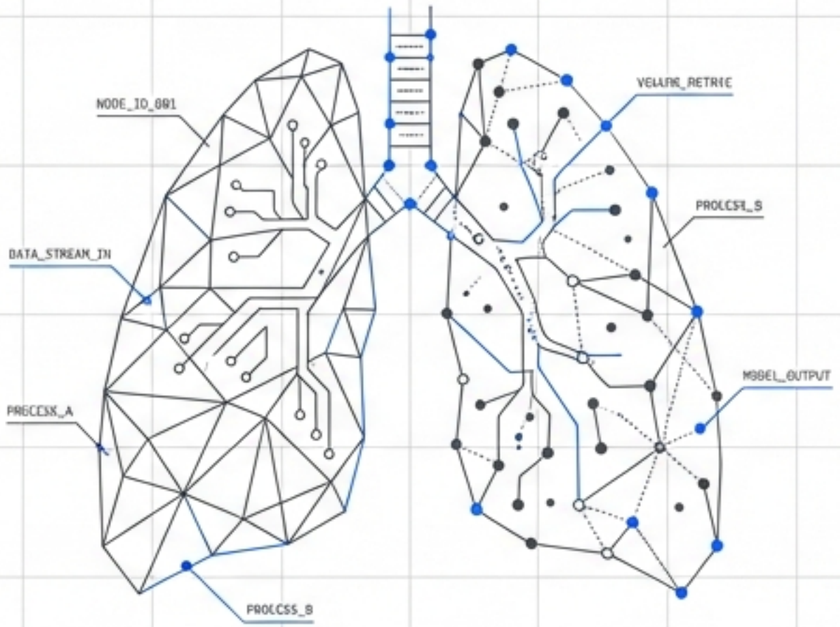
HAID

Health AI Data Resource



Curated, Preprocessed, Research-Ready Clinical Imaging

v1.0.0 | January 2026

THE MISSION

DEMOCRATIZING HIGH-QUALITY MEDICAL IMAGING

HAID bridges the gap between raw clinical data and AI model development. By providing standardized workflows and preprocessed formats, we **eliminate the bottleneck** of data wrangling, allowing researchers to focus on model architecture and clinical impact.

13

CURATED DATASETS SPANNING MULTIPLE PATHOLOGIES

Standardized Preprocessing

Uniform NIfTI formats and spacing

Documentation

Provenance and clinical context

Annotations

Detection, segmentation, classification

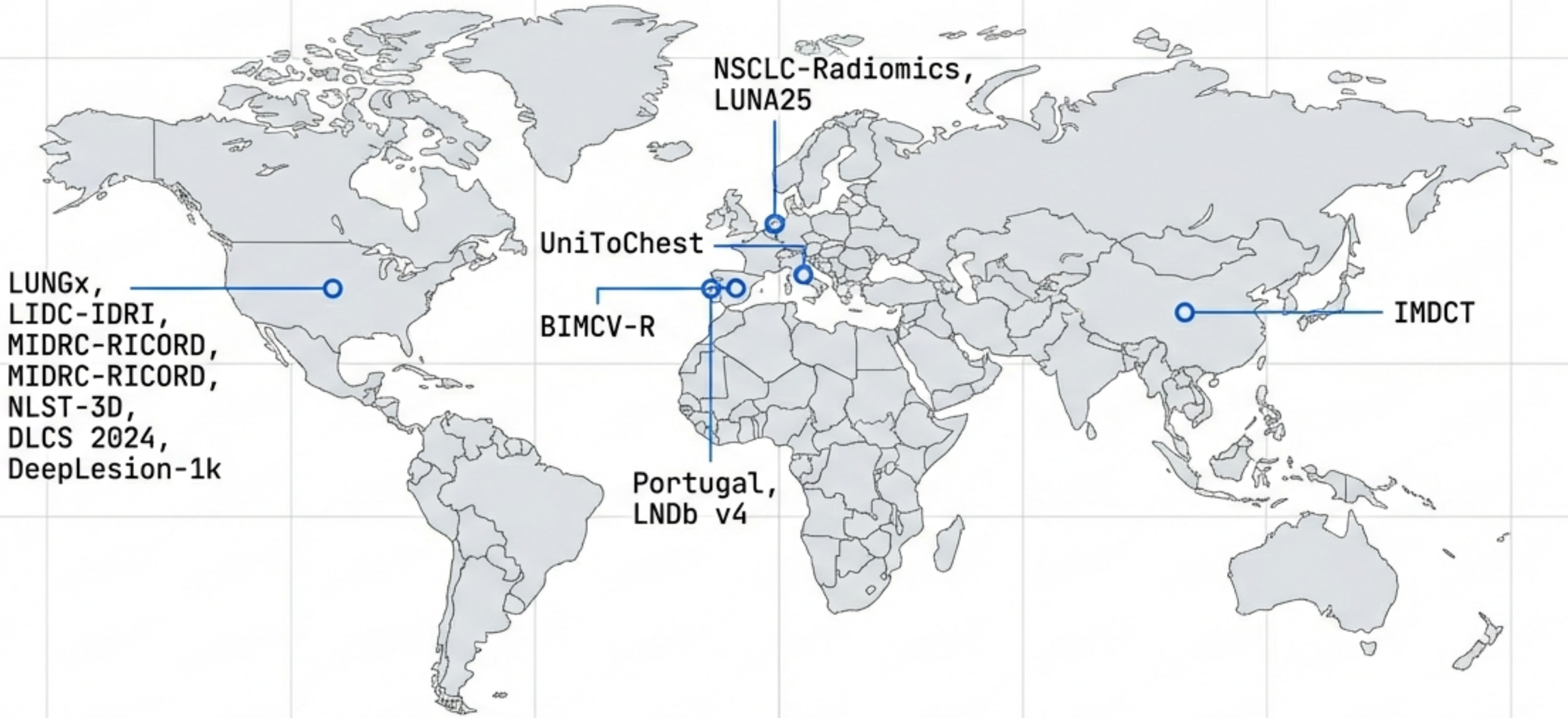
Benchmarking

Train/Validation/Test splits

Reference Notebooks

Jupyter implementations

GLOBAL DATA FOOTPRINT



Insight: Data curated from diverse patient demographics and varying scanner technologies (Philips, GE, Siemens), ensuring robust model generalization.

01 | NSCLC-Radiomics (NSCLCR) Netherlands

VITAL STATS

Patients: 421

Modality:
CT (Pre-treatment)

Condition:
Non-Small Cell Lung Cancer

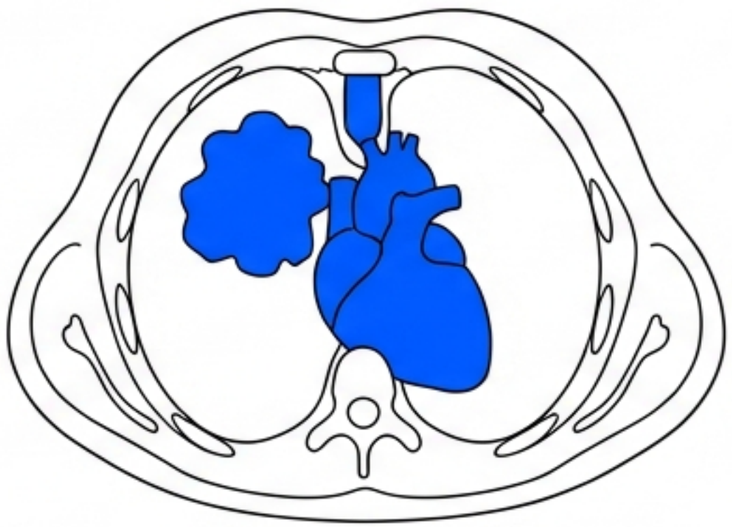
Voxel Size:
512×512×[88-176]

Slice Thickness:
3mm

KEY INNOVATION: MULTI-LABEL SEGMENTATION

This dataset goes beyond simple tumor detection. It includes precise segmentation for Organs at Risk (OAR) crucial for radiotherapy planning.

- Gross Tumor Volume (GTV-1), Esophagus, Heart, Spinal Cord, Lungs.



Includes TNM staging, histology, and survival time metadata.

02 | UniToChest

Italy

VITAL STATS

Patients:
623 (715 Scans)

Annotations:
10,071 Nodules

Condition:
Lung Nodules

Mean Diameter:
21.4 ± 20.2 mm

KEY INNOVATION: HIGH-VOLUME ANNOTATION

A massive collection of over **10,000** segmentation masks created by expert radiologists. This high volume allows for robust training of deep segmentation networks.

STANDARDIZATION NOTE

Original variable slice thicknesses have been strictly resampled to a uniform standard.

Variable Thickness

Uniform [0.7, 0.7, 1.25] mm

03 | IMDCT

China

VITAL STATS

Patients:
2,032

Malignancy Rate:
78.6%

Nodule Size:
10-30+ mm

Condition:
Indeterminate Nodules

KEY INNOVATION: THE GOLD STANDARD

Unlike datasets relying on radiologist consensus, IMDCT features 100% histopathology confirmation. Every label is backed by biopsy or surgical resection, providing ground-truth certainty.

Single target per scan for simplified analysis. Includes 3D bounding boxes.

04 | LNDb v4

Portugal

VITAL STATS

Patients:
294

Instances:
1,235 Nodules

Condition:
Pulmonary Nodules

Size Focus:
Sub-centimeter
(3-10mm)

KEY INNOVATION: RICH CLINICAL ATTRIBUTES

Beyond simple detection, this dataset offers deep texture analysis with **'Fleischner score'** compatibility. Features are rated on 1-6 scales for radiomics research.

Texture	Calcification	Sphericity
Malignancy	Subtlety	Spiculation
Lobulation	Margin	Internal Structure

Multi-reader annotations (up to 4 experts per case).

05 | BIMCV-R

Spain

VITAL STATS

Patients: 5,340

Scans: 8,069

Scale: 2M+ 2D Slices

Scale: 2M+ 2D Slices

Focus: COVID-19 & Pneumonia

KEY INNOVATION: BILINGUAL & MULTI-LABEL

A massive scale resource containing paired radiology reports in both Spanish and English. Ideal for vision-language models.

Strict split separation: No patient overlap across Train/Val/Test.

06 | LUNGx

USA

VITAL STATS

Patients: 83

Task: Classification

Classes: Benign vs. Malignant

Source: SPIE-AAPM-NCI Challenge

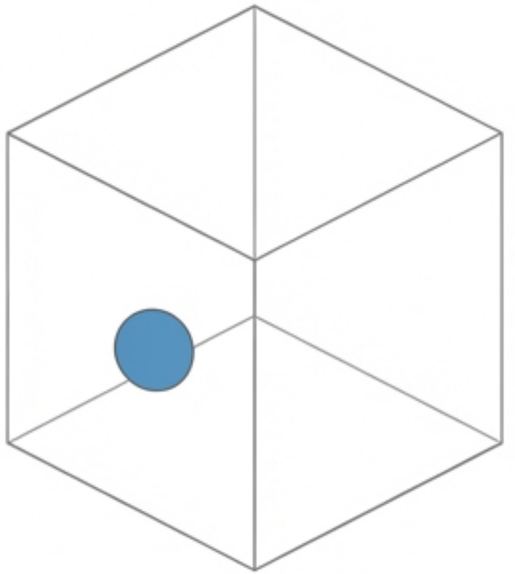
KEY INNOVATION: THE BENCHMARK

Designed specifically for the SPIE-AAPM-NCI challenge. This dataset serves as a **rigorous test bed for classification algorithms, backed by 100% histological confirmation.**

PREPROCESSING

Diagnostic patches extracted and centered.

64x64x64 voxels



Converted from DICOM pixel space to World Coordinates.

07 | LIDC-IDRI

USA

VITAL STATS

Patients: 875

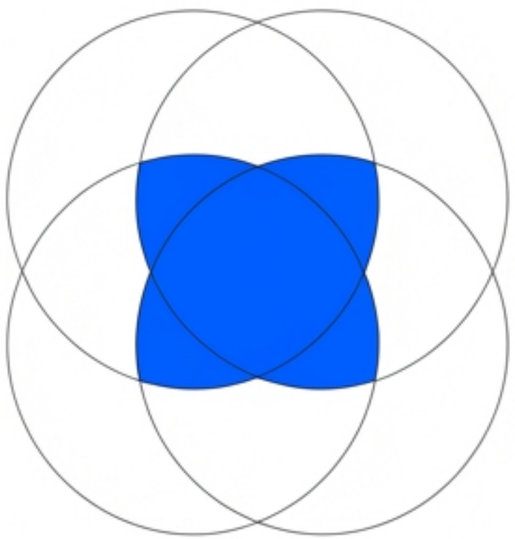
Scans: 1,010

Condition: Nodule Detection

Status: Industry Standard

KEY INNOVATION: 4-RADIOLOGIST CONSENSUS

The defining classic of medical imaging. Annotations are derived from a two-phase process (blinded + unblinded) involving four separate radiologists.



Union of Annotations = Complete Coverage

Label 23 (Nodule),
Label 200 (Body),
Vista3D Organs.

VITAL STATS

Patients: 2,020

Nodules: 6,163

Base:
Extended LUNA16

Tech:
Deep Learning Generated

KEY INNOVATION: MODERN AI EXTENSION

An evolution of the LUNA16 dataset. We used state-of-the-art foundation models to generate dense segmentation masks where only bounding boxes existed before.

PiNS Model	MONAI Vista3D
JetBrains Mono Precise Nodule Boundaries (3x3x3 morphological erosion)	JetBrains Mono Multi-organ context (Airways, Vessels, Heart)

Multi-class NIffTI files.

09 | MIDRC-RICORD

USA

VITAL STATS

Patients: 227

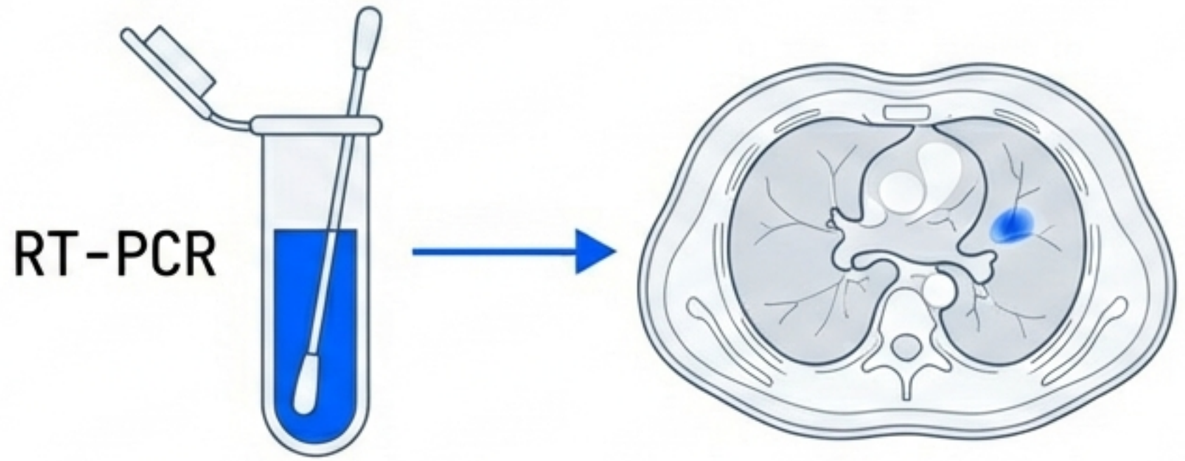
Condition:
COVID-19

Era:
2020-2021

Format:
NIfTI (from DICOM)

KEY INNOVATION: PCR-CONFIRMED GOLD STANDARD

Contains the RICORD-1A (Positive) and RICORD-1B (Negative Control) datasets. Crucially, all positive cases are confirmed via RT-PCR, ensuring high label fidelity.



Multi-institutional acquisition.
Slice Thickness: Variable (1.25mm - 3.0mm).

VITAL STATS

Volumes: 12,000+ CTs

Size: 52.5 GB

Format: TFRecords (96x160x160)

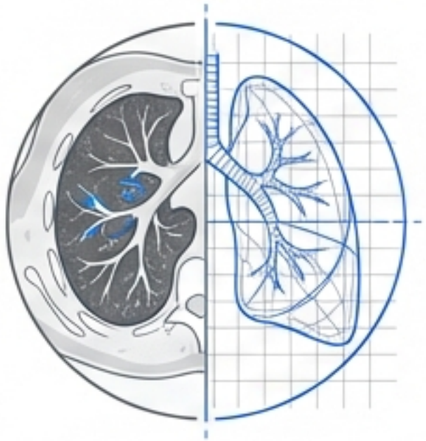
Sources: 10 Public Datasets

KEY INNOVATION: AGGREGATION & VIRTUAL TRIALS

The 'Mega-Collection'. Aggregates sources like BIMCV, LIDC, LIDC, MIDRC, and MosMed into a single unified resource. Also includes 'Virtual Imaging Trials' (VIT) data.

INCLUDES SYNTHETIC DATA

Utilizes XCAT phantoms and DukeSim framework.



Purpose: Robustness testing and AI transparency.

11 | NLST-3D

USA

VITAL STATS

Patients: 900+

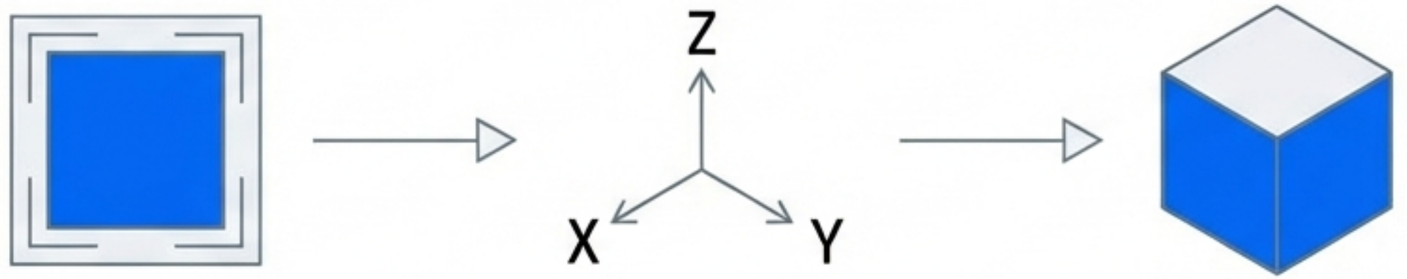
Annotations:
1,192 3D Nodules

Context:
Screening
(Smokers 55-74)

Mortality Reduction:
20%

KEY INNOVATION: 2D TO 3D CONVERSION

The National Lung Screening Trial is a landmark study. We modernized it by converting over 9,000 legacy 2D slice-level bounding boxes into fully verified 3D volumetric annotations.



Merged overlapping consecutive 2D annotations.

12 | DLCS 2024

USA (Duke)

VITAL STATS

Patients: 2,061

Nodules: 2,487

Tech Era:
2015-2021

Format:
Low-Dose CT

KEY INNOVATION: CONTEMPORARY TECH

Most datasets rely on scans from the early 2000s. DLCS reflects modern clinical reality with data from 2015-2021 scanners.

SEMIAUTOMATED ANNOTATION

Pipeline: MONAI Detection -> Radiologist Review.

>90% Time Reduction
98% Accuracy

Includes Lung-RADS scores and cancer outcomes.

13 | DeepLesion-1kTest3D

USA (NIH)

VITAL STATS

Volumes: 1,000 3D NfTI

Lesions: 4,927

Scope:
8 Body Regions

Origin:
PACS Bookmarks

KEY INNOVATION: UNIVERSAL LESION DETECTION

Moving beyond the lung. This dataset covers the entire body. Originally 2D key-slices, we have converted them into full 3D volumes.

ANATOMY GRID

Lung	Liver	Kidney	Bone
Pelvis	Mediastinum	Abdomen	Soft Tissue

Corrected Hounsfield Units from original PNG sources.

The Preprocessing Standard

Why HAID is Plug-and-Play

Raw DICOM / PNG → **HAID Standardizer** → **Unified NIfTI**

Format

Unified NIfTI files compatible with PyTorch/TensorFlow.

Spacing

Resampled to uniform voxel spacing (~0.7mm x 0.7mm x 1.25mm).

Coordinates

DICOM pixel space converted to World Coordinates.

Intensity

Hounsfield Unit (HU) restoration.

Impact: Eliminates custom dataloaders. Enables immediate cross-dataset training.

How to Use HAID

1 BROWSE

Review the dataset catalog for specific modalities/conditions.

2 DOWNLOAD

Access preprocessed **NIIfTI** data via direct **Zenodo** links.

3 CONFIGURE

Load pre-defined **CSV**/**JSON** splits (**Train**/**Val**/**Test**).

4 EXECUTE

Run provided **Jupyter notebooks** for reference implementation.

COMMUNITY

Open Source (**Apache 2.0**).

Pull Requests welcome for new datasets or documentation.

Credits & Contact

Maintainer:

Fakrul Islam Tushar, PhD (@fitushar)

License:

Repository code under Apache 2.0.
Datasets retain original licenses (CC BY 4.0, MIT, etc.).

Acknowledgments:

TCIA (The Cancer Imaging Archive), RSNA, MIDRC, and all original dataset authors.

Scan for Repo

github.com/fitushar/HAID