Data-Centric AI for Chest X-Ray Analysis in Resource-Constrained Settings[*]

Yasmeena Akhter
Indian Institute of Technology Jodhpur, India
akhter.1@iitj.ac.in

Abstract

With approximately 2 billion chest X-ray examinations conducted globally each year, the demand for radiological interpretation far surpasses the available expertise, particularly in resource-constrained regions. Recent advancements in artificial intelligence and computer vision present promising automated chest X-ray analysis solutions. Nevertheless, integrating AI-driven diagnostics into clinical practice encounters several challenges, including *data-centric issues*, *implementation barriers*, *deployment complexities*, and the need for *trustworthy AI*. This dissertation focuses on the *data-centric* aspect, making significant contributions through enhanced data collection, creating novel datasets, algorithm development, privacy-preserving collaborative learning, and modelling for low-resolution data. It offers practical methodologies for embedding AI into chest radiology workflows, particularly addressing underserved conditions and healthcare settings with limited data availability. Furthermore, this work illustrates how tailored AI solutions can democratize access to high-quality radiological care while balancing privacy considerations and operational constraints across diverse environments.

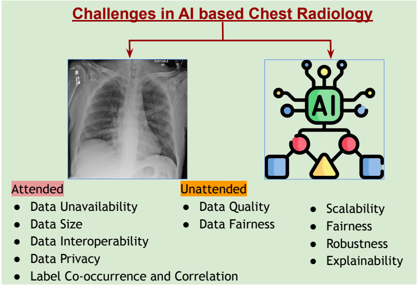

Figure 1: Illustrates the challenges of enabling AI-based automatic diagnosis in chest radiology. This dissertation explores the *Data-Centric* aspect and proposes novel solutions to address these challenges.

1 Introduction

The field of Chest X-ray (CXR) analysis has progressed with the advent of datasets for lung diseases, enabling diverse research tasks such as image enhancement, lung and disease segmentation, pathology classification, automated report generation, and trustworthy AI development. These tasks address clinical needs and advance computer-aided diagnosis for thoracic conditions by automating processes that reduce radiologist burnout and alleviate information overload. However, creating large medical image datasets is challenging due to privacy constraints and the need for specialized annotation. While existing CXR datasets support research on tuberculosis and pneumonia, they often lack sufficient samples for advanced deep-learning applications. Accurate disease classification, detection, and explainability rely on localized ground truth labeling. Moreover, the scarcity of public CXR datasets for pneumoconiosis, COPD, and lung cancer poses significant research barriers, often relegating these conditions to small-sample studies.

The integration of AI into medical imaging has led to a transformative shift in disease diagnostics, offering automated, accurate interpretation of complex radiological images while reducing analysis time, a significant advantage for resource-limited healthcare settings with few radiologists. However, some critical challenges impede widespread adoption. Our review [Akhter *et al.*, 2023b] systematically analyzed diagnostic tasks across pulmonary pathologies such as pneumonia, tuberculosis, COVID-19, lung cancer, and pneumoconiosis, laying the conceptual groundwork for our technical contributions and highlighting promising research directions. This analysis identified key challenges, as illustrated in Figure 1, which shape the focus of this work. Below, we outline research questions to which this work is tied.

RQ1: What approaches enable effective **zero-shot** AI-based detection of early-stage pneumoconiosis without publicly available annotated datasets? **RQ2:** To what extent do **lightweight deep learning models** improve diagnostic accuracy for silicosis in chest radiography under **resource-limited conditions**? **RQ3:** What are the capabilities and limitations of **few-shot learning methods and model re-usage** in achieving robust generalization for chest X-ray diagnostics with minimal training samples? **RQ4:** In what ways

[*] Advisors: Mayank Vatsa and Richa Singh, IIT Jodhpur, India

can **privacy-preserving DL frameworks** (such as differential privacy, etc) balance diagnostic accuracy, usability, and data security in chest radiography? **RQ5:** How effective are AI-based **image-resolution and interpretability techniques** in enhancing diagnostic reliability from low-resolution chest X-ray images in resource-constrained healthcare settings?

2 Our Contributions

The research contributions toward the above-highlighted research gaps are discussed below:

RQ1: Early Stage Pneumoconiosis Diagnosis in Scarce Publicly Accessible Datasets

Pneumoconiosis is a group of occupational lung diseases resulting from inhaled workplace dust. Silicosis is a common and fatal variant caused by crystalline silica exposure in industries such as construction and mining. Research on diagnosing Pneumoconiosis using CXRs relies on small, in-house datasets, with no publicly available datasets [Akhter *et al.*, 2023b]. Our research introduced *Silicodata*, a novel expert-annotated CXR dataset specifically designed for *silicosis* detection. The dataset includes silicosis-related conditions like Silicotuberculosis, Tuberculosis, and healthy lung images. Radiologists provide detailed lung and disease region segmentation annotations in the dataset [Akhter *et al.*, 2025b]. Our collaboration with health centers and NGOs in Rajasthan creates a data collection and implementation pipeline, supporting UN SDGs 3 and 8 to improve health and working conditions [Akhter *et al.*, 2023a].

RQ2: Lightweight Framework for Silicosis Diagnosis

To address the diagnostic challenges of silicosis in resource-limited rural areas, we introduced *SHIELD* (Self-supervised, Silicosis-focused Hierarchical Imaging framework for Early Lung Disease), a computational approach tailored for low compute environments. *SHIELD* employs a multi-resolution strategy that begins with self-supervised learning, enabling the model to learn meaningful representations from limited chest radiograph data without requiring extensive labelled samples. The approach ensures high diagnostic accuracy with minimal computational needs, ideal for settings with limited infrastructure [Akhter *et al.*, 2025a].

RQ3: AI Model Generalization for Small Sample CXRs

To improve CXR model generalization over small datasets in multiclass settings without relying on training data, we introduce Class Aware Selective Knowledge Amalgamation (CASKA). This novel method enhances model generalizability while safeguarding patient privacy. CASKA integrates expertise from teacher models specialized in distinct diseases, ensuring robust diagnostic performance even when training data are isolated across medical institutions. This approach resolves the conflict between collaborative model development and strict patient privacy standards. It also addresses the challenge of requiring extensive training data for model generalization.

RQ4: Safeguarding Patient Privacy in AI-Enabled Chest Radiography Diagnostics

CXR enables the development of chest biometrics. CXR-AI systems face privacy challenges such as incomplete image de-identification, re-identification through dataset cross-referencing, and data transmission or storage vulnerabilities compromising patient information. We propose PrivDiff-Net, a diffusion-based framework that de-identifies CXRs while retaining diagnostic utility. PrivDiff-Net includes two modules: Selective Attribute Suppression, which filters sensitive information via orthogonal projection in cross-attention, and Selective Privacy Guidance, which penalizes identity markers during reverse diffusion. Evaluations on the ChestX-ray14 dataset demonstrate reduced re-identification risk with preserved diagnostic accuracy.

RQ5: Enhancing Low-Resolution Chest Radiograph Diagnostics

We address the challenge of low-resolution chest radiographs in resource-limited clinical settings using the Multi-level Collaborative Attention Knowledge (MLCAK) methodology. The MLCAK method uses Vision Transformers' self-attention mechanism to transfer diagnostic knowledge from high-resolution images, improving low-resolution CXR diagnostics. This approach integrates local pathological findings to improve model explainability, enabling accurate global predictions in a multi-task framework and expanding access to quality diagnostics in diverse healthcare settings [Akhter *et al.*, 2024].

3 Ongoing and Future Work

In our future work, we foresee expanding the current work through a comprehensive multimodal approach to Pneumoconiosis detection, integrating multiple data modalities, including CXR images, detailed radiological interpretations, and textual reports. We will further focus on developing scalability methodologies that overcome the current computational limitations, emphasizing on enhancing model generalization capabilities. Our Future work extends to advance robustness and fairness, ensuring that our deep learning-based diagnostic algorithms demonstrate consistent performance across varied demographic and clinical contexts. We aim to establish a redefined, adaptable, and clinically reliable diagnostic framework for thoracic disorders from Chest X-rays.

References

[Akhter *et al.*, 2023a] Yasmeena Akhter, Rishabh Ranjan, et al. On AI-assisted pneumoconiosis detection from chest X-rays. In *IJCAI*, 2023.

[Akhter *et al.*, 2023b] Yasmeena Akhter, Richa Singh, and Mayank Vatsa. AI-based radiodiagnosis using chest X-rays: A review. *Frontiers in big data*, 6:1120989, 2023.

[Akhter *et al.*, 2024] Yasmeena Akhter, Rishabh Ranjan, et al. Low-resolution chest X-ray classification via knowledge distillation and multi-task learning. In *IEEE ISBI*, 2024.

[Akhter *et al.*, 2025a] Yasmeena Akhter, Rishabh Ranjan, et al. SHIELD: A Self-supervised, silicosis-focused Hierarchical Imaging framework for occupational lung disease Diagnosis. In *IJCAI*, 2025.

[Akhter *et al.*, 2025b] Yasmeena Akhter, Rishabh Ranjan, et al. SilicoData: An Annotated Benchmark CXR Dataset for Silicosis Detection. *Nature Scientific Data*, 2025.