Deep Learning for Healthcare Diagnostics: Improving Model Accuracy While Ensuring Ethical and Explainable AI

(PhD Dissertation Example – Computer Science / Biomedical AI · For Educational Use)

ABSTRACT

Medical imaging has become one of the most promising areas for deep learning applications, particularly for early disease detection. However, concerns about algorithmic bias, opacity, and uneven performance across demographic sub-groups limit real-world adoption.
This doctoral study investigates how a convolutional neural network (CNN)–based diagnostic model can improve classification accuracy for lung-disease detection while ensuring transparency, fairness, and ethical integrity.

A dataset of 18,400 annotated chest X-rays was pre-processed, augmented, and split into training, validation, and test sets.
A baseline CNN was compared with a fine-tuned ResNet-50 architecture. Grad-CAM, SHAP values, subgroup performance audits, and five domain-expert interviews were used to evaluate accuracy, interpretability, and perceived clinical trustworthiness.

Findings show that the fine-tuned model achieved a 6.8% increase in accuracy and 9.3% increase in sensitivity, particularly for early-stage abnormalities. Explainability tools improved clinician confidence, but highlighted ethical concerns around dataset imbalance and opacity of model reasoning.
The study concludes that technical performance alone is insufficient for safe clinical deployment and proposes a structured framework for accuracy, fairness, and ethical governance in medical AI.

CHAPTER 1: INTRODUCTION

Deep learning is increasingly used to support clinical decision-making, yet its adoption in frontline healthcare remains limited. Hospitals face three main barriers:
(1) uneven accuracy across demographic groups,
(2) poor explainability of model decisions, and
(3) uncertainty about compliance with ethical and regulatory standards.

Research Aim

To develop and evaluate a deep learning diagnostic model that balances performance improvements with transparency, fairness, and ethical oversight.

Research Questions

1. How accurately can a deep learning model classify lung-disease patterns from chest X-rays compared with traditional baselines?
2. What factors most influence clinician trust in AI diagnostic tools?
3. How can explainability and fairness audits strengthen ethical compliance in medical AI systems?

Research Objectives

- Build and evaluate a CNN model using a large X-ray dataset.
- Compare baseline and fine-tuned architectures for diagnostic accuracy.
- Apply interpretability tools to assess transparency.
- Conduct fairness and subgroup-performance audits.
- Propose an ethical governance framework for medical AI.

CHAPTER 2: LITERATURE REVIEW

Existing literature shows strong performance of CNNs for tasks such as pneumonia detection, tumour classification, and anomaly segmentation. However, three gaps persist:

2.1 Diagnostic Accuracy

Earlier studies achieved 85–95% accuracy using architectures such as VGG, ResNet and DenseNet.
Yet, few studies benchmark performance against fairness or trust metrics, relying solely on quantitative accuracy.

2.2 Explainability & Clinician Trust

Research consistently shows that clinicians hesitate to use AI models that operate as "black boxes."
Tools such as Grad-CAM, LIME, and SHAP improve transparency, but many systems still lack interpretable decision pathways.

2.3 Ethical Concerns & Dataset Imbalance

Common issues include:

- Skewed representation (e.g., age, ethnicity, disease prevalence)
- Hidden biases in annotation
- Lack of transparency in training data provenance
- Insufficient analysis of false-positive and false-negative risks

2.4 Research Gap

No documented PhD-level study integrates:

- accuracy optimisation
- interpretability
- fairness
- ethical oversight
  in a single evaluation framework.

This dissertation addresses that gap.

CHAPTER 3: METHODOLOGY

This study follows a mixed-methods design combining:

1. **Quantitative modelling** (deep learning pipeline)
2. **Qualitative interviews** with clinicians
3. **Ethical + fairness audits**

3.1 Dataset

- 18,400 anonymised chest X-rays (normal vs abnormal)
- Source: public medical datasets with ethics approval
- Pre-processing: resizing, noise reduction, normalisation, augmentation

3.2 Model Development

Two architectures:

Model A: Baseline CNN

- 5 convolutional layers
- ReLU activation
- Max-pooling
- Dense softmax output

Model B: Fine-tuned ResNet-50

- Pretrained on ImageNet
- Final layers retrained on medical dataset
- Early stopping, dropout, and LR scheduling

3.3 Training Configuration

- Optimiser: Adam
- Learning rate: 0.0001
- Batch size: 32
- Train/val/test: 70/15/15 split
- Evaluation metrics: Accuracy, Precision, Recall, F1, AUC

3.4 Explainability & Fairness Tools

- Grad-CAM: heatmap visualisation of model attention
- SHAP: feature attribution
- Subgroup audits: age, gender, ethnicity
- Interview instrument: semi-structured clinical trust survey

3.5 Ethical Approval

Ethics clearance granted under “non-interventional use of anonymised secondary data.”
Clinician interviews followed informed consent procedures.

CHAPTER 4: RESULTS

4.1 Model Performance

Metric	Baseline CNN	Fine-tuned ResNet-50
Accuracy	87.2%	**93.8%**
Sensitivity	81.4%	**90.7%**
Specificity	88.9%	**92.3%**
AUC	0.91	**0.96**

4.2 Explainability Findings

Grad-CAM heatmaps aligned well with pathological regions identified by clinicians.
SHAP visualisations revealed sensitivity to texture gradients and opacity clusters.

4.3 Fairness & Bias Insights

Subgroup analysis showed:

- Slightly lower sensitivity for older age groups
- Minimal disparity across gender
- Higher false-negatives for under-represented ethnic groups

4.4 Clinician Interviews

Themes:

- Higher trust with transparent reasoning
- Desire for on-screen explanations
- Concerns about over-reliance on automated outputs

CHAPTER 5: DISCUSSION

The fine-tuned model significantly outperformed the baseline CNN, confirming the value of transfer learning for medical imaging.
However, raw accuracy did not guarantee clinical trust.
Clinicians responded positively to interpretability tools, yet highlighted:

- ethical risks of dataset imbalance
- need for reproducible validation

- importance of transparent reporting
- compliance with GDPR and medical-device regulation (MDR)

This supports the argument that **ethical AI must be both accurate and accountable**.

CHAPTER 6: CONCLUSION

Deep learning can meaningfully enhance diagnostic accuracy, but cannot be deployed safely without fairness checks, transparency mechanisms, and clear governance procedures.
This study proposes a five-pillar framework:

1. Accuracy benchmarking
2. Interpretability integration
3. Fairness audits
4. Ethical governance
5. Clinician-centred evaluation

Future research should incorporate multimodal datasets and real-time decision support.

NOTE TO STUDENTS

This is a partial sample.
To receive the **full PhD dissertation PDF (all chapters + extended results):**
Request it through the website or WhatsApp.