

Research Methodology in Computer Science - Assignment 1

The Impact of Artificial Intelligence on Medical Diagnosis Accuracy: A Literature Review.

Paper 1

Deep Learning for Chest Radiograph Diagnosis

Summary

This study aimed to evaluate whether deep learning could match or surpass radiologists in diagnosing pneumonia from chest X-rays. Researchers used a 121-layer convolutional neural network trained on the NIH ChestX-ray14 dataset. CheXNet achieved performance comparable to practicing radiologists on pneumonia detection.

Critical Analysis

Strengths

- Extremely large dataset (over 100,000 images).
- Direct comparison with multiple certified radiologists.
- High external validity for pneumonia detection.

Weaknesses

- Limited to pneumonia; generalizability to other thoracic diseases less clear.
- Dataset labeling relied partly on NLP reports, introducing potential labeling noise.

Connections/Conflicts

Findings align with Papers 2 and 4, which also show DL outperforming clinicians for specific tasks.

Paper 2

Artificial Intelligence Versus Clinicians: Systematic Review

Summary

This systematic review evaluated over 20 studies comparing AI systems to human clinicians in diagnostic performance. Across imaging-based tasks (radiology, dermatology, ophthalmology), AI matched clinicians in accuracy in most studies and exceeded them in some specialized tasks.

Critical Analysis

Strengths

- Comprehensive and systematic coverage of multiple medical domains.
- Evaluates real-world performance comparisons of AI and clinicians.

Weaknesses

- Many included studies suffered from small sample sizes.
- Lack of real-clinical deployment environments reduces ecological validity.

Connections/Conflicts

Consistent with Papers 1 and 3 showing parity or superiority of AI in imaging diagnostics. Conflicts with Paper 5 on issues of generalizability.

Paper 3

Deep Learning for Dermatologist-Level Skin Cancer Classification

Summary

This research trained a deep CNN on 129,450 clinical images covering over 2,000 skin disease classes. The AI system achieved dermatologist-level performance in detecting melanoma and carcinoma.

Critical Analysis

Strengths

- One of the largest dermatology image datasets available.
- Demonstrated human-level performance using only images.

Weaknesses

- Lack of demographic diversity in the dataset; may not generalize across skin tones.
- Clinical workflow integration not addressed.

Connections/Conflicts

Supports Papers 1 and 2 by reinforcing AI's strong performance in image-based diagnosis. Contrasts with Paper 6's ethical concerns regarding bias.

Paper 4

AI-Based Diabetic Retinopathy Detection

Summary

Google's deep learning algorithm was evaluated for diabetic retinopathy screening. Using 128,000 retinal images for training, the AI achieved sensitivity of 97.5% and specificity of 93.4%, rivaling expert ophthalmologists.

Critical Analysis

Strengths

- Strong clinical validation with multiple test sets.
- High diagnostic accuracy makes it suitable for screening programs.

Weaknesses

- Limited explainability of the model decisions.
- Requires high-quality images, limiting use in low-resource settings.

Connections/Conflicts

Complements Papers 1–3 showing strong AI performance in medical imaging. Highlights the need for consistency in image quality (ties to Paper 5).

Paper 5

Generalizability Challenges of AI in Medical Imaging

Summary

This study examined whether deep learning models for pneumonia detection generalize to data from different hospitals. The models trained on one hospital dataset performed poorly on external datasets due to confounding factors (e.g., device-specific artifacts).

Critical Analysis

Strengths

- First major study revealing fragility and bias in medical imaging AI.
- Strong methodology evaluating multiple cross-hospital datasets.

Weaknesses

- Focused only on chest X-ray models.
- Did not evaluate potential correction methods.

Connections/Conflicts

Contradicts Papers 1 and 4 by showing AI accuracy decreases on new datasets.
Supports ethical concerns mentioned in Paper 6.

Paper 6

Ethical Implications of AI in Diagnostics: A Review

Summary

This paper reviews ethical, legal, and practical implications of using AI in clinical diagnosis. Highlights bias, lack of transparency, accountability issues, and risks of overreliance by clinicians.

Critical Analysis

Strengths

- Broad ethical and policy perspective.
- Identifies key challenges for clinical adoption.

Weaknesses

- Lacks empirical data; mostly conceptual analysis.
- No concrete evaluation of specific AI models.

Connections/Conflicts

Adds context to Papers 1–4 by highlighting non-technical barriers. Supports Paper 5 with discussion on bias and generalizability.

SYNTHESIS OF FINDINGS

Across the six papers, several themes emerge

1. AI Achieves or Exceeds Clinician-Level Accuracy

Papers 1–4 consistently demonstrate that deep learning models perform at or above clinician level in image-based diagnosis (chest X-rays, retinal scans, dermatology images).

2. Challenges in Generalizability

Paper 5 highlights that despite high accuracy on internal data, models may fail when applied to new populations or imaging devices. This challenges the broader claims of universality in Papers 1–4.

3. Ethical and Practical Barriers

Paper 6 emphasizes that performance alone is not enough—bias, explainability, regulation, and accountability must be addressed before wide adoption.

4. Strongest Consensus

AI excels in image analysis and is a valuable adjunct tool, but must be validated broadly to avoid deployment risks.

5. Key Conflict

Claimed performance advantages (Papers 1–4) are limited by real-world constraints (Papers 5–6), requiring careful deployment.

COMPARISON TABLE OF REVIEWED PAPERS

Year	Author(s)	Title	Methodology	Key Findings	Weaknesses / Limitations
2017	Rajpurkar et al.	CheXNet Pneumonia Detection	CNN on large chest X-ray dataset	AI matches/exceeds radiologists	Labeling noise; limited to pneumonia
2019	Topol AI vs. Clinicians Review	AI vs. Clinicians Review	Systematic review	AI matches clinicians across domains	Many studies use small datasets
2017	Esteva et al.	Skin Cancer Classification	CNN trained on 129k images	Dermatologist-level accuracy	Poor demographic diversity
2016	Gulshan et al.	Diabetic Retinopathy AI	DL on retinal images	High sensitivity/specificity	Needs high-quality images
2018	Zech et al.	Generalizability of AI	Cross-hospital evaluation	Poor external performance	Limited disease scope
2018	Char et al.	Ethics of AI Diagnostics	Conceptual/ethical review	Highlights bias & accountability issues	No empirical validation

CONCLUSION

The reviewed literature demonstrates that AI has transformative potential in diagnostic medicine, achieving performance comparable to expert clinicians in several imaging-based tasks. However, generalizability, bias, lack of explainability, and ethical concerns significantly limit its immediate clinical adoption. The synthesis of these six papers shows that while AI can enhance diagnostic accuracy, successful deployment requires robust validation, transparency, and regulatory oversight.