

Universidade Federal de Campina Grande
Programa de Pós-graduação em Engenharia Elétrica - PPgEE

Mestrado em Engenharia Elétrica
Linha de pesquisa: MXAI

Relatório Técnico

Conjuntos de Dados Multimodais em IA Médica

Mestrando: Alysson Machado de Oliveira Barbosa
Orientador: Prof. Dr. Wamberto José Lira de Queiroz
Coorientadora: Profa. Dra. Luciana Ribeiro Veloso

Campina Grande – PB
2026

Sumário

1 Visão geral **2**

2 MIMIC: MIMIC-CXR, MIMIC-IV e MS-CXR **2**

3 CheXpert Plus **3**

4 PadChest **3**

5 ROCOv2 (*Radiology Objects in COntext*) **3**

6 VQA-RAD **4**

7 SLAKE **5**

8 TCGA (*The Cancer Genome Atlas*) **5**

9 Outras fontes para explorar conjuntos de dados **6**

Referências **6**

1 Visão geral

Este relatório reúne conjuntos de dados médicos multimodais para pesquisa em inteligência artificial. Eles combinam imagens com texto, registros clínicos ou metadados, e variam nas regras de acesso.

Tabela 1: Visão comparativa dos conjuntos selecionados.

Conjunto	**Modalidades e anotações**	**Escala aproximada**	**Acesso**
MIMIC	Radiografias, laudos, prontuário estruturado e caixas delimitadoras com frases clínicas no MS-CXR.	377 mil imagens e 1.162 pares MS-CXR.	Credenciamento no PhysioNet.
CheXpert Plus	Radiografias, laudos em texto, dados demográficos e etiquetas de patologias.	223 mil pares de imagem e laudo.	Stanford AIMI e Redivis.
PadChest	Radiografias, laudos em espanhol, dados demográficos e rótulos clínicos.	160 mil imagens de 67 mil pacientes.	Solicitação no BIMCV.
ROCOv2	Imagens radiológicas, legendas e conceitos UMLS.	79.789 imagens.	Zenodo, acesso público.
VQA-RAD	Imagens radiológicas com perguntas e respostas clínicas.	315 imagens e 3.515 pares.	OSF, acesso público.
SLAKE	Imagens, perguntas e respostas bilíngues, máscaras, caixas delimitadoras e grafo de conhecimento.	642 imagens e cerca de 14 mil pares.	Página dos autores.
TCGA	Dados clínicos e moleculares. Algumas coortes também incluem histopatologia.	Varia por projeto.	GDC, acesso aberto ou controlado.

Nem todo paciente possui dados em todas as modalidades. Antes de combinar arquivos, é importante conferir as chaves de ligação entre as entidades de dados multimodais.

2 MIMIC: MIMIC-CXR, MIMIC-IV e MS-CXR

O MIMIC-CXR reúne 377.110 radiografias de tórax correspondentes a 227.835 estudos, acompanhadas de laudos em texto. O MIMIC-IV complementa esse material com registros estruturados de internações e UTI, como exames laboratoriais, diagnósticos, medicamentos e medições clínicas. As notas clínicas em texto são distribuídas em um projeto separado, MIMIC-IV-Note. Os projetos podem ser relacionados por identificadores desidentificados, mas representam subconjuntos diferentes [1], [2].

O MS-CXR contém 1.162 pares imagem-frase com caixas delimitadoras para oito achados cardiopulmonares. As imagens pertencem ao MIMIC-CXR. Para usá-las, também é necessário obter acesso ao MIMIC-CXR [3].

O acesso é controlado e individual. Crie uma conta no PhysioNet, conclua o credenciamento e o treinamento exigido sobre pesquisa com participantes humanos e assine o acordo de uso de dados (*Data Use Agreement*, DUA) de cada projeto. A página do PhysioNet orienta sobre o curso CITI Data or Specimens Only Research.

1. Abra a página da versão desejada e aceite o DUA: MIMIC-CXR 2.1.0, MIMIC-IV 3.1 ou MS-CXR 1.1.0.

2. Na seção *Files*, após toda a etapa de conclusão do credenciamento, certificação e aceite do DUA ter sido concluída, os *datasets* ficarão disponíveis para download.

3 CheXpert Plus

CheXpert Plus reúne 223.462 pares de radiografias de tórax e laudos, distribuídos por 187.711 estudos de 64.725 pacientes. As imagens são fornecidas em DICOM e incluem metadados técnicos e demográficos. Os laudos são organizados em 11 partes e acompanham etiquetas para 14 patologias torácicas [4], [5].

O conjunto amplia o CheXpert com os textos completos dos laudos, dados demográficos e anotações clínicas. Isso permite estudar tarefas multimodais e também avaliar modelos em grupos demográficos distintos [4].

Para acessar o conjunto, abra a página do CheXpert Plus no Stanford AIMI e selecione a opção de download. O link leva ao Redivis, onde é necessário entrar ou criar uma conta e consultar os requisitos de acesso e os termos de uso. O Stanford AIMI disponibiliza seus dados sem custo para pesquisa não comercial, sujeito aos termos de cada conjunto [5], [6].

1. Na página do Stanford AIMI, clique em *Download here* para abrir o registro no Redivis.
2. Entre ou crie uma conta no Redivis e confira os requisitos de acesso e os termos aplicáveis.
3. Exporte os arquivos desejados pela plataforma. A página do conjunto descreve os formatos disponíveis e as informações incluídas.

4 PadChest

PadChest reúne mais de 160 mil radiografias de tórax de cerca de 67 mil pacientes atendidos na Espanha. As imagens vêm acompanhadas de laudos em espanhol, informações demográficas e rótulos clínicos. A base inclui seis posições radiográficas e anotações para 174 achados, 19 diagnósticos diferenciais e 104 localizações anatômicas. Parte dos laudos foi anotada manualmente por médicos e os demais foram rotulados por um método automatizado [7], [8].

O conjunto completo pode ser solicitado sem custo para pesquisa. O acesso exige o preenchimento do formulário e o aceite do acordo de uso do BIMCV. O acordo proíbe compartilhar os dados e tentar reidentificar pacientes [8], [9].

1. Acesse a página do PadChest no BIMCV e escolha a opção para solicitar o conjunto completo.
2. Preencha o formulário com seus dados e objetivo de pesquisa. Leia e aceite o acordo de uso antes de baixar os arquivos.
3. Reserve espaço em disco. A página do projeto informa que o conjunto completo ocupa cerca de 1,02 TB. Também há uma amostra menor para exploração inicial.

5 ROCOv2 (*Radiology Objects in COntext*)

ROCOv2 associa 79.789 imagens radiológicas a legendas e conceitos médicos. As imagens vieram do subconjunto de acesso aberto do PubMed Central e abrangem sete modalidades clínicas. Os conceitos UMLS também permitem tarefas de classificação multi-rótulo [10].

1. Acesse o registro do conjunto no Zenodo.
2. Baixe os arquivos de imagem dos splits necessários (`train`, `valid` e/ou `test`) e os CSVs correspondentes de legendas e conceitos. O pacote completo tem cerca de 6,4 GB.
3. Extraia as imagens e relacione os registros pelo nome do arquivo. Baixe também `cui_mapping.csv` e consulte `license_information.csv` antes de reutilizar imagens.

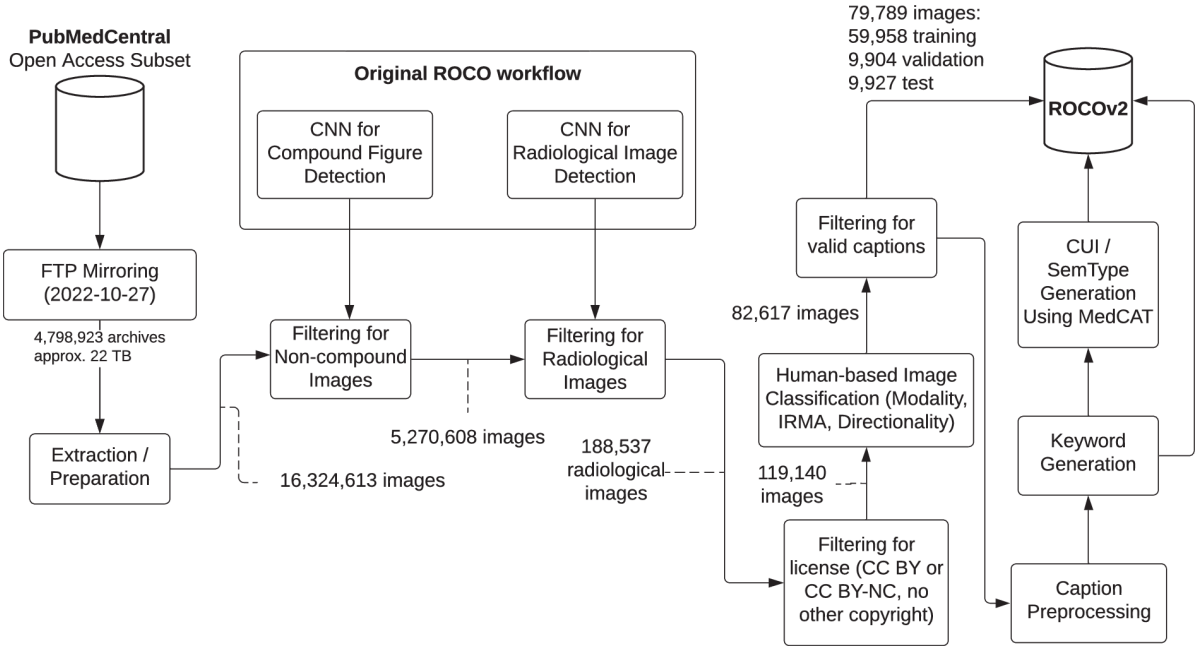


Figura 1: Etapas de seleção e organização das imagens e anotações do ROCOv2 [10].

6 VQA-RAD

VQA-RAD foi montado por profissionais da área médica para avaliar *visual question answering* em radiologia. A versão descrita no artigo reúne 315 imagens de cabeça, tórax e abdômen e 3.515 pares de pergunta e resposta, com perguntas abertas e fechadas em linguagem clínica [11].

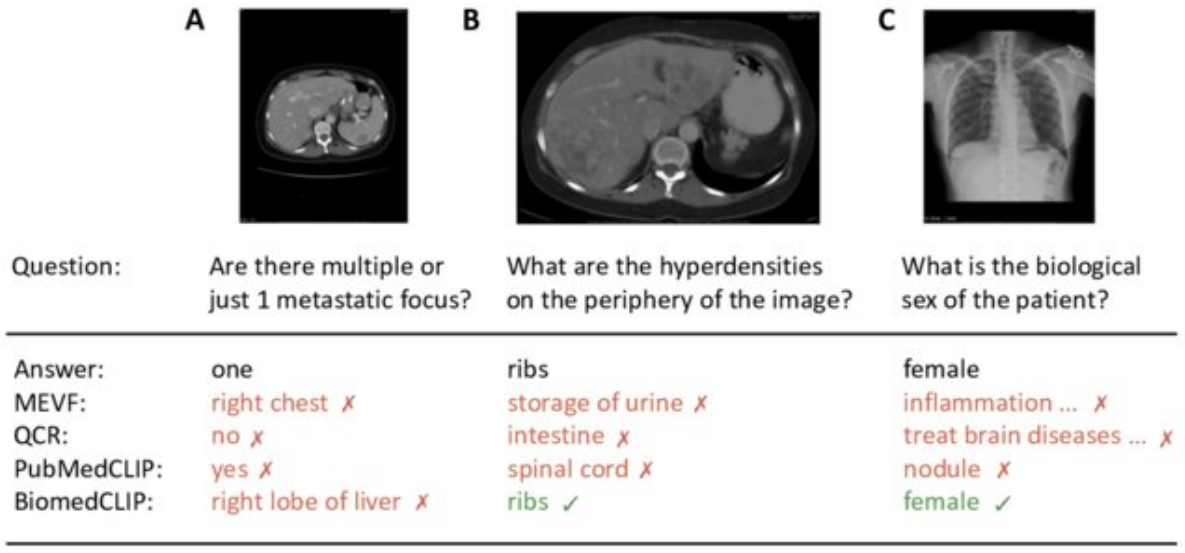


Figura 2: Exemplos de perguntas e respostas do VQA-RAD e respostas de modelos [12].

1. Abra o projeto no Open Science Framework (OSF).
2. Baixe os arquivos do projeto, incluindo a tabela de perguntas e respostas e a pasta de imagens.
3. Preserve os nomes originais das imagens, pois os registros fazem referência a eles para formar os pares pergunta e imagem.

7 SLAKE

SLAKE combina 642 imagens médicas com cerca de 14 mil perguntas e respostas em inglês e chinês. Além dos pares para *visual question answering*, inclui máscaras de segmentação, caixas delimitadoras e um grafo de conhecimento médico. As perguntas podem exigir resposta visual ou consulta ao conhecimento associado [13].

1. Acesse a página dos autores, SLAKE, e escolha um mirror: download direto, Google Drive ou Hugging Face.
2. Baixe o pacote disponível e extraia-o preservando a estrutura de pastas.
3. Consulte os arquivos de divisão e o README. A página do projeto informa que a versão limpa mais recente pode diferir ligeiramente da versão descrita no artigo.

8 TCGA (*The Cancer Genome Atlas*)

TCGA reúne projetos de diferentes tipos de câncer, não um único arquivo uniforme. Há dados clínicos e de amostras, perfis moleculares (como expressão gênica, variantes e metilação) e imagens histopatológicas. A disponibilidade muda entre projetos e pacientes. Os autores recomendam fazer o pareamento pelos identificadores de caso e amostra e confirmar nos metadados [14], [15].

Clinical Characteristics	Number of Samples
Patients with Serous Ovarian Carcinoma having Clinical Data	$N = 587$
Stage I	17 (2.90%)
Stage II	30 (5.11%)
Stage III	446 (75.98%)
Stage IV	89 (15.16 %)
Stage Not Available	5 (0.85 %)
Data Category	**Number of Attributes**
Gene expressions	6426
Copy number variants	24,776
Pathology images for each sample	1–10 images
Total pathology images for all samples.	1375 images

Figura 3: Amostra doTCGA-OV, com variáveis clínicas, dados moleculares e imagens histopatológicas [15].

1. Abra o GDC Data Portal e selecione um projeto, por exemplo, `TCGA-OV` ou `TCGA-BRCA`.
2. No repositório, filtre por categoria, tipo de dado, estratégia experimental e nível de acesso. Adicione os arquivos ao carrinho e escolha *Download Cart* → *Manifest*.
3. Para volumes grandes, instale o GDC Data Transfer Tool e execute:

```
gdc-client download -m gdc_manifest.txt
```

Arquivos controlados exigem autorização no dbGaP e um token do GDC. Após obter a autorização e baixar o token no portal, use:

```
gdc-client download -m gdc_manifest.txt -t gdc-user-token.txt
```

Arquivos abertos não exigem token. Confira a coluna *Access* antes do download e não compartilhe o token.

9 Outras fontes para explorar conjuntos de dados

Para ampliar a busca e ter uma ideia da variedade de conjuntos de dados disponíveis, as fontes abaixo reúnem desde catálogos de bases clínicas até conjuntos específicos de radiografias de tórax. Consulte cada página para verificar as modalidades, as anotações, os critérios de acesso e os termos de uso.

1. **PhysioNet**: catálogo de conjuntos de dados clínicos e biomédicos.
2. **NIH Clinical Center - ChestX-ray14**: conjunto de radiografias de tórax.
3. **Stanford AIMI - CheXpert**: conjunto de radiografias de tórax com anotações de patologias.
4. **PadChest**: conjunto de radiografias de tórax acompanhado de laudos e rótulos clínicos.

Referências

[1] A. Johnson, T. Pollard, R. Mark, S. Berkowitz e S. Horng, "MIMIC-CXR Database," *PhysioNet*, jul. de 2024, Version 2.1.0. DOI: 10.13026/4jqj-jw95 endereço: https://doi.org/10.13026/4jqj-jw95

[2] A. Johnson et al., "MIMIC-IV," *PhysioNet*, out. de 2024, Version 3.1. DOI: 10.13026/kpb9-mt58 endereço: https://doi.org/10.13026/kpb9-mt58

[3] B. Boecking et al., "MS-CXR: Making the Most of Text Semantics to Improve Biomedical Vision-Language Processing," *PhysioNet*, nov. de 2024, Version 1.1.0. DOI: 10.13026/9g2z-jg61 endereço: https://doi.org/10.13026/9g2z-jg61

[4] P. Chambon et al., *CheXpert Plus: Augmenting a Large Chest X-ray Dataset with Text Radiology Reports, Patient Demographics and Additional Image Formats*, 2024. DOI: 10.48550/arXiv.2405.19538 arXiv: 2405.19538. endereço: https://arxiv.org/abs/2405.19538

[5] Stanford Center for Artificial Intelligence in Medicine and Imaging, *CheXpert Plus*, Stanford AIMI, Accessed: Oct. 20, 2026, 2026. endereço: https://aimi.stanford.edu/datasets/chexpert-plus

[6] Stanford Center for Artificial Intelligence in Medicine and Imaging, *AIMI Shared Datasets*, Stanford AIMI, Accessed: Oct. 20, 2026, 2026. endereço: https://aimi.stanford.edu/data

[7] A. Bustos, A. Pertusa, J.-M. Salinas e M. de la Iglesia-Vayá, "PadChest: A large chest X-ray image dataset with multi-label annotated reports," *Medical Image Analysis*, v. 66, p. 101 797, 2020. DOI: 10.1016/j.media.2020.101797 endereço: https://doi.org/10.1016/j.media.2020.101797

[8] BIMCV, *PadChest*, BIMCV, Accessed: Oct. 20, 2026, 2026. endereço: https://bimcv.cipf.es/bimcv-projects/padchest/

[9] BIMCV, *PadChest Dataset Research Use Agreement*, BIMCV, Accessed: Oct. 20, 2026, 2026. endereço: https://bimcv.cipf.es/bimcv-projects/padchest/padchest-dataset-research-use-agreement/

[10] J. Rückert et al., "ROCOv2: Radiology Objects in COntext Version 2, an Updated Multimodal Image Dataset," *Scientific Data*, v. 11, p. 688, 2024. DOI: 10.1038/s41597-024-03496-6 endereço: https://doi.org/10.1038/s41597-024-03496-6

[11] J. J. Lau, S. Gayen, A. Ben Abacha e D. Demner-Fushman, "A dataset of clinically generated visual questions and answers about radiology images," *Scientific Data*, v. 5, p. 180 251, 2018. DOI: 10.1038/sdata.2018.251 endereço: https://doi.org/10.1038/sdata.2018.251

[12] S. Zhang et al., *Large-Scale Domain-Specific Pretraining for Biomedical Vision-Language Processing*, 2023. DOI: 10.48550/arXiv.2303.00915 arXiv: 2303.00915. endereço: https://arxiv.org/abs/2303.00915

[13] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang e X.-M. Wu, "Slake: A Semantically-Labeled Knowledge-Enhanced Dataset For Medical Visual Question Answering," em *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 1650–1654. DOI: 10.1109/ISBI48211.2021.9434010

[14] J. H. Hong, H. W. Cho, Y.-T. Ouh, J. K. Lee, Y. Chun e J.-A. Gim, "Genomic landscape of advanced endometrial cancer analyzed by targeted next-generation sequencing and the cancer genome atlas (TCGA) dataset," *Journal of Gynecologic Oncology*, v. 33, n. 3, e29, 2022. DOI: 10.3802/jgo.2022.33.e29 endereço: https://pmc.ncbi.nlm.nih.gov/articles/PMC9024183/

[15] R. M. Ghoniem, A. D. Algarni, B. Refky e A. A. Ewees, "Multi-Modal Evolutionary Deep Learning Model for Ovarian Cancer Diagnosis," *Symmetry*, v. 13, n. 4, p. 643, 2021. DOI: 10.3390/sym13040643 endereço: https://www.mdpi.com/2073-8994/13/4/643