Researh Proporsal

Preparing high-quality X-ray images for AI-based diagnostic systems

Group: SE - 2307

Names: Saulet Kabdrakhmanov, Galymzhan Aliakbar, Aruzhan Atelova and Bekzhan Nurallin

Table of contents

Introduction

Background and Context

Artificial intelligence (AI) and computer vision are increasingly used to support diagnosis of lung diseases from chest X-rays and CT scans [7], [8], [9], [12]. Reviews report rapid progress but also note that data quality and consistent processing are critical for reliable results [8], [9], [10], [11]. In practice, poor contrast, noise, or low resolution can reduce model accuracy, so preparing high-quality images before AI analysis is important [8], [10].

Problem Statement

Hospitals use different X-ray machines and protocols, which leads to variable image quality. Models trained on one dataset may perform worse on images from another source [8], [9], [11]. Although many studies show strong accuracy for tasks such as pneumonia, TB, and COVID-19 detection, performance depends on clear and standardised data and preprocessing [1], [6], [10]. Recent reviews also emphasise that limited standardisation makes it hard to compare methods and slows clinical adoption [8], [9], [12]. These issues point to the need for better image preparation and quality control before using AI diagnostic tools.

Research Questions

1) How does chest X-ray image quality affect AI model performance?

2) Which preprocessing techniques most effectively improve images before AI analysis?
3) Can a standardised preparation pipeline make AI results more reliable across sites?

Relevance and Importance of the Research

Improving X-ray quality is a practical way to make AI diagnostics more accurate and trustworthy. Clear guidance on preprocessing and reporting can support consistent results and safer use in real clinical settings [8], [9], [12].

Literature review

Key Concepts, Theories and Studies

Lately, computer vision and artificial intelligence (AI) have started to change how doctors detect and study lung diseases. The way doctors read chest X-rays or CT scans is often based on the doctor's experience, which can mean that different experts get different results. AI helps reduce these problems by analyzing images automatically and finding disease patterns more accurately. Sindhu et al. [1] explain that deep learning (DL) and machine learning (ML) methods can help doctors detect lung problems such as pneumonia, lung cancer, and COPD faster and more reliably. Mustafa and Nsour [2] tested a computer vision model called YOLO on chest X-ray images and showed that it can detect abnormalities in real time with good accuracy. Felder and Walsh [3] describe new approaches like quantitative computed tomography (QCT) and deep learning systems such as CALIPER, SOFIA, and data-driven texture analysis (DTA), which can measure disease severity and predict how it will progress. Together, these studies show that AI can play a big role in improving lung disease diagnosis and treatment planning in the future.

Key Debates and Controversies

One of the most important debates in computer vision for lung disease detection concerns the quality of the data used to train artificial intelligence (AI) models. Mustafa and Nsour (2023) highlight that "*Chest X-ray images can exhibit significant variations in terms of exposure, contrast, positioning, and the presence of artefacts. These variations can have a direct impact on the performance of automated algorithms*" [2, p.6]. This means that when imaging conditions are inconsistent — for example, when images are blurry, unevenly lit, or captured during patient movement — the accuracy of AI systems decreases. Models trained on noisy or low-quality data often fail to identify subtle lung abnormalities correctly.

The authors further explain that "*subtle abnormalities might be challenging to detect in poor-quality images due to noise or the loss of relevant details*" [2, p.6]. This observation supports the goal of our project, which is to improve the quality of data sets and eliminate factors that cause images to become unclear. By enhancing the clarity and consistency of chest X-ray images, we can reduce noise and emphasise important diagnostic features, such as nodules or opacities. This improves the reliability of AI-based detection.Another key issue concerns explainability — the ability to understand how AI makes decisions.Felder and Walsh (2023) highlight that *"explainable AI will be essential to develop trust within the medical community and facilitate implementation in routine clinical practice"* [3, p.2]. Our project supports this idea by improving the clarity and reliability of the data that AI models learn from. When the data used to train the model is good and consistent, doctors can understand the model's predictions better. Simply put, good data helps build trust, which is needed for AI to be used in actual medical settings.

In addition, Mustafa and Nsour note that "*the chest region comprises intricate and overlapping anatomical structures, including the heart, lungs, ribs, and blood vessels*" [2, p.6]. These overlapping structures make it even more important to work with high-quality and well-preprocessed datasets so that algorithms can accurately distinguish between normal anatomy and pathological findings. Our project directly addresses this challenge by optimizing dataset quality and reducing unnecessary visual interference, allowing AI models to learn from cleaner and more standardized images.

Overall, the authors findings underline that data quality remains one of the main barriers to reliable AI in medical imaging. By tackling this issue, our project contributes to solving a critical problem identified by Mustafa and Nsour (2023) — ensuring that AI models are trained on consistent, high-quality datasets to achieve trustworthy and accurate lung disease detection.

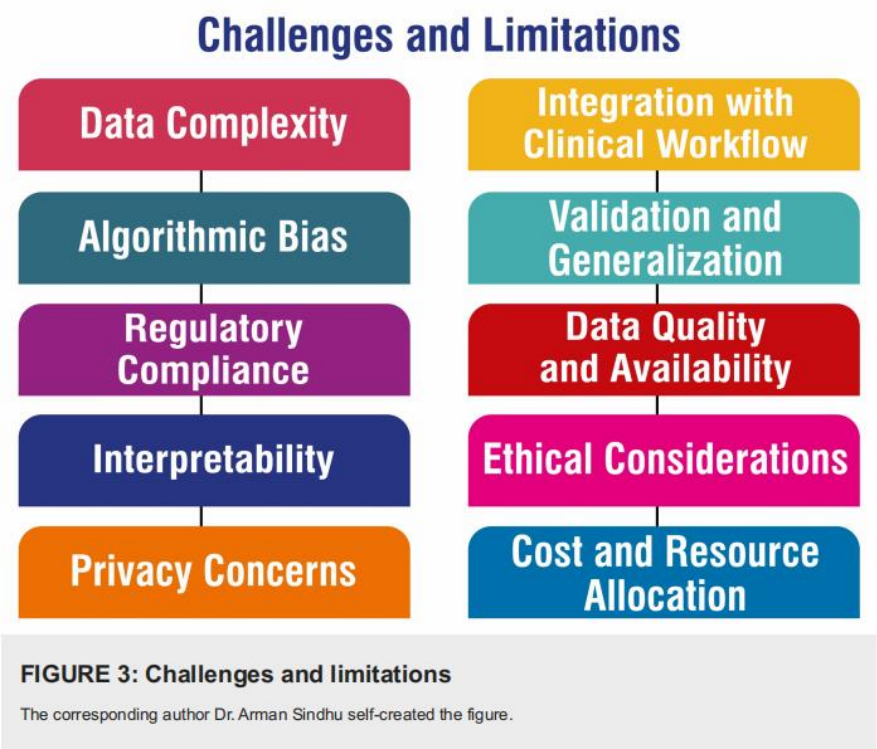

Figure 3. *Challenges and Limitations of AI in Lung Imaging (adapted from Sindhu et al., 2024).*

Gaps in Existing Knowledge

Even though many researchers have studied the use of artificial intelligence (AI) in lung imaging, there are still gaps that limit real progress. Sindhu et al. (2024) explain that current systems face problems with "*data quality, algorithmic bias, transparency, and regulatory compliance*" [1, p.7]. This means that many AI models are trained on incomplete or inconsistent datasets, which makes their results less reliable.

Similarly, Mustafa and Nsour (2023) point out that "*Chest X-ray images can exhibit significant variations in terms of exposure, contrast, positioning, and the presence of artefacts*" [2, p.6]. These differences in image quality make it hard for AI systems to detect small or complex lung abnormalities. Felder and Walsh (2023) also mention ongoing "*issues related to data management and data sharing*" [3, p.2], showing that the medical field still lacks standardized and clean datasets that can be safely used across hospitals.

The aim of our project is to address these issues by enhancing the quality of the dataset and eliminating unwanted interference in chest X-ray images. By creating sharper, clearer, and more consistent datasets, we facilitate the learning process of AI models from reliable data, enabling more accurate predictions. This work strengthens the foundation for future research and supports the safe use of AI in diagnosing lung diseases

Research Design and Methods

To answer the research question, this study will use a Systematic Literature Review. This way works best for looking at already existing info, comparing different research results, and judging how those studies were done, instead of gathering our own information.

Methods and Sources

A. How We'll Search and Where We'll Look

- **Databases:** We'll search big academic databases like Google Scholar, PubMed, IEEE Xplore, and Web of Science. We'll focus on articles from January 2016 up to today (2025).
- **Keywords:** We'll search by these keywords: deep learning, convolutional neural network, computer vision, artificial intelligence and lung disease, pneumonia, tuberculosis or lung cancer and chest X-ray.

B. Selection and Inclusion/Exclusion Criteria

- **Inclusion Criteria:** Articles must focus on the application of deep learning/computer vision to diagnose or classify lung diseases using X-ray images. They must be peer-reviewed or highly cited preprints and be published in English.
- **Exclusion Criteria:** Articles based solely on magnetic resonance (MR) or Positron Emission Tomography (PET) images, review articles, editorials, and non-peer-reviewed conference abstracts will be excluded.

C. Data Extraction and Analysis

The data will be analyzed qualitatively to identify patterns in model effectiveness

Building a research proposal methodology

Research type	This study will use a **qualitative synthesis** approach, relying on existing **secondary sources** like articles, papers, and preprints. The main idea is **descriptive and critical** the current state of AI models, looking at how they're used and where they fall short.
Research methods	A **systematic literature review**. Our "data collection" does not involve collecting new data on a topic, but rather systematically identifying, selecting, and extracting information from existing research articles. Because the project's goal is to critically synthesize and compare the large volume of existing, published studies to assess the reliability and clinical applicability of AI models, which is the most rigorous method for identifying precise research gaps.
Practicalities	• How much time will you need? ~ 5 – 10 weeks • A main challenge is that reports vary across studies. To fix this, we'll create a standard way to score articles based on how clear they are about where their data comes from and how well it's labeled. • Since we won't collect new data, ethical concerns are low. Still, the study will cover the ethical and legal sides of the AI models discussed in the sources, as found when we first looked into the topic. • A limit of this study is that it depends on already available data. That means any biases in those sources will show up in this review.

Research schedule

Week	Research Phase	Objectives / Tasks	Deadline
Week 6	**Data Review and Preparation**	- Collect all required chest X-ray datasets. - Check image quality and identify noise, artefacts, and incorrect labels. - Create a catalog separating “clean” and “noisy” images.	End of Week 6
Week 7	**Data Cleaning and Enhancement** **&** **Writing introduction chapter**	- Apply preprocessing and enhancement techniques (e.g., noise reduction, contrast adjustment). - Remove low-quality or duplicate images. - Develop a script for automated data cleaning.	End of Week 7

		-Read and analyze relevant extra articles(if necessery) -Equaly separate task among the team -Discass with team given task and try to find common solution	
Week 8	**Model Testing with Improved Dataset**	- Train a baseline CNN model using the improved dataset. - Compare model accuracy with results from the original dataset. - Analyze how data quality affects model performance.	End of Week 8
Week 9	**Evaluation and Validation** **&** **Writing methodology**	- Evaluate model performance using metrics . - Visualize results with graphs and confusion matrices. - Summarize findings and performance improvements. -Read and analyze relevant extra articles(if necessery) -Equaly separate task among the team -Discass with team given task and try to find common solution	End of Week 9
Week 10	**Reporting and Final Review** **&** **Create a final paper**	- Write the final report (results, conclusions, limitations). - Review and format the document according to submission guidelines. - Prepare the presentation or	End of Week 10

		defense of the project.	

Implications and contributions to knowledge

This project is important because it addresses one of the key challenges in AI-based medical diagnostics — the quality of X-ray images used for training and testing computer vision models. As several studies have shown, the effectiveness of deep learning systems in medical imaging strongly depends on the quality of the input data (Elyan et al., 2022, 34). Poor-quality images, affected by noise or blurriness, can lead to inaccurate model predictions and limit the clinical reliability of AI systems (Olveres et al., 2020, 3835).

By focusing on preparing and enhancing high-quality X-ray images, this project contributes to both practical medical applications and theoretical development in the field of AI for healthcare.

Practical Implications

The results of this research can directly improve the diagnostic process in medicine. Enhancing the quality of X-ray images before they are used by AI systems can increase diagnostic accuracy, reduce errors, and speed up the decision-making process for doctors and radiologists. As Liu et al. (2022, 179- 188- 192) noted, the performance of deep learning systems can be dramatically improved when trained on large-scale, high-quality, and well-annotated imaging datasets.

In practice, this means that hospitals and AI developers could implement better preprocessing pipelines for medical images, leading to more reliable diagnostic tools. In regions with limited medical resources, such improvements could make AI-based diagnosis more accessible and effective. Furthermore, this research can inform policies and standards for data collection and image preparation in medical AI systems.

Theoretical Implications

From a theoretical perspective, this work strengthens the existing understanding that data quality is one of the most critical factors in machine learning performance (Elyan et al., 2022, 34). It supports and extends previous studies showing that unclear or noisy medical images reduce the interpretability and accuracy of AI models.

Moreover, this research provides a foundation for future work on integrating image enhancement techniques directly into the architecture of diagnostic models. By demonstrating

that improving image quality at the preprocessing stage significantly boosts model performance, it may inspire the development of new hybrid models that combine image restoration and disease detection. In this way, the project contributes to building a stronger theoretical basis for designing more explainable and clinically reliable AI systems.

Budget

The proposed project will require a moderate budget mainly focused on data acquisition, computational resources, and software tools. The total estimated budget is **$4,800**.

Item	Cost (USD)	Justification	Source / Calculation
1. Data access and licensing	**$1,000**	To access high-quality, annotated chest X-ray image datasets . These datasets are essential for training and testing AI models.	Based on dataset licensing fees listed by public repositories and data providers.
2. Computing resources (cloud GPU access)	**$2,000**	Needed for training and testing deep learning models using large-scale image data. High-performance GPUs significantly reduce training time and improve experimental efficiency.	Estimated from average hourly rates of NVIDIA GPU cloud services (Google Colab Pro+, AWS, or Kaggle).
3. Software and image preprocessing tools	**$800**	For software licenses and image enhancement tools (e.g., MATLAB, Adobe Photoshop for visualization, and paid AI libraries if required).	Based on annual academic licenses and subscription prices.
4. Travel for collaboration / data collection	**$500**	Covers travel to a local medical institution or research center to discuss data annotation standards and validate the research approach with professionals.	Estimated from local transport and short-distance travel costs.

5. Miscellaneous expenses	**$500**	For data storage devices (external SSDs), backups, and documentation materials.	Based on current market prices for storage hardware.

Data taken from articles :

Michael E. Kim, K. Ramadass, C. Gao, et al., “Scalable, reproducible, and cost-effective processing of large-scale medical imaging datasets,”https://www.researchgate.net/publication/383460554_Scalable_reproducible_and_cost-effective_processing_of_large-scale_medical_imaging_datasets

G. Yeung, D. Borowiec, A. Friday, R. Harper, P. Garraghan, “Towards GPU Utilization Prediction for Cloud Deep Learning,” *Proceedings of the USENIX Workshop on HotCloud*, 2020.https://dl.acm.org/doi/10.5555/3485849.3485855

References

[1]A. Sindhu, U. Jadhav, Babaji Ghewade, J. Bhanushali, and P. Yadav, “Revolutionizing Pulmonary Diagnostics: A Narrative Review of Artificial Intelligence Applications in Lung Imaging,” *Cureus*, Apr. 2024, doi: https://doi.org/10.7759/cureus.57657.

[2] Z. Mustafa and H. Nsour, "Using computer vision techniques to automatically detect abnormalities in chest X-rays," *Diagnostics*, vol. 13, no. 18, p. 2979, 2023. DOI: 10.3390/diagnostics13182979.

[3]F. N. Felder and S. L. F. Walsh, “Exploring computer-based imaging analysis in interstitial lung disease: opportunities and challenges,” *ERJ Open Research*, vol. 9, no. 4, Jul. 2023, doi: https://doi.org/10.1183/23120541.00145-2023.

[4] K. D. Lumamba, G. Wells, D. Naicker, T. Naidoo, C. Steyn, and Mandlenkosi Gwetu, “Computer vision applications for the detection or analysis of tuberculosis using digitised human lung tissue images - a systematic review,” *BMC Medical Imaging*, vol. 24, no. 1, Nov. 2024, doi: https://doi.org/10.1186/s12880-024-01443-w

[5] Nurmukhammed Ernestov and Dim Shaiakhmetov, “Development and Deployment of a Computer Vision Model for Diagnosing Lung Diseases from X-Ray Images,” May 2025, doi: https://doi.org/10.20944/preprints202505.0601.v1.

[6] N. Divya, “Detection of COVID-19 and Pneumonia using Chest X-ray Scans with Deep Learning,” *Advances in Computational Sciences and Technology*, vol. 16, no. 1, pp. 27–34, Jun. 2023, doi: https://doi.org/10.37622/acst/16.1.2023.27-34.

[7] H. J. Yoon, Y. J. Jeong, H. Kang, J. E. Jeong, and D.-Y. Kang, “Medical Image Analysis Using Artificial Intelligence,” *Korean Society of Medical Physics*, vol. 30, no. 2, pp. 49–58, Jun. 2019, doi: https://doi.org/10.14316/pmp.2019.30.2.49.

[8] E. Elyan, P. Vuttipittayamongkol, P. Johnston, K. Martin, K. McPherson, C. F. Moreno-García, C. Jayne, and M. M. K. Sarker, “Computer vision and machine learning for medical image analysis: Recent advances, challenges, and way forward,” *Artif. Intell. Surg.*, vol. 2, no. 1, pp. 24–45, 2022. [Online]. Available: https://www.oaepublish.com/articles/ais.2021.15

[9] J. Olveres, G. González, F. Torres, J. C. Moreno-Tagle, E. Carbajal-Degante, A. Valencia-Rodríguez, N. Méndez-Sánchez, and B. Escalante-Ramírez, “What is new in computer vision and artificial intelligence in medical image analysis applications,” *Ann. Transl. Med.*, vol. 9, no. 23, p.1750, 2021. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8245941/#abstract1

[10] T. Liu, E. Siegel, and D. Shen, “Deep learning and medical image analysis for COVID-19 diagnosis and prediction,” *Annu. Rev. Biomed. Eng.*, vol. 24, pp. 179–201, 2022. [Online]. Available: https://www.annualreviews.org/content/journals/10.1146/annurev-bioeng-110220-012203

[11] J. Gao, Y. Yang, P. Lin, and D. S. Park, “Computer vision in healthcare applications,” *J. Healthc. Eng.*, vol. 2018, Article ID 5157020, 2018. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC5857319/

[12] D. Khemasuwan, J. S. Sorensen, and H. G. Colt, “Artificial intelligence in pulmonary medicine:Computer vision, predictive model and COVID-19,” *Eur. Respir. Rev.*, vol. 29, no. 157, p. 200181, 2020. [Online]. Available: https://publications.ersnet.org/content/errev/29/157/200181.abstract

[13] Michael E. Kim, K. Ramadass, C. Gao, et al., “Scalable, reproducible, and cost-effective processing of large-scale medical imaging datasets,”https://www.researchgate.net/publication/383460554_Scalable_reproducible_and_cost-effective_processing_of_large-scale_medical_imaging_datasets

[14] G. Yeung, D. Borowiec, A. Friday, R. Harper, P. Garraghan, “Towards GPU Utilization Prediction for Cloud Deep Learning,” *Proceedings of the USENIX Workshop on HotCloud*, 2020.https://dl.acm.org/doi/10.5555/3485849.3485855