

RAG Multimodal

Maria Fernanda Bosco e Gabriel Carvalho Freitas

27 de novembro de 2025

Abstract

Este trabalho apresenta o desenvolvimento de um sistema de recuperação de casos clínicos (retrieval) baseado em Radiografias de Tórax (CXR) e laudos médicos, utilizando representações multimodais extraídas pelo modelo pré-treinado MedSigLip. O método inclui pré-processamento textual e visual, extração e normalização de embeddings para cada modalidade, combinação multimodal por média ponderada e construção de vector stores independentes com FAISS. O sistema permite consultas por texto, imagem ou ambas, retornando casos similares do dataset. A avaliação, baseada em critérios de relevância multirrotulo e métricas de precisão e similaridade, demonstra que a representação multimodal melhora a recuperação de estudos clinicamente relacionados em comparação aos sistemas unimodais.

1 Introdução

A radiografia de tórax é o exame de imagem mais comum em todo o mundo, sendo fundamental para o diagnóstico e tratamento adequado de muitas doenças potencialmente fatais. Porém, o percentual de radiologistas está diminuindo ao redor do mundo, e o cenário é ainda pior em países subdesenvolvidos. Por isso, a interpretação automatizada de radiografias de tórax, pode trazer benefícios em diversos contextos médicos.

O objetivo do projeto é a implementação um sistema de RAG indexado de forma multimodal a partir de imagens de CXR e laudos médicos, obtidos através do dataset MIMIC-CXR, e usando um modelo pré-treinado. Dessa forma, dado um *input* de entrada (texto ou imagem) o sistema pode recomendar casos similares, auxiliando no diagnóstico.

2 Dataset

O dataset utilizado no projeto, MIMIC-CXR, foi obtido através do exame de radiografia do tórax de 65.379 pacientes diferentes. Cada paciente pode estar associado a mais de um estudo, totalizando 227.835 estudos totais. Um estudo é composto por um laudo textual, escrito por um médico especialista, e uma ou

mais imagens de radiografia do tórax (geralmente uma frontal e outra lateral), totalizando 377.110 imagens.

3 Metodologia

O dataset utilizado neste estudo foi submetido a um processo estruturado de pré-processamento, visando reduzir inconsistências e assegurar a qualidade necessária para a extração confiável de embeddings multimodais. As etapas metodológicas contemplaram o tratamento dos laudos médicos, a preparação das imagens de radiografia, a geração de embeddings por meio do modelo MedSigLip e a construção de diferentes mecanismos de recuperação baseados em FAISS.

3.1 Pré-processamento dos laudos médicos

Os laudos textuais foram inicialmente submetidos a uma etapa de limpeza e normalização. Para garantir que apenas o conteúdo clínico mais relevante fosse utilizado durante a extração dos embeddings, foi desenvolvida uma função dedicada à identificação e extração da seção findings, reconhecida como a porção do laudo na qual se concentram os achados diagnósticos pertinentes ao exame.

Essa etapa mostrou-se essencial devido à limitação do modelo MedSigLip, que admite no máximo 64 tokens como entrada textual, exigindo assim a seleção criteriosa das informações clínicas fundamentais. Após a extração da seção relevante, cada texto foi processado pelo do modelo, que foi normalizado por norma L2.

3.2 Pré-processamento das imagens

As imagens de radiografia de tórax associadas a cada estudo foram processadas individualmente pelo MedSigLip, gerando um embedding por imagem. Nos casos em que um estudo incluía múltiplas imagens, adotou-se um procedimento de agregação baseado em pooling (média) sobre os embeddings individuais, de modo a produzir uma única representação vetorial por estudo. O embedding resultante passou igualmente por normalização L2.

3.3 Combinação multimodal de embeddings

Com o objetivo de produzir uma representação unificada capaz de capturar simultaneamente informações textuais e visuais, os embeddings de texto e imagem foram combinados por meio de uma média ponderada, formalizada como:

$$final_emb = \alpha \cdot text_emb + (1 - \alpha) \cdot image_emb \quad (1)$$

Diferentes valores de α foram avaliados empiricamente, considerando o desempenho obtido nas métricas de recuperação. Os experimentos indicaram que o melhor equilíbrio entre modalidades foi alcançado com $\alpha = 0.625$ (gráfico 3).

O embedding multimodal final foi normalizado para garantir compatibilidade com os métodos de busca baseados em similaridade coseno.

3.4 Construção da vector store e sistema RAG

Com os embeddings finais disponíveis, foram construídas três vector stores independentes utilizando a biblioteca FAISS, cada uma correspondente a uma modalidade de representação:

- Textual — composta exclusivamente pelos embeddings derivados dos laudos;
- Visual — contendo apenas os embeddings das imagens;
- Multimodal — formada pelos embeddings resultantes da combinação texto–imagem.

Cada vector store originou um sistema de recuperação distinto, permitindo consultas conforme a natureza da entrada fornecida pelo usuário: texto, imagem ou entrada multimodal.

Essa estrutura possibilitou a avaliação comparativa das diferentes modalidades de consulta e permitiu investigar em que medida a combinação das informações visuais e textuais contribui para a recuperação de casos clinicamente similares.

3.5 Avaliação

Para avaliar o desempenho do sistema de recuperação, foram utilizadas as anotações de múltiplos rótulos clínicos presentes no dataset. Cada estudo possuía um conjunto de labels correspondentes às seguintes condições:

- Atelectasis
- Cardiomegaly
- Consolidation
- Edema
- Enlarged Cardiomedastinum
- Fracture
- Lung Lesion
- Lung Opacity
- No Finding
- Pleural Effusion
- Pleural Other

- Pneumonia
- Pneumothorax
- Support Devices

Cada estudo podia receber múltiplos labels simultaneamente, caracterizando um problema de classificação multirrótulo. Assim, a relevância entre um estudo de consulta e os estudos retornados pelo sistema foi definida com base no grau de interseção entre seus conjuntos de labels. Foram considerados dois critérios distintos de relevância:

- **ANY**: um estudo retornado é considerado relevante caso compartilhe pelo menos um label com o estudo de referência.
- **ALL**: um estudo retornado é considerado relevante apenas se apresentar exatamente o mesmo conjunto de labels do estudo de referência.

Esses critérios permitem avaliar tanto a capacidade do sistema de recuperar casos parcialmente semelhantes (ANY) quanto a sua precisão em identificar estudos clinicamente equivalentes (ALL).

Com base nessas definições, foram adotadas três métricas de avaliação:

- **PRECISION_ANY@K**: proporção de estudos relevantes segundo o critério ANY entre os K primeiros resultados retornados.
- **PRECISION_ALL@K**: proporção de estudos relevantes segundo o critério ALL entre os K primeiros resultados.
- **JACCARD@K**: média do coeficiente de Jaccard entre o estudo de referência e cada um dos K resultados, avaliando a similaridade entre os conjuntos de labels independentemente da definição de relevância. O coeficiente de Jaccard é definido como a intersecção entre os dois conjuntos (casos relevantes e casos retornados) dividida pela união entre os mesmos.

Essas métricas combinam perspectivas complementares — precisão estrita, precisão parcial e similaridade estrutural entre conjuntos de labels — permitindo uma análise abrangente do desempenho dos sistemas de retrieval unimodais e multimodais.

4 Resultados

Os 227.835 estudos foram submetidos aos três sistemas de recuperação descritos na seção 3.4, buscando sempre os cinco estudos com maior similaridade de cosseno ($K = 5$), de acordo com qual era o embedding de interesse: texto, imagem ou estudo completo. Sendo assim, temos um dataset final com um total de 1.139.175 retornos, 5 para cada estudo de referência.

4.1 Métricas de retrieval

Avaliamos a performance dos retornos para cada estudo a partir das métricas descritas na seção 3.5. O gráfico 1 apresenta a média de cada uma das métricas para o dataset completo.

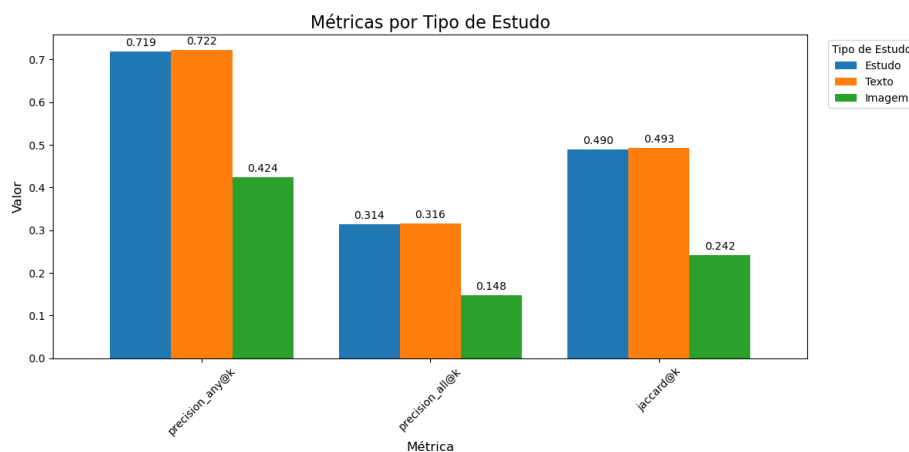


Figure 1: Performance dos diferentes sistemas de busca para o dataset completo.

A análise dos resultados indica que a modalidade textual apresenta o melhor desempenho entre os três sistemas avaliados. Isso ocorre porque os laudos médicos, especialmente após a extração da seção findings, concentram de forma explícita as informações clínicas relevantes, o que permite ao modelo captar de maneira direta os achados associados a cada estudo. Como as métricas utilizadas se baseiam em interseções entre conjuntos de labels, o embedding textual tende a refletir com maior precisão a presença ou ausência de condições clínicas específicas. Assim, o sistema de recuperação baseado apenas em texto demonstra uma eficiência maior para identificar estudos com similaridade clínica, apresentando os maiores valores nas métricas de interesse.

A modalidade visual apresenta desempenho significativamente inferior. Os embeddings derivados exclusivamente das imagens capturam padrões visuais, mas muitos dos achados clínicos do dataset são sutis e dependem de interpretação contextual de um médico especialista. Dessa forma, o sistema puramente visual se mostra menos eficiente para recuperar estudos clinicamente similares, ainda que apresente alguma capacidade de identificação, com uma média de 14% dos retornos tendo todos os labels iguais ao estudo de referência.

O sistema multimodal, construído pela combinação ponderada entre embeddings textuais e visuais, apresenta um desempenho similar a busca estritamente textual. Embora não alcance o desempenho máximo observado na modalidade textual, o modelo multimodal supera com folga a abordagem baseada apenas em imagens. Isso demonstra que a fusão de modalidades é capaz de incorporar sinais visuais complementares sem sacrificar de forma substancial a informação

clínica contida no texto. O valor de $\alpha = 0.625$ foi capaz de preservar a estrutura semântica textual dominante, ao mesmo tempo em que adiciona elementos visuais relevantes. Porém, como a modalidade de busca via imagem apresenta desempenho consideravelmente inferior de forma isolada, a agregação dos embeddings acaba por reduzir levemente a precisão alcançada quando se utiliza somente texto. Ainda assim, o sistema multimodal tende a recuperar estudos que não seriam identificados por texto sozinho, o que sugere maior robustez em cenários nos quais laudos estejam incompletos ou ambíguos, ou quando a consulta seja feita a partir de uma entrada visual. Assim, a abordagem multimodal funciona como um compromisso entre preservação da precisão textual e incorporação da variabilidade visual, oferecendo um modelo mais flexível para aplicações reais de recuperação em ambientes clínicos.

Para complementar a análise de desempenho agregado, analisamos cada estudo individualmente, comparando qual modalidade, textual ou multimodal, recupera a maior quantidade de casos clinicamente relevantes segundo o critério ALL. Entre os 227.835 estudos avaliados, o sistema multimodal retornou mais casos relevantes do que o sistema textual em 23.805 casos, o que corresponde a aproximadamente 10,4% do total. Por outro lado, o sistema baseado exclusivamente em texto apresentou melhor desempenho em 25.955 estudos, representando cerca de 11,4% da base.

Além disso, dentre os 23.805 casos em que o multimodal obteve mais retornos relevantes, identificamos que 7.696 deles apresentaram zero resultados relevantes quando a busca foi realizada apenas com embeddings textuais. Isso indica que a inclusão da informação visual permite recuperar casos que seriam completamente perdidos por uma abordagem puramente textual. Esse resultado reforça a contribuição complementar dos embeddings de imagem, que, embora apresente desempenho inferior de forma combinada, adiciona informação útil que ampliam a cobertura da busca em cenários específicos.

4.2 Métricas de classificação

Além da avaliação baseada em métricas de recuperação, realizamos um teste adicional no qual o sistema foi analisado como se fosse um classificador multirrótulo. Dessa maneira, cada estudo de referência é tratado como “rótulo verdadeiro”, e os estudos retornados pelo sistema são interpretados como as predições do modelo. Considerando que cada consulta retorna cinco resultados, o dataset final para essa análise contém os 1.139.175 milhão de retornos como amostra. A partir desse conjunto, calculamos o F1-score por classe para os três sistemas: multimodal (estudo), texto e imagem.

Os resultados apresentados no gráfico 2 evidenciam um padrão robusto e consistente com o que vimos nas métricas de busca: o modelo textual apresenta o melhor F1-score na grande maioria das classes, seguido pelo modelo multimodal, enquanto o modelo puramente visual apresenta desempenho substancialmente inferior.

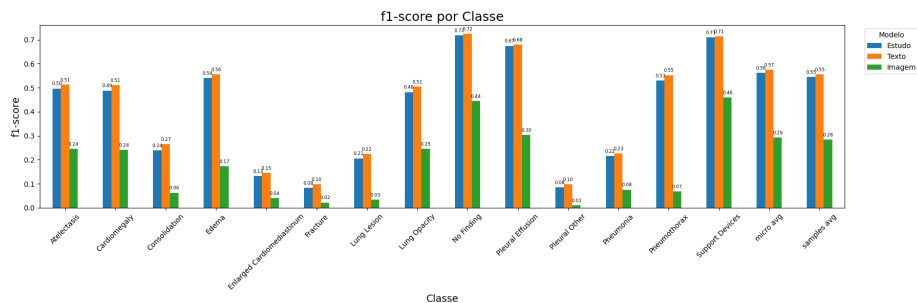


Figure 2: f1-score para cada um dos labels do dataset, considerando a busca como um classificador multi-rótulo.

5 Conclusão

Os resultados analisados demonstram que os embeddings textuais são, de forma consistente, a modalidade mais precisa para recuperar estudos clinicamente semelhantes, refletindo a forte relação entre o conteúdo dos laudos e os rótulos utilizados como referência. Embora o desempenho da modalidade visual seja inferior, ela acrescenta informações complementares que permitem ao sistema multimodal recuperar casos que seriam totalmente perdidos por uma abordagem baseada apenas em texto. Esse efeito aparece tanto nas métricas de retrieval quanto na análise de F1-score tratada como classificação.

Assim, apesar de o texto apresentar o melhor desempenho médio, o modelo multimodal mostra um potencial de robustez e abrangência, recuperando um conjunto maior de estudos relevantes em cenários nos quais as informações textuais são insuficientes. Em resumo, a combinação de embeddings não supera o texto em precisão, mas oferece uma complementaridade, se mostrando como uma alternativa equilibrada para aplicações reais de recuperação de estudos clínicos.

6 Trabalhos Futuros

Os resultados obtidos indicam potencial na metodologia multimodal e abrem possibilidades de evolução do sistema, tanto no nível dos embeddings quanto no desenho do mecanismo de recuperação. Uma primeira ideia de exploração futura seria investigar novas estratégias de combinação entre embeddings textuais e visuais, substituindo a média ponderada por outros métodos, como fusão não linear, projeções aprendidas ou módulos de atenção entre modalidades. Na mesma linha, é importante explorar abordagens diferentes de embedding para imagem, uma vez que combinação as imagens também via uma média.

Outra ideia de trabalho futuro seria a aplicação de fine-tuning supervisionado para o embedding multimodal, ajustando o espaço latente de forma a aproximar estudos clinicamente semelhantes e distanciar casos irrelevantes. Estratégias

como contrastive learning ou fine-tuning orientado por classificação multirrótulo podem melhorar a representação dos embeddings de estudo.

Inicialmente, não exploramos muito abordagens de melhoria na busca. Um próximo passo seria aprimorar o mecanismo de recuperação, incorporando reranking supervisionado, por exemplo. Além disso, poderíamos ter uma quarta abordagem de busca, retornando os top-K casos similares considerando tanto o embedding de estudo quanto o embedding estritamente de texto.

References

- [1] Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. MIMIC-CXR Database (version 2.1.0). PhysioNet, 2024. RRID:SCR_007345.
- [2] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.
- [3] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *arXiv preprint*, 2023.

A Imagens Adicionais

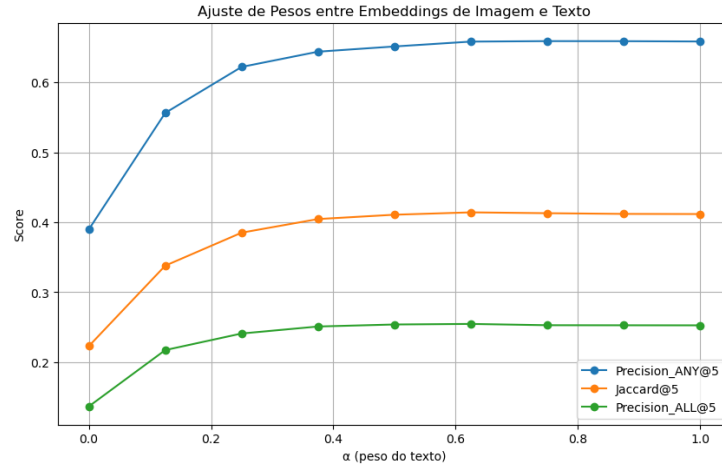


Figure 3: Avaliação de performance, em um conjunto de validação, do sistema de busca por estudo, para diferentes valores de α