanto a outra matriz abrange a região confinada do campo receptivo. O filtro é espacialmente menor que uma imagem, mas é mais profundo. Consequentemente, quando a imagem de entrada compreende, por exemplo, três canais (RGB), a altura e a largura do filtro permanecem espacialmente compactas, enquanto sua profundidade se estende por todos os três canais [9].

O filtro percorre a imagem em altura e largura, gerando uma representação da área receptiva. Esta operação resulta na criação de uma representação de imagem bidimensional denominada "mapa de ativação". Este mapa fornece informações sobre a resposta do filtro em cada localização espacial da imagem. A distância que o filtro percorre durante esse processo de "deslizamento"é chamada de passo. Esse processo é demonstrado através da figura 2.1 [9].

### 2.2.2 Camada de *pooling*

A camada de *pooling* substitui saídas de rede específicas por uma estatística resumida derivada de saídas próximas. Isso serve para diminuir as dimensões espaciais da representação, reduzindo consequentemente a carga de trabalho computacional. É importante ressaltar que a operação de *pooling* é conduzida separadamente em cada fatia da entrada [9].

Existem inúmeras funções de *pooling*, incluindo a média de uma vizinhança retangular, o cálculo da norma L2 dentro de uma vizinhança retangular e o cálculo de uma média ponderada considerando a distância do pixel central. No entanto, a técnica mais amplamente adotada é o *pooling* máximo, que identifica o maior valor dentro da vizinhança [9]. A imagem 2.2 ilustra a operação de *pooling* máximo.

### 2.2.3 Camada densa

Os neurônios nesta camada têm conectividade total com todos os neurônios da camada anterior e seguinte. Dessa forma pode ser calculado normalmente por uma multiplicação de matrizes seguida por um efeito de polarização. A camada densa ajuda a mapear a representação entre a entrada e a saída. O objetivo da camada densa é ajustar os parâmetros de peso para criar uma representação de probabilidade estocástica de cada classe com base nos mapas de ativação gerados pela concatenação de camadas convolucionais,

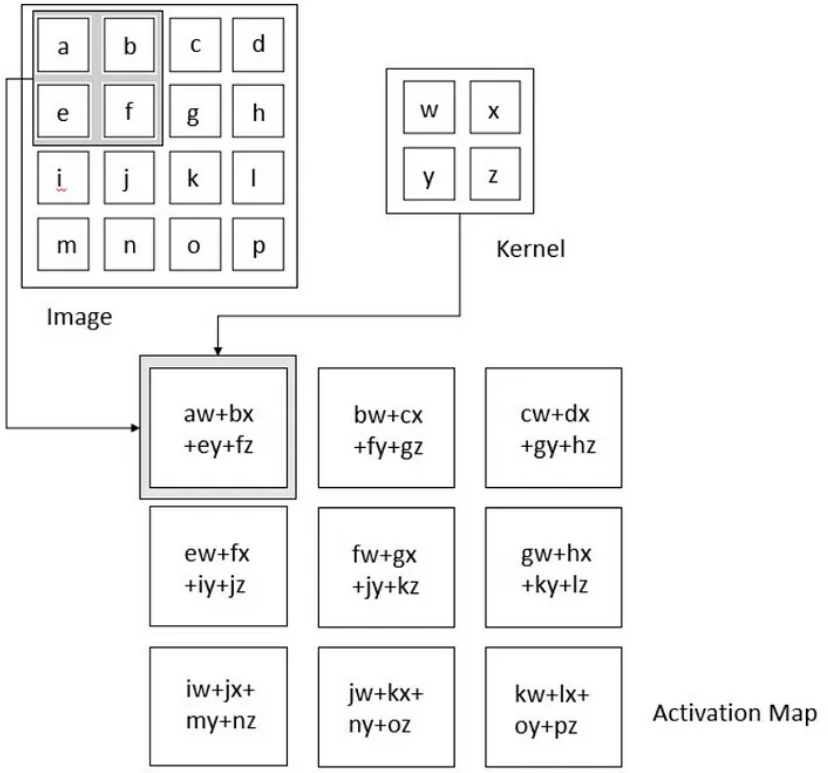


**Figura 2.1** Operação de convolução [9].

não linearidades, retificação e *pooling* [9].

## 2.3 Interpretabilidade

A interpretabilidade em sistemas baseados em IA é um aspecto vital, especialmente em aplicações onde as decisões do modelo acarretam consequências significativas, como saúde, direção autônoma e finanças. Alcançar uma compreensão mais profunda de como os modelos chegam às suas previsões é crucial por vários motivos.

Em primeiro lugar, confiança e responsabilidade: em cenários de alto risco, como o diagnóstico médico, onde os sistemas auxiliam os profissionais de saúde, os modelos interpretáveis fornecem aos médicos informações sobre a razão pela qual um determinado diagnóstico ou recomendação foi feito. Isto promove a confiança e garante que as decisões de saúde sejam tomadas de forma colaborativa entre o modelo e o especialista humano [2]. Em segundo lugar, preconceito e justiça: compreender como as técnicas tomam decisões é vital para identificar e mitigar preconceitos nas suas previsões. Os métodos de interpretabilidade podem revelar se um modelo está tomando decisões com base em características não intencionais, levando potencialmente a resultados injustos ou tendenciosos. Ao diagnosticar e corrigir essas questões, os sistemas podem fornecer resultados mais justos e equitativos [1]. Esses *insights* podem orientar melhorias e ajustes do mo-

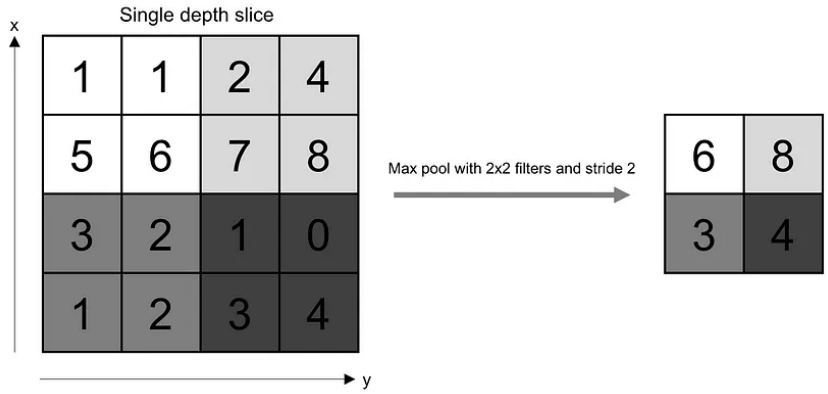


**Figura 2.2** Operação de *pooling* máximo [9].

delo, destacando regiões ou recursos específicos nos dados de entrada que podem precisar de atenção especial. Assim, a interpretabilidade auxilia na depuração e aprimoramento do modelo [24]. Além disso, conformidade e regulamentação: Em setores regulamentados como finanças ou saúde, os modelos interpretáveis simplificam a conformidade com os requisitos de transparência. Os órgãos reguladores muitas vezes exigem a capacidade de explicar as decisões do modelo, e os sistemas baseados em IA interpretáveis facilitam a adesão a esses padrões legais e éticos [16].

Para alcançar a interpretabilidade nas CNNs, vários métodos foram desenvolvidos:

- ***Feature Visualization***: esta técnica gera imagens que ativam ao máximo neurônios específicos na rede, fornecendo informações visuais sobre o que neurônios individuais estão detectando [17];
- **Grad-CAM**: esta técnica destaca as regiões cruciais em uma imagem de entrada que mais contribuem para a previsão de uma classe específica, ajudando a entender onde o modelo está concentrando sua atenção [21]. Este método será detalhado na seção 3.3;
- **LIME**: esta técnica perturba os dados de entrada e observa como as previsões da CNN mudam, oferecendo explicações locais do comportamento do modelo [19];
- **Mapas de saliência**: esta técnica destaca regiões em uma imagem de entrada que são mais relevantes para a decisão do modelo, calculando gradientes da saída em relação à entrada [22];

## 2.4 Auditoria em IA

O processo que visa avaliar o comportamento, as decisões e os resultados produzidos pelos sistemas de IA fazendo o uso da interpretabilidade é chamado de auditoria. O objetivo deste processo é garantir que os sistemas baseados em IA atendam a critérios, padrões e considerações éticas específicos. É um componente crítico do desenvolvimento e implantação responsável da IA, especialmente à medida que os sistemas de IA se tornam mais

integrados em vários aspectos da sociedade. Além de garantir a explicabilidade e transparência, a auditoria de sistemas baseados em IA também avaliam os modelos do ponto de vista de privacidade, garantindo que dados pessoais sejam tratados em conformidade com as normas regulamentadoras; do ponto de vista de segurança, avaliando pontos de vulnerabilidade a ataques e possível uso indevido; do ponto de vista da conformidade do sistema em relação às normas regulamentadoras e padrões do setor; verifica o desempenho e a precisão do sistema para garantir que atendam aos padrões e objetivos desejados; também são estabelecidos monitoramento contínuo e ciclos de *feedback* para monitorar o desempenho dos sistemas de IA em cenários do mundo real, ajudando a identificar problemas e fazer os ajustes necessários; e define a documentação necessária no processo de desenvolvimento do sistema [24].

Os desafios na interpretabilidade e auditoria em IA permanecem, incluindo o tratamento de dados de alta dimensão e a navegação nos compromissos entre a complexidade do modelo e a interpretabilidade. A investigação nesta área continua em evolução, procurando sistemas de IA mais transparentes, responsáveis e justos [24].

Capítulo 3

# Trabalhos Relacionados

## 3.1 CheXnet

Numerosos modelos foram concebidos para aprimorar os CDSS no contexto da radiografia de tórax [23, 11, 13, 18]. Notavelmente, um destaque em termos de classificação de patologias é o modelo CheXnet. Este modelo avança na precisão diagnóstica ao apresentar uma visão das probabilidades associadas a várias condições patológicas. Ele alcança isso aproveitando uma técnica de interpretação de redes neurais chamada Mapas de Ativação de Classes (*Class Activation Mapping* - CAM) [25], que destaca regiões específicas nas imagens de raios-X do tórax que possuem pistas fundamentais para cada patologia.

A arquitetura do CheXnet é formada por uma rede convolucional DenseNet [10] com uma profundidade de 121 camadas. Adaptações são feitas na arquitetura convencional, incluindo a substituição da camada final totalmente conectada por uma camada global de *pooling* médio infundida com uma não-linearidade sigmoide.

Para a realização dos testes da CheXnet foi coletado um conjunto de teste de 420 radiografias de tórax frontal. As anotações foram obtidas independentemente de quatro radiologistas praticantes da Universidade de Stanford, que foram solicitados a rotular todas as 14 patologias. Os radiologistas tinham 4, 7, 25 e 28 anos de experiência. Os radiologistas não tiveram acesso a nenhuma informação do paciente ou conhecimento da prevalência da doença nos dados. É possível ver os resultados da CheXnet a partir da tabela 3.1 e 3.2, onde na tabela 1 mostra-se a área sob a curva ROC (*Area Under the Roc Curve* - AUROC) para cada classe e na tabela 2 é feita uma comparação do F1 score de radiologistas com o modelo [18].

## 3.2 *Class Activation Mapping* - CAM

Conforme destacado anteriormente, o CheXnet apresenta a capacidade de geração de imagens de raios X enriquecidas com um mapa de calor que identifica as regiões mais significativas para o diagnóstico de cada patologia. Esta capacidade é alcançada através da aplicação de uma técnica de interpretabilidade de rede neural conhecida como Mapas de Ativação de Classe (*Class Activation Mapping* - CAM). Esta técnica é fundamentada nos princípios da localização de objetos não supervisionada, propondo um algoritmo autodidata para localização de objetos. Em sua essência, a técnica CAM é adaptada para arquiteturas que dependem predominantemente de camadas convolucionais. As etapas principais desta técnica se desenrolam pouco antes da camada de saída final (geralmente

| **Patologia** | **AUROC** |
|---|---|
| Atelectasia | 0.8094 |
| Consolidação | 0.7901 |
| Infiltração | 0.7345 |
| Pneumotórax | 0.8887 |
| Edema | 0.8878 |
| Enfisema | 0.9371 |
| Fibrose | 0.8047 |
| Efusão | 0.8638 |
| Pneumonia | 0.7680 |
| Espessamento Pleural | 0.8062 |
| Cardiomegalia | 0.9248 |
| Nódulo | 0.7802 |
| Massa | 0.8676 |
| Hérnia | 0.9164 |

**Tabela 3.1** Tabela de resultados da CheXnet

| | **F1 Score (95% de Intervalo de Confiança)** |
|---|---|
| Radiologista 1 | 0.383 (0.309, 0.453) |
| Radiologista 2 | 0.356 (0.282, 0.428) |
| Radiologista 3 | 0.365 (0.291, 0.435) |
| Radiologista 4 | 0.442 (0.390, 0.492) |
| Média de Radiologistas | 0.387 (0.330, 0.442) |
| CheXnet | 0.435 (0.387, 0.481) |

**Tabela 3.2** Tabela comparação da CheXnet com radiologistas

uma camada softmax em tarefas de categorização). Aqui, o *pooling* médio global é executado nos mapas de ativação da última camada convolucional. Este processo é então aproveitado para alimentar uma camada totalmente conectada que gera o resultado desejado, seja categórico ou não. Ilustrado na Figura 3.1 , o procedimento de *pooling* médio global calcula a média espacial do mapa de ativações para cada unidade dentro da camada convolucional final. Uma soma ponderada destas médias constitui a base para o resultado final [25].

Matematicamente, para uma determinada imagem, considere que $f_k(x, y)$ representa a ativação da unidade $k$ na última camada convolucional na localização espacial $(x, y)$. Então, para a unidade $k$, o resultado da realização do *pooling* médio global, $F_k$ é $\sum_{x,y} f_k(x, y)$. Assim, para uma dada classe $c$, a entrada para o softmax, $S_c$, é $\sum_k w_k^c F_k$ onde $w_k^c$ é o peso correspondente à classe $c$ para a unidade $k$. Essencialmente, $w_k^c$ indica a importância de $F_k$ para a classe $c$. Finalmente, a saída do softmax para a classe $c$, $P_c$ é dada por

$\frac{exp(Sc)}{\sum_c exp(Sc)}$. Ao inserir $F_k = \sum_{x,y} f_k(x,y)$ na pontuação da classe, $Sc$, obtem-se

$$S_c = \sum_k w_k^c \sum_{x,y} f_k(x,y) = \sum_{x,y} \sum_k w_k^c f_k(x,y)$$

Define-se $Mc$ como o mapa de ativação de classe para a classe $c$, onde cada elemento espacial é dado por

$$M_c(x,y) = \sum_k w_k^c f_k(x,y)$$

$M_c(x,y)$ indica diretamente a importância da ativação na grade espacial $(x,y)$ levando à classificação de uma imagem na classe $c$ [25].

Entretanto, esta habilidade para localização de objetos não supervisionados diminui em cenários onde camadas totalmente conectadas precedem a camada de *pooling* [25]. É importante notar que a arquitetura básica do CheXnet caracteriza-se por conexões densas interpostas entre camadas convolucionais [18]. Em teoria, esta interligação poderia impactar a capacidade de gerar CAMs utilizando a técnica de interpretabilidade, atenuando potencialmente a capacidade de localização não supervisionada.

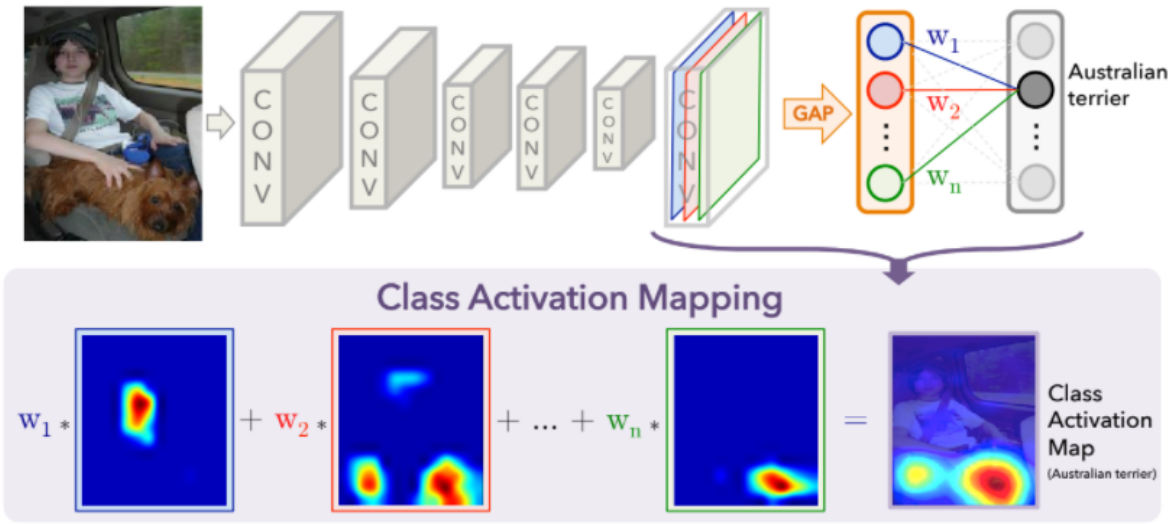


**Figura 3.1** *Class Activation Mapping*: a pontuação de classe prevista é mapeada de volta para a camada convolucional anterior para gerar os CAMs. O CAM destaca as regiões discriminativas específicas da classe [25].

## 3.3 Grad-Cam

Um algoritmo de localização não supervisionado adicional no domínio da IA explicável é o Grad-Cam, que propõe uma abordagem para gerar "explicações visuais"para decisões fornecidas por uma gama diversificada de modelos baseados em CNN. Denominada

Mapeamento de Ativação de Classe Ponderada por Gradiente (*Gradient-weighted Class Activation Mapping*), esta metodologia aproveita os gradientes inerentes originados de qualquer conceito de destino designado - um conceito que a rede é treinada para reconhecer, como 'cachorro' em uma rede de classificação ou uma sequência de palavras em uma rede de legendas. Esses gradientes são aproveitados como um recurso para fabricar um mapa de localização, destacando as regiões salientes dentro de uma imagem que exercem influência máxima sobre a previsão da rede em relação ao conceito designado [21].

O funcionamento do Grad-CAM depende do papel central das camadas convolucionais na extração e abstração de informações nas CNNs. Ao alinhar-se com a camada convolucional final, o Grad-CAM combina as sutilezas do fluxo gradiente, permitindo a identificação de regiões específicas da imagem que são importantes na determinação da previsão da rede. Como consequência, as "explicações visuais" do Grad-CAM servem como uma ponte entre os cálculos da rede e as capacidades cognitivas humanas, facilitando a comunicação de decisões baseadas em IA de uma forma inteligível [21]. Ao contrário de abordagens precedentes, como a CAM, Grad-CAM, sem mudanças arquitetônicas ou retreinamento, é aplicável a uma grande variedade de famílias de modelos CNN:

- CNNs com camadas totalmente conectadas (por exemplo, VGG);
- CNNs usados para saídas estruturadas (por exemplo, legendagem);
- CNNs usados em tarefas com entradas multimodais (por exemplo, resposta visual de perguntas);
- Aprendizagem por reforço [21].

Conforme ilustrado na figura 3.2, o processo operacional da técnica Grad-CAM começa com o fornecimento de uma imagem e uma classe de destino específica como entradas. A imagem é canalizada através da arquitetura da CNN, aproveitando suas camadas para desvendar a essência da informação visual. Posteriormente, cálculos específicos da tarefa são empregados para derivar uma pontuação bruta correspondente à categoria de interesse designada. Nesta condição, ocorre uma manipulação dos gradientes, este sinal é retro propagado para os mapas de ativações convolucionais retificados que encapsula recursos dentro da rede [21].

A convergência desses gradientes retro propagados nos mapas de ativações convolucionais produz a base para gerar a localização grosseira do Grad-CAM. Este esquema de localização sintetiza características salientes, destacando regiões da imagem essenciais para a tomada de decisão do modelo em relação à classe alvo. Este destaque de áreas críticas da imagem acentua onde o modelo deve concentrar a sua atenção para chegar a uma previsão específica [21].

Matematicamente, como mostrado na figura 3.2, para obter o mapa de localização discriminativo de classe Grad-CAM $L^c_{Grad-CAM} \in \mathbb{R}^{u \times v}$ de largura $u$ e altura $v$ para qualquer classe $c$, primeiro calcula-se o gradiente da pontuação para classe $c$, $y^c$ (antes do softmax), em relação às ativações do mapa de ativações $A^k$ de uma camada convolucional, ou seja, $\frac{\partial y^c}{\partial A^k}$. Esses gradientes que fluem de volta são agrupados pela média global nas dimensões de largura e altura (indexados por i e j respectivamente) para obter os pesos

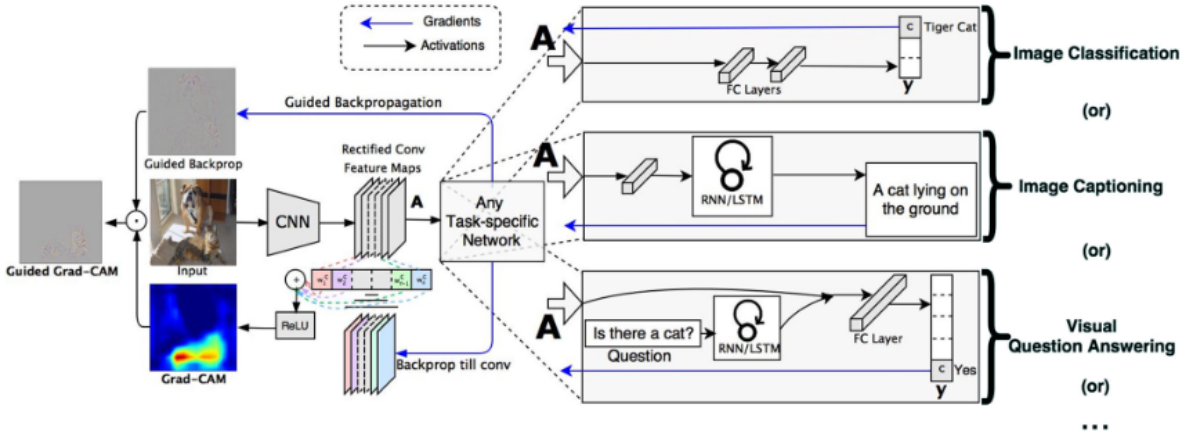


**Figura 3.2** Visão geral do Grad-CAM [21]

de importância dos neurônios $\alpha_k^c$:

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\textit{pooling}\text{ médio global}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradientes por retroprop}}$$

Durante o cálculo de $\alpha_k^c$ durante a retro propagação de gradientes em relação às ativações, o valor exato do cálculo equivale a produtos matriciais sucessivos das matrizes de peso e o gradiente em relação às funções de ativação até a camada de convolução final para a qual os gradientes estão sendo propagados. Portanto, este peso $\alpha_k^c$ representa uma linearização parcial da rede profunda a jusante de A e captura a "importância"do mapa de ativações $k$ para uma classe alvo $c$. É realizada uma combinação ponderada de mapas de ativação direta através de um ReLU para obter,

$$L_{Grad-CAM}^c = ReLU \underbrace{(\sum_k \alpha_k^c A^k)}_{\text{combinação linear}}$$

Observe que isso resulta em um mapa de calor grosseiro do mesmo tamanho dos mapas de ativação convolucionais. A ReLU é aplicada à combinação linear de mapas porque o interesse é apenas nas características que têm uma influência positiva na classe de interesse, ou seja, pixeis cuja intensidade deve ser aumentada para aumentar $y^c$ [21].

Em uma fusão em camadas de técnicas de interpretabilidade, o mapa de calor derivado da retro propagação guiada entra em cena. Este mapa de calor, que incorpora *insights* obtidos da retro propagação guiada, está entrelaçado com as visualizações geradas pelo Grad-CAM. A combinação dessas técnicas culmina na criação de visualizações guiadas Grad-CAM [21].

É possível perceber conexões, ou seja, similaridades de funcionamento, entre o Grad-CAM e o CAM e possível provar formalmente que o Grad-CAM generaliza o CAM para

uma ampla variedade de arquiteturas baseadas em CNN, inclusive CNNs caracterizadas por conexões densas interpostas entre camadas convolucionais [21].

Capítulo 4

# Metodologia

O objetivo deste trabalho é, a partir do treinamento da CheXnet, a comparação de duas técnicas de interpretabilidade empregadas na geração de mapas de calor - CAM e Grad-CAM. A metodologia proposta abrange a implantação de técnicas de interpretabilidade para fornecer dois mecanismos distintos de visualização. Então, é realizada uma avaliação e análise comparativa das capacidades de localização destas duas técnicas.

Através de experimentos, este estudo visa verificar se a integração do Grad-Cam com a CheXnet aumenta a sua capacidade de interpretabilidade. A hipótese subjacente a este esforço postula que a abordagem de visualização do Grad-Cam pode produzir mapas de calor mais precisos, aumentando assim a profundidade dos *insights* obtidos no processo de tomada de decisão do CheXnet. Esta hipótese está ancorada na antecipação de que o alinhamento do Grad-Cam com camadas convolucionais poderá permitir superar certas limitações encontradas quando a técnica CAM é aplicada a modelos com camadas densas, como é o caso do CheXnet [25, 21, 18].

Espera-se que as descobertas experimentais revelem novos *insights* sobre as complexidades das técnicas de interpretabilidade e suas interações com a arquitetura do modelo subjacente. Notavelmente, pretende elucidar se as melhorias do Grad-Cam na interpretabilidade do CheXnet estão alinhadas com o seu desempenho de classificação, validando assim o seu potencial como uma ferramenta para enriquecer os diagnósticos médicos baseados em IA.

A figura 4.1 demonstra de forma resumida os passos para a realização do estudo. Inicialmente é feita a retirada, da base de dados, as imagens de treinamento, teste, validação e os *bouding boxes* utilizados para os testes de localização. Em seguida, as imagens de treinamento são balanceadas através de um processo que será detalhado posteriormente. Então é feito o treinamento do modelo CheXnet para a classificação de patologias em imagens de radiografias do tórax. Logo após, realiza-se a aplicação das técnicas de interpretabilidade (CAM e Grad-CAM) nas imagens que possuem *bounding boxes* anotados. Por fim, ocorre a comparação utilizando a métrica MIoU para a avaliação dos mapas de calor gerados pelas técnicas no quesito precisão de localização.

## 4.1 Base de dados

O banco de dados utilizado para o treinamento da CheXnet será o ChestX-ray14 [23]. Um *dataset* que compreende uma compilação de 112.120 imagens de radiografias de tórax frontal provenientes de 30.805 pacientes únicos. Este conjunto de dados possui anotações

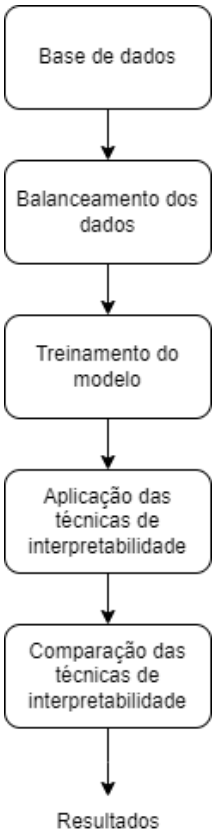


**Figura 4.1** Diagrama geral do estudo.

selecionadas que abrangem a presença ou ausência de 14 observações radiográficas de tórax comuns. A lista das 14 condições patológicas rotuladas no ChestX-ray14 são:

- Atelectasia;
- Consolidação;
- Infiltração;
- Pneumotórax;
- Edema;
- Enfisema;
- Fibrose;
- Efusão;
- Pneumonia;
- Espessamento Pleural;
- Cardiomegalia;
- Nódulo;
- Massa;
- Hérnia.

Estas condições, por sua vez, podem ser utilizadas pelos profissionais da saúde para

diagnosticar 8 doenças diferentes [23]. Além disto, o ChestX-ray14 fornece aproximadamente 1000 *bounding boxes* rotulados com uma das condições listadas, localizando na imagem a observação radiográfica que causou o diagnóstico [23]. A frequência de classes do ChestX-ray14 é encontrada na figura 4.2, assim como a frequência de rótulos positivos e negativos para cada classe encontra-se na figura 4.3. A quantidade de imagens com *bounding boxes* agrupadas por classe é encontrada na tabela 4.1.

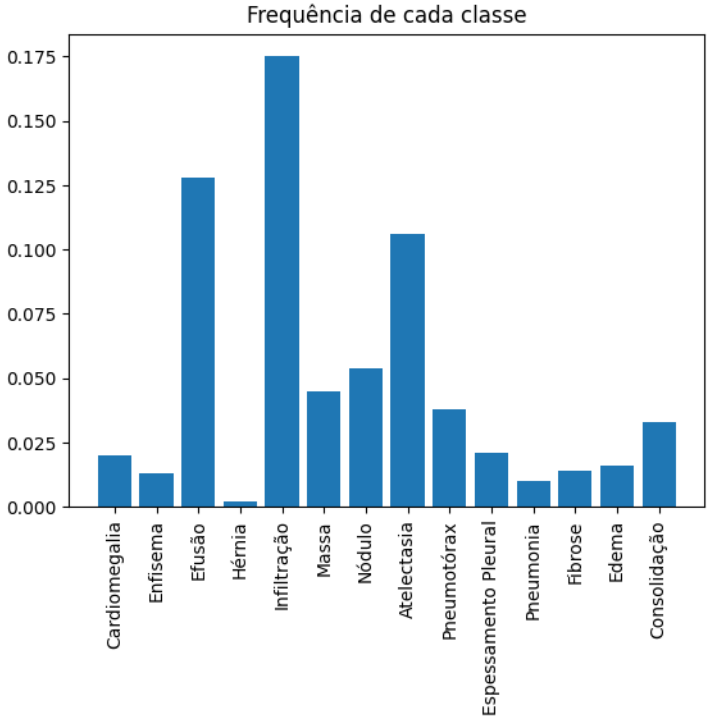


**Figura 4.2** Frequência das classes do ChestX-ray14

## 4.2 Balanceamento

Como observado na figura 4.3, há um desbalanceamento entre a frequência de rótulos positivos e negativos para cada classe do ChestX-ray14. Na CheXnet o problema é solucionado a partir da modificação da função de perda do modelo (*Binary Cross Entropy*). Para que o balanceamento entre as classes seja feito na função de perda, é preciso que

$$w_p \times freq_p = w_n \times freq_n$$

onde $w_p$ é o peso positivo da função de perda, $w_n$ o peso negativo, $freq_p$ a frequência de rótulos positivos da classe e $w_{neg}$ a frequência de rótulos negativos da classe. Para isso basta que

$$w_p = freq_n$$

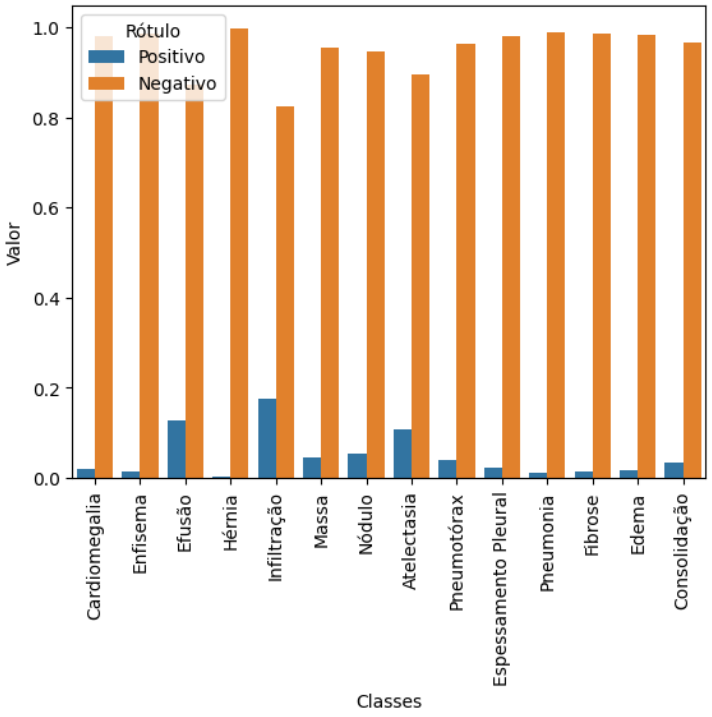


**Figura 4.3** Frequência de rótulos positivos e negativos para cada classe do ChestX-ray14

$$w_n = freq_p$$

. Dessa forma a função de perda é definida da seguinte maneira:

$$\mathcal{L}^{w}_{cross-entropy}(x) = -(w_p y \log(f(x)) + w_n(1-y)\log(1-f(x))).$$

## 4.3 Treinamento

O processo de treinamento do CheXnet começa com sua inicialização de pesos, onde os pesos pré-treinados de um modelo treinado no conjunto de dados ImageNet [6] servem como um ponto de partida [18]. É utilizado o otimizador Adam com parâmetros padrão ($\beta 1 = 0{,}9$ e $\beta 2 = 0{,}999$) e adota mini-*batches* de tamanho 16. É utilizada uma taxa de aprendizado inicial de 0,001 que é diminuída por um fator de 10 cada vez que a perda de validação se estabiliza após uma época. O modelo que apresenta a menor perda de validação é selecionado. Antes de inserir as imagens na rede, a escala das imagens é reduzida para 224×224 e é normalizada com base na média e no desvio padrão das imagens no conjunto de treinamento ImageNet [18].

A CheXnet é desenvolvida e treinada através de uma máquina virtual (VM) hospedada na plataforma Google Colab com a configuração padrão para VMs com GPU (Nvidia K80, 12GB). É utilizado um subconjunto do ChestX-ray14 de 2100 imagens separadas obedecendo ao padrão 70% para treino, 20% para teste e 10% para validação utilizado na CheXnet. A seleção é feita garantindo que o subconjunto mantenha as frequências das

| Classe | Quantidade de Imagens |
|---|---|
| Atelectasia | 180 |
| Infiltração | 123 |
| Pneumotórax | 98 |
| Efusão | 153 |
| Pneumonia | 120 |
| Cardiomegalia | 146 |
| Nódulo | 79 |
| Massa | 85 |

**Tabela 4.1** Quantidade de imagens com *bounding box* agrupadas por classe no ChestX-ray14

classe assim como a ocorrência de positivos e negativos para cada classe e a unicidade de pacientes, com o objetivo de reproduzir o treinamento original o tanto quanto possível. Dessa forma são 1470 imagens para treino, 420 para teste e 210 para validação. O desenvolvimento do CAM e do Grad-Cam são feitos seguindo o seu método de funcionamento descrito nas seções 3.2 e 3.3.

## 4.4 Aplicação

Para avaliar e comparar a eficácia das duas técnicas de interpretabilidade, será aproveitado um conjunto de dados composto por 984 imagens equipadas com *bounding boxes* do banco de dados ChestX-ray14 [23]. Dentro desta estrutura experimental, tanto o CAM quanto o Grad-CAM serão aplicados independentemente a cada imagem. O resultado será a geração de mapas de calor, que por sua vez servirão como modelos para delinear *bounding boxes* por meio da segmentação de mapas de calor. Esses *bounding boxes*, assim gerados, serão justapostos aos *bounding boxes* anotados no conjunto de dados ChestX-ray14.

## 4.5 Comparação

A análise comparativa será baseada na utilização da métrica média da intersecção sob união (*Mean Intersection Over Union* (MIoU)) – um critério de avaliação empregado para avaliar a precisão dos detectores de objetos dentro de um determinado conjunto de dados. O princípio do MIoU está na medida Interseção sobre União (IoU), que gira em torno do cálculo da razão entre a área de interseção de dois *bounding boxes* e a área de união que elas ocupam coletivamente. MIoU agrega os valores de IoU em uma média para todas as imagens dentro do conjunto de dados de teste, culminando em uma avaliação da capacidade de localização das técnicas de interpretabilidade.

Capítulo 5

# Resultados

## 5.1 Treinamento

A avaliação do treino da CheXnet foi feita a partir das métricas acurácia, sensibilidade, especificidade e AUROC, como e mostrado na tabela 5.1. O grafico 5.1 demonstra as curvas ROC para cada uma das classes treinadas.

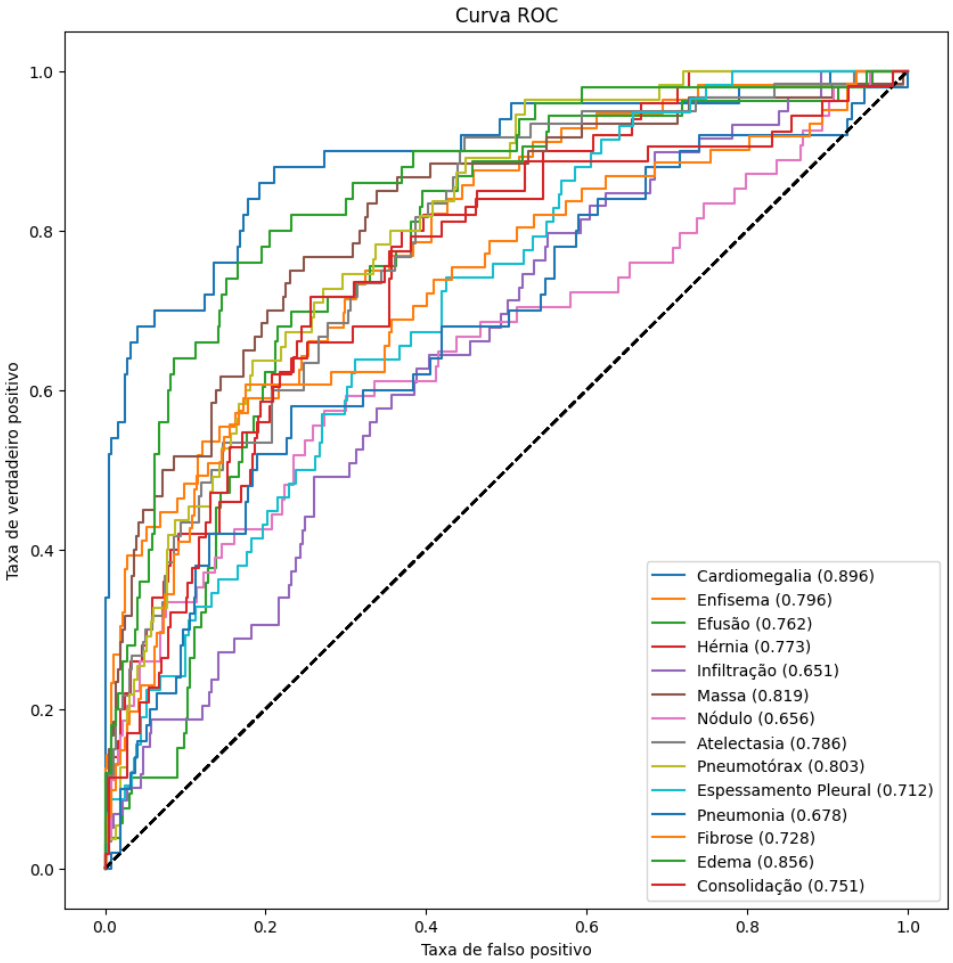


**Figura 5.1** Curvas ROC resultantes do treinamento para cada classe.

| Classe | Acurácia | Sensibilidade | Especificidade | AUROC |
|---|---|---|---|---|
| Atelectasia | 0.671 | 0.733 | 0.661 | 0.786 |
| Consolidação | 0.61 | 0.811 | 0.58 | 0.751 |
| Infiltração | 0.602 | 0.644 | 0.596 | 0.651 |
| **Pneumotórax** | **0.702** | **0.745** | **0.696** | **0.803** |
| **Edema** | **0.781** | **0.8** | **0.778** | **0.856** |
| Enfisema | 0.76 | 0.607 | 0.783 | 0.796 |
| Fibrose | 0.648 | 0.623 | 0.652 | 0.728 |
| Efusão | 0.688 | 0.736 | 0.681 | 0.762 |
| Pneumonia | 0.64 | 0.6 | 0.646 | 0.678 |
| Espessamento Pleural | 0.61 | 0.672 | 0.599 | 0.712 |
| **Cardiomegalia** | **0.826** | **0.76** | **0.835** | **0.896** |
| Nódulo | 0.657 | 0.593 | 0.667 | 0.656 |
| **Massa** | **0.767** | **0.75** | **0.769** | **0.819** |
| Hérnia | 0.698 | 0.66 | 0.703 | 0.773 |

**Tabela 5.1** Tabela de resultados do treinamento. As linhas destacadas representam as classes que obtiveram os melhores resultados.

## 5.2 Comparação

Uma avaliação da MIoU foi conduzida em todo o conjunto de dados ChestX-ray14, abrangendo imagens com anotações e os *bounding-box* gerados a partir da segmentação dos mapas de calor. Diversos valores de *thresholds* para a segmentação foram testados, para ambas as técnicas (CAM e Grad-CAM), porém o que mostrou melhores resultados foi o valor de 20%.

As imagens foram organizadas com base nas classes prescritas pelo próprio conjunto de dados. Durante a execução da avaliação, surgiu um padrão notável: uma tendência discernível para valores mais elevados de MIoU tornou-se evidente quando as imagens foram filtradas com base nas probabilidades de saída do CheXnet pertencentes à classe designada. Em essência, foi observada uma propensão em que o desempenho do MIoU mostrou melhoria à medida que o processo de filtragem selecionava imagens onde o CheXnet exibia um maior grau de "certeza"em suas classificações - casos em que a probabilidade resultante era maior.

Esta observação apontou para uma relação entre o desempenho da técnica de interpretabilidade e os níveis de confiança preditiva do modelo. Imagens onde o CheXnet demonstrou maior confiança preditiva pareciam exibir uma maior capacidade de localização por meio das técnicas de interpretabilidade estudadas. Este fenômeno sugeriu uma possível ligação entre as previsões robustas do modelo e a subsequente eficácia das metodologias de localização utilizadas.

Com o propósito de resumir essas conclusões, foi compilada a tabela 5.2. A tabela 5.2 delineia as pontuações MIoU para imagens classificadas em classes distintas indicadas pelo

conjunto de dados. Os valores MIoU são estratificados em diferentes graus de confiança de saída da rede, para demonstrar o aumento da performance de localização com o aumento da certeza de classificação do modelo. Esse valor foi denominado Grau de confiança de saída da rede (GCSR). Esses níveis de GCSR abrangem faixas de probabilidade incluindo $> 0\%, > 50\%, > 60\%, > 70\%, > 80\% e > 90\%$.

| Técnica | GCSR | Atelectasia | Cardiomegalia | Efusão | Infiltração | Massa | Nódulo | Pneumonia | Pneumotórax |
|---|---|---|---|---|---|---|---|---|---|
| CAM | > 0% | 0.09 | **0.61** | 0.12 | **0.13** | 0.12 | 0.03 | **0.17** | 0.08 |
| | > 50% | 0.10 | **0.63** | 0.12 | **0.14** | 0.13 | 0.03 | **0.18** | 0.09 |
| | > 60% | 0.10 | **0.63** | 0.12 | **0.15** | 0.14 | 0.02 | **0.20** | 0.09 |
| | > 70% | 0.11 | **0.64** | 0.12 | **0.18** | 0.15 | 0.03 | **0.23** | 0.09 |
| | > 80% | 0.11 | **0.65** | 0.12 | **0.22** | 0.15 | 0.03 | **0.28** | 0.09 |
| | > 90% | 0.12 | **0.65** | 0.13 | **0.45** | 0.17 | 0.02 | **0.30** | 0.10 |
| Grad-Cam | > 0% | 0.10 | **0.64** | 0.13 | **0.14** | 0.12 | 0.02 | **0.18** | 0.08 |
| | > 50% | 0.10 | **0.66** | 0.13 | **0.15** | 0.14 | 0.03 | **0.20** | 0.09 |
| | > 60% | 0.11 | **0.66** | 0.13 | **0.16** | 0.14 | 0.03 | **0.23** | 0.09 |
| | > 70% | 0.12 | **0.67** | 0.13 | **0.19** | 0.15 | 0.03 | **0.26** | 0.09 |
| | > 80% | 0.12 | **0.67** | 0.13 | **0.24** | 0.16 | 0.03 | **0.30** | 0.09 |
| | > 90% | 0.13 | **0.67** | 0.14 | **0.53** | 0.18 | 0.03 | **0.37** | 0.10 |

**Tabela 5.2** Tabela de comparação das MIoU do CAM e Grad-cam, agrupadas pelas patologias para cada GCSR (Grau de confiança de saída da rede). As colunas destacadas representam as classes que obtiveram os melhores resultados em termos de melhoria na comparação Grad-CAM vs CAM.

A partir da tabela 5.2 foi observada uma melhoria percentual média de 7.5% na aplicação do Grad-CAM em comparação com o CAM. Esse valor é calculado a partir da melhoria percentual de cada classe, que é calculada através da equação: $\frac{MGC-MC}{MC} * 100\%$. Sendo MGC o resultado do MIoU na aplicação do Grad-CAM e o MC o resultado do MIoU na aplicação do CAM.

## 5.3 Aplicação

### 5.3.1 Arquitetura

A arquitetura do sistema inteligente projetado para o diagnóstico e identificação de regiões de interesse nas radiografias de tórax é representada na figura 5.2. Este sistema opera através de uma interface web, acessível através de dispositivos móveis ou computadores padrão. A interface do usuário possui dois campos de entrada distintos, cada um servindo a uma finalidade específica. O primeiro campo de entrada é designado para a entrada de um código de acesso ao sistema, uma sequência alfanumérica de 16 caracteres. Este código é fornecido exclusivamente pelos administradores do sistema, garantindo o acesso controlado às funcionalidades. O segundo campo de entrada serve como um portal para os usuários enviarem suas imagens de radiografia de tórax para análise. Após o envio do formulário, o servidor processa a solicitação, iniciando uma série de tarefas. Em primeiro lugar, valida o código de acesso introduzido, garantindo que apenas pessoal autorizado possa interagir com o sistema. Posteriormente, o servidor realiza a tarefa de processamento de imagens. A radiografia de tórax submetida passa por uma análise,

culminando na geração de resultados diagnósticos potencialmente múltiplos, utilizando a combinação de técnicas estudas neste trabalho, o CheXnet e o Grad-CAM. Além disso o sistema possui uma interface de administração para o gerenciamento das chaves de acesso entregues aos usuário, assim como o gerenciamento dos dados de envio e resposta das requisições (imagens enviadas e seus resultados).

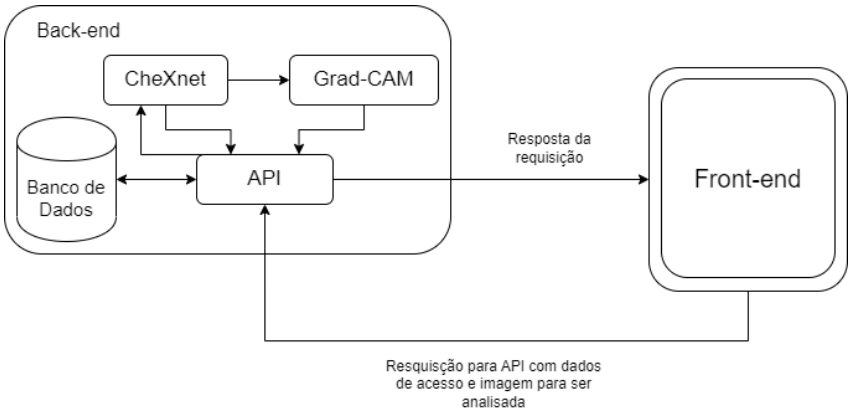


**Figura 5.2** Arquitetura da aplicação.

Para cada possibilidade de diagnóstico, o sistema produz um mapa de calor. Este mapa de calor destaca as áreas específicas de interesse na imagem da radiografia de tórax, auxiliando os profissionais de saúde a concentrarem sua atenção em regiões críticas. Além disso, o sistema fornece uma pontuação de confiança baseada em porcentagem, oferecendo *insights* sobre o nível de certeza do sistema em relação a cada diagnóstico.

O sistema foi elaborado utilizando um conjunto de tecnologias. O componente *front-end* foi desenvolvido utilizando o *framework* Vue.js utilizando JavaScript como linguagem de programação principal. No *back-end*, o sistema conta com Django, uma estrutura robusta de desenvolvimento web, e um complemento de pacotes de desenvolvimento web dentro do ecossistema Django. Integrado com o *back-end* encontram-se os algoritmos inteligentes desenvolvidos que foram construídos utilizando bibliotecas de python para desenvolvimento de IA, como pandas, numpy, Keras, TensorFlow e scikit-learn.

### 5.3.2 *Back-end*

#### 5.3.2.1 API

A autenticação do sistema é feita automaticamente pela biblioteca do Django *"rest framework api key"* onde a autorização do usuário é feita através do método de chave de API (*Application Programming Interface*), a qual a chave entregue para o usuário é uma chave de API. Os *endpoins* da API se resumem a dois: /admin e /*inference*. O *endpoint* /admin é responsável por lidar com as requisições referentes à página de administração, a qual é construída automaticamente pelas ferramentas disponibilizadas pelo Django. O *endpoint* de /*inference* é o responsável por lidar, através do método HTTP POST, com os pedidos de suporte em diagnóstico para uma dada imagem de entrada enviada pelo

usuário. Através dos campos *key* e originalImage a requisição é processada. O campo *key* representa a chave de acesso do usuário e o originalImage representa o arquivo binário que armazena a imagem a ser analizada.

#### 5.3.2.2 Diagrama de classes

O diagrama de classes da camada de serviço da aplicação encontra-se na figura 5.3. A classe Resultados representa o resultado criado e armazenado de cada requisição. Ela possui os seguintes atributos:

- *key* - Do tipo *Key* (será detalhado posteriormente) e representa a chave de acesso que requisitou o resultado;
- originalImage - Do tipo Imagem e representa a imagem a ser analisada para a geração dos diagnósticos;
- heatmaplinks - Do tipo *array* de HeatmapLink (será detalhado posteriormente) e representa as referências com cada mapa de calor gerados pelos algoritmos inteligentes;
- created_at - Do tipo Data e representa a data e horário em que o resultado foi gerado.

A classe *Key* representa as chaves dos usuário criadas pelo administrador. Ela possui os seguintes atributos:

- *key* - *string* que representa os 16 caracteres alfanuméricos. Essa *string* é armazenada no banco criptografada e só é disponibilizada no momento de sua criação;
- *counter* - número inteiro que representa a quantidade de vezes que uma requisição de resultado foi bem sucedida a partir da chave. O objetivo disso é limitar o número máximo de requisições por chave;
- created_at - Do tipo Data e representa a data e horário em que a chave foi criada.

A classe HeatmapLink representa as referências em que um resultado tem com todos os mapas de calor resultantes de sua inferência. Seus atributos são:

- *pathology* - *string* que representa a classe pertencente ao mapa de calor;
- *link* - *string* que representa o diretório da imagem no banco de dados onde mapa de calor está armazenado.

#### 5.3.2.3 Máquina de estados

Caso a autorização do usuário tenha sido bem sucedida, o sistema segue a máquina de estados apresentada na figura 5.4. Inicialmente é criado um objeto que representa os resultados da inferência a partir da classe Resultado. Após isso é feita a inferência baseada na imagem enviada pelo usuário através da CheXnet e, também, são produzidos os mapas de calor a partir dos mapas de ativação da CheXNet através do Grad-CAM.

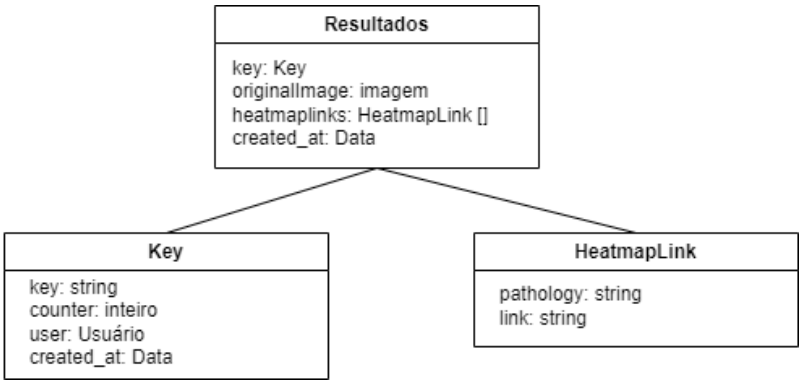


**Figura 5.3** Diagrama de classes da camada de serviço da aplicação.

São criados e armazenados os links de mapas de calor a partir da classe HeatmapLink e assim os resultados são agrupados e ordenados no objeto resultado. O resultado é então armazenado no banco de dados e a resposta da requisição é enviada com os resultados dos algoritmos inteligentes. Caso algum erro ocorra durante esse processo a máquina de estado é finalizada e uma mensagem de erro é enviada como resposta da requisição.

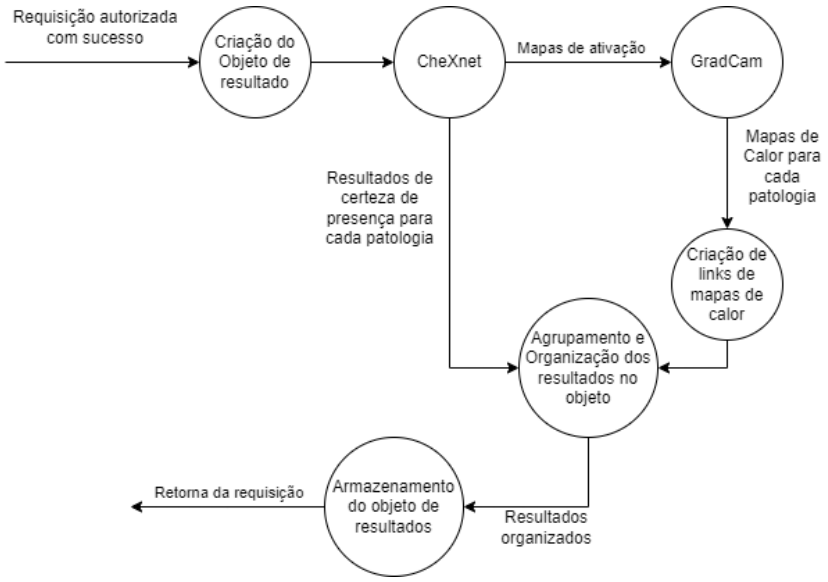


**Figura 5.4** Máqiuna de estados do *back-end*.

### 5.3.3 *Front-end*

Representações visuais do sistema podem ser encontradas nas figuras 5.5 e 5.6, oferecendo uma visão mais abrangente de sua interface de usuário e funcionalidades. É essencial observar que os únicos dados retidos pelo sistema compreendem a imagem radiográfica enviada, juntamente com metadados de data/hora indicando a data e hora do envio. Além disso, o acesso ao sistema é controlado por meio de chaves de acesso únicas, que

são geradas aleatoriamente e disponibilizadas aos interessados em testar o sistema. Para facilitar o gerenciamento de usuários e a integridade do sistema, a geração de chaves de acesso envolve a captura do nome e informações de contato do usuário.


**Figura 5.5** Tela inicial da aplicação.

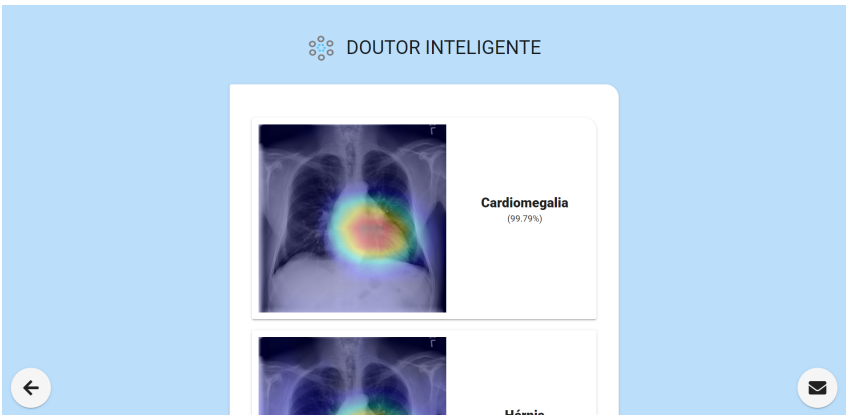


**Figura 5.6** Tela de resultados da aplicação.

Capítulo 6

# Conclusão e trabalhos futuros

As descobertas feitas neste estudo fundamentam a hipótese inicial, que afirma que a abordagem de visualização do Grad-Cam pode produzir mapas de calor mais precisos. Isso revela melhorias discerníveis alcançadas pelo Grad-CAM em relação ao CAM em várias classes médicas, através da comparação da MIoU como métrica de avaliação das duas técnicas. Como é visto na seção 5.2, a melhoria média percentual apresentou o valor aproximado de 7.5%.

Apesar disso, a avaliação também ressalta uma grande limitação dos algoritmos de localização não supervisionados, que, embora sejam adeptos de destacar áreas potencialmente influentes para a classificação, ficam aquém em termos de identificação de locais precisos devido à ausência de treinamento explícito para esta tarefa. Consequentemente, um modelo de localização supervisionada surge como a escolha preferida para localização precisa de patologias em imagens. No entanto, a escassez de dados rotulados adaptados para localização supervisionada representa um desafio, pois a geração de tais anotações exige não apenas o diagnóstico, mas também a localização precisa de patologias, tornando o processo intensivo em recursos.

Uma observação fundamental obtida da análise diz respeito à correlação entre o desempenho da técnica de interpretabilidade e a confiança preditiva do modelo. O aumento da certeza do modelo na classificação de uma imagem correlaciona-se com uma capacidade de localização mais elevada através das técnicas de interpretabilidade.

## 6.1 Trabalhos futuros

No contexto desse estudo, técnicas de localização não supervisionadas encontram utilidade potencial na automação da anotação de imagens radiográficas. Isto implica aproveitar a correlação observada entre a certeza do modelo e a precisão da localização para identificar automaticamente regiões de interesse para fins de anotação. Essas anotações automatizadas poderiam contribuir potencialmente para ampliar o conjunto de dados disponíveis para o treinamento de modelos de localização supervisionados, potencialmente avançando no desenvolvimento de ferramentas de diagnóstico mais precisas e confiáveis.

Além disto, o desenvolvimento da aplicação representa um passo crucial para completar a lacuna entre a investigação e a utilidade prática. Ao encapsular as funcionalidades das técnicas de interpretabilidade, como o Grad-CAM, numa ferramenta de software de fácil utilização. Através de testes juntamente com profissionais da saúde, a aplicação poderá ser usada para o estudo da avaliação de CDSS do ponto de vista da usabilidade

dos usuários, e não somente do ponto de vista analítico, visando reproduzir testes para a melhoria das capacidades interpretativas da ferramenta. Esta aplicação tem o potencial de melhorar a interpretabilidade dos diagnósticos baseados em IA para radiografias de tórax, contribuindo, em última análise, para decisões clínicas mais informadas e melhor atendimento ao paciente em ambiente real. Em essência, esta aplicação de software serve como um canal para traduzir a excelência da investigação em benefícios tangíveis para a comunidade médica, reforçando a integração de técnicas avançadas de IA na prática clínica de rotina.

# Referências Bibliográficas


[1] Miltiadis Allamanis, Earl T. Barr, Premkumar Devanbu, and Charles Sutton. A survey of machine learning for big code and naturalness, 2018.

[2] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 1721–1730, New York, NY, USA, 2015. Association for Computing Machinery.

[3] Gwang Hyeon Choi, Jihye Yun, and Jonggi Choi. Development of machine learning-based clinical decision support system for hepatocellular carcinoma. September 2020.

[4] Langlotz CP. Will artificial intelligence replace radiologists? 2018.

[5] Community Preventive Services Task Force CPSTF. The guide to community preventive services. cardiovascular disease: Clinical decision-support systems (cdss). 2013.

[6] 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, 2009.

[7] Chartrand G, Cheng PM, and Vorontsov E. Deep learning: A primer for radiologists. 2018.

[8] Rubin GD. Artificial intelligence in medical imaging: Challenges and opportunities. 2018.

[9] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[10] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018.

[11] Jeremy Irvin, Pranav Rajpurkar, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. January 2019.

[12] Kawamoto K, Houlihan CA, Balas EA, and Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. 2005.

[13] Ashkan Khakzar, Shadi Albarqouni, and Nassir Navab. Learning interpretable features via adversarially robust optimization. August 2019.

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. January 2017.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[16] Zachary C. Lipton. The mythos of model interpretability, 2017.

[17] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. https://distill.pub/2017/feature-visualization.

[18] Pranav Rajpurkar, Jeremy Irvin, and Kaylie Zhu. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. December 2017.

[19] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning, 2016.

[20] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016.

[21] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, and Dhruv Batra Devi Parikh. Grad-cam: Visual explanations from deep networks via gradient-based localization. December 2019.

[22] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

[23] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald Summers. Chestx-ray14: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 3462–3471, 2017.

[24] Quanshi Zhang and Song-Chun Zhu. Visual interpretability for deep learning: a survey, 2018.

[25] Bolei Zhou, Aditya Khosla, and Agata Lapedriza. Learning deep features for discriminative localization. December 2015.