Optimizing and Harnessing 20 Million NLST CT Lung Screening Images for Robust Foundation Model Training

Md. Enamul Hoq[1*] and Fred Prior[2]

[1] Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA; mhoq@uams.edu

[2] Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA; fwprior@uams.edu

*Correspondence: mhoq@uams.edu

Abstract: We are developing a Foundation Model (FM) for lung CT, aiming to advance cancer screening and biomarker research. We have prepared a well curated AI-ready training dataset comprised of 20 million DICOM lung cancer screening images from the National Lung Screening Trial (NLST) dataset, hosted on the National Cancer Institute's Imaging Data Commons (NCI-IDC). These data were then preprocessed using the Medical Open Network for AI (MONAI) framework and PyDicom. The data were then migrated to the University of Arkansas for Medical Sciences (UAMS) High-Performance Computing (HPC) storage system, scalable and robust computational infrastructure, needed for training our model. Our commitment to open science principles and ethical considerations in data handling underpins our efforts to foster collaboration, accelerate AI adoption in cancer research, and promote equitable access to AI benefits in oncology. Through comprehensive efforts to make lung cancer screening images AI-ready, we aim to enhance early lung cancer detection.

Keywords: Foundation Model, NLST, NCI-IDC, DICOM, MONAI, Precision Oncology, PyDICOM

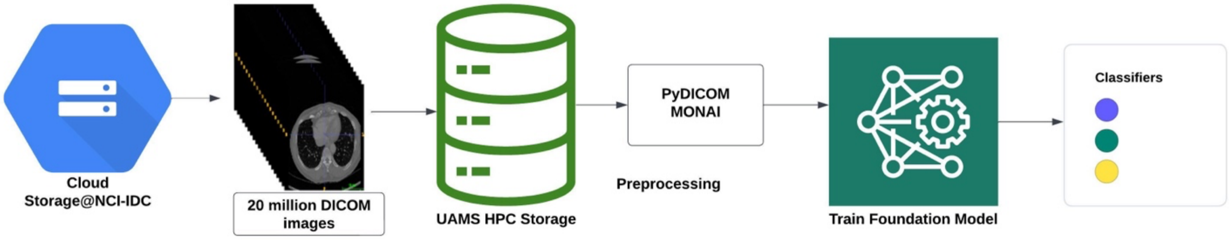

Figure 1: The figure depicts the overall processing pipeline where NLST screening data collected from NCI-IDC