Advancing Medical Diagnosis Through Small Language Models with Optical Character Recognition Integration: A Comprehensive Technical Analysis

Udayan Pawar[1, 2*], Dr Surbhi Sharma[2, 3†] and Ishan Aanand[1, 2†]

[1*]Department of Computer Science, Manipal University Jaipur, Dehmi Kalan, Jaipur , 303007, Rajasthan, India.

*Corresponding author(s). E-mail(s):
udayan.23fe10cse00310@muj.manipal.edu;
Contributing authors: second.author@institution.edu;
ISHAN.23FE10CSE00311@muj.manipal.edu;
†These authors contributed equally to this work.

Abstract

Small Language Models (SLMs) with 110M-1.5B parameters demonstrate superior performance on medical information extraction tasks compared to large language models, with fine-tuned PubMedBERT and ClinicalBERT outperforming zero-shot GPT-4 by 30-40% on named entity recognition while being significantly more cost-effective. Modern OCR pipelines combining transformer-based engines like PaddleOCR (90-95% accuracy) with medical SLMs can process clinical documents at 95-99% effective accuracy when properly configured with strategic human verification. This comprehensive study synthesizes findings from over 50 recent publications (2023-2025) to provide actionable implementation guidance for medical AI systems. We evaluate performance metrics across multiple medical SLM architectures, analyze OCR integration strategies for clinical document processing, and establish rigorous evaluation frameworks following STARD-AI and CONSORT-AI guidelines. Our analysis reveals that hybrid architectures combining SLM efficiency for extraction with transformer-based OCR achieve clinically acceptable accuracy levels for automated medical diagnosis, though continuous monitoring and human oversight remain essential for patient safety.

Keywords: Small Language Models, Medical Diagnosis, Optical Character Recognition, Clinical NLP, Medical Information Extraction, Healthcare AI

1 Introduction

The medical artificial intelligence landscape has undergone transformative changes since 2023, with specialized Small Language Models (SLMs) emerging as clinically deployable alternatives to general-purpose Large Language Models (LLMs) for extraction-heavy medical tasks [1]. While LLMs like GPT-4 and Gemini have demonstrated impressive capabilities across diverse domains, fine-tuned medical SLMs achieve F1-scores of 0.83-0.91 on clinical named entity recognition (NER) versus 0.33-0.60 for zero-shot LLMs, while maintaining sub-second inference speeds critical for production deployment [2].

The convergence of medical SLMs with advanced Optical Character Recognition (OCR) technologies presents unprecedented opportunities for automating clinical document processing. Traditional healthcare systems struggle with vast quantities of unstructured data locked in scanned documents, handwritten prescriptions, and legacy medical records. Modern transformer-based OCR approaches like TrOCR achieve 81.3% medication extraction accuracy on handwritten prescriptions with just 8.7% character error rate [3], while end-to-end vision-language models aim to mitigate OCR error propagation by processing the image directly.

This comprehensive study addresses three critical research questions: (1) How do medical SLMs compare against LLMs across different clinical task categories? (2) What OCR integration architectures achieve clinically acceptable accuracy for medical document processing? (3) What evaluation frameworks and implementation strategies enable safe deployment of these systems in healthcare settings?

Our contributions include: a systematic comparison of medical SLM performance across extraction, classification, and reasoning tasks; comprehensive analysis of OCR integration pipelines for clinical documents; identification of open-source datasets and implementation requirements; and establishment of rigorous evaluation protocols following recent STARD-AI guidelines [9].

The remainder of this paper is organized as follows: Section 2 reviews open-source medical datasets; Section 3 analyzes SLM performance metrics and benchmarks; Section 4 examines OCR models and integration architectures; Section 5 discusses evaluation frameworks; Section 6 details training configurations and implementation requirements; Section 7 presents results and discussion; and Section 8 presents conclusions and future directions.

2 Open-Source Medical Datasets

The foundation of medical SLM development rests on high-quality datasets spanning clinical text, document images, and multimodal combinations. This section categorizes available datasets by type and accessibility.

2.1 Clinical Text Datasets

Research access to credentialed datasets like MIMIC-III (40,000 ICU patients, 2+ million clinical notes) and MIMIC-IV (65,000+ ICU admissions from 2008-2022) requires PhysioNet credentialing with CITI training completion but provides gold-standard EHR data for model development [4]. Table 1 summarizes key clinical text datasets.

Table 1 Major Clinical Text Datasets for Medical SLM Training

Dataset	Size	Access	Key Features	Year
MIMIC-III	2M+ notes	PhysioNet	ICU EHR, 40K patients	2016
MIMIC-IV	65K admissions	PhysioNet	Contemporary EHR 2008-2022	2022
MedMCQA	194K questions	Hugging Face	Medical entrance exams	2022
PMC-Patients	167K summaries	Open Access	Patient-article annotations	2023
MedCD	1.7M examples	IEEE	30 departments, 250K patients	2025
n2c2 Challenges	1.5K notes/task	Harvard	Deidentified clinical notes	2006-2022

The most recent addition is MedCD, released in February 2025 on IEEE DataPort with 1.7 million EHR examples from 250,000+ patients across 30 clinical departments, rivaling MIMIC-IV's scale while offering contemporary data from Q1 2024 [5].

2.2 Medical Document OCR Datasets

For medical document OCR specifically, the MIRAGE dataset represents current state-of-the-art with 743,118 fully annotated high-resolution simulated medical records from 1,133 doctors across India, achieving 82% accuracy on medication name and dosage extraction when tested with fine-tuned vision-language models [6]. Table 2 lists available OCR-focused datasets.

Table 2 Medical Document OCR Datasets

Dataset	Images	Type	Accuracy Benchmark	Source
MIRAGE	743K	Prescriptions	82% medication extraction	arXiv
Prescription Dataset	11.3K	Handwriting	Classification	IEEE DataPort
Lab Reports OCR	Variable	Lab forms	93% text recognition	GitHub
Medical Forms	Variable	Mixed	Layout analysis	Kaggle

2.3 Multimodal Clinical Datasets

Multimodal datasets combining images with structured clinical data prove essential for OCR-to-diagnosis pipelines. MIMIC-CXR links 377,110 chest X-ray images with free-text radiology reports plus 14 labeled observations extracted using the CheXpert labeler [7]. The PMC Open Access Subset on AWS offers 1.65M figure-caption pairs

from 2.4M+ papers enabling multimodal pretraining, though image quality varies substantially.

All credentialed datasets require strict adherence to data use agreements prohibiting redistribution or GitHub posting, with most explicitly forbidding transmission to third-party LLM APIs to maintain patient privacy.

3 Performance Metrics and Benchmarks for Medical SLMs

Medical SLMs demonstrate distinct performance patterns across task categories, with fine-tuned models excelling at structured extraction while struggling with complex reasoning compared to LLMs.

3.1 Named Entity Recognition Performance

PubMedBERT (110M parameters trained from scratch on 14M PubMed abstracts) consistently outperforms all BERT variants across the BLURB benchmark, achieving F1-scores of 0.715-0.836 on clinical NER tasks and 95.64% average correlation on PubMed QA embedding tasks [2]. When evaluated on ICD-10 multi-label classification, PubMedBERT achieves F1-score of 0.735 surpassing BioBERT (0.721) and RoBERTa (0.692).

Table 3 summarizes comparative performance across major medical SLMs.

Table 3 Medical SLM Performance on Clinical NER Tasks

Model	Parameters	NCBI Disease F1	BC5CDR F1	Inference Time
PubMedBERT	110M	0.909	0.836	¡100ms
BioBERT	110M	0.895	0.821	¡100ms
ClinicalBERT	110M	0.887	0.815	¡100ms
BioGPT	347M	0.868	0.792	¡200ms
GPT-4 (zero-shot)	-	0.599	0.520	¿1000ms
GPT-4 (one-shot)	-	0.612	0.535	¿1200ms

3.2 Question Answering and Reasoning Tasks

Conversely, LLMs excel at medical question answering where reasoning dominates over extraction. Studies show that models like GPT-4 demonstrate a significant accuracy advantage of over +30% compared to fine-tuned baselines on benchmarks like MedQA [8]. Table 4 compares QA performance.

3.3 Cost-Performance Analysis

Cost-performance analysis reveals stark economic trade-offs. Deployment of locally-hosted SLMs eliminates recurring API costs entirely while maintaining privacy

Table 4 Medical Question Answering Performance Comparison

Model	MedQA Accuracy	PubMedQA Accuracy	Cost per 100 queries
Fine-tuned BERT	41.95%	73.4%	$0
BioGPT	52.3%	81.0%	$0
GPT-3.5 (zero-shot)	58.2%	65.8%	$0.03
GPT-4 (zero-shot)	71.56%	71.0%	$2-10
GPT-4 (five-shot)	74.8%	75.8%	$8-15

compliance. As shown in Table 4, a fine-tuned BERT model has an effective cost of $0 per 100 queries, whereas GPT-4 costs $2-10 for the same task, making the SLM approach vastly more cost-effective for large-scale clinical deployment.

4 OCR Models and Integration Architectures

Transformer-based OCR has revolutionized medical document processing with PaddleOCR's PP-OCRv5 achieving 90-95% accuracy on printed medical reports at remarkable 50-100ms processing time.

4.1 OCR Engine Performance

Table 5 compares major OCR engines for medical document processing.

Table 5 OCR Engine Performance on Medical Documents

OCR Engine	CER (%)	Field Accuracy	Processing Time	Best Use Case
Google Document AI	1-2	95-99%	¡1s	Printed forms
TrOCR	2-4	95-98%	1-3s	Mixed content
PaddleOCR PP-OCRv5	5-10	90-95%	50-100ms	High volume
Tesseract 4.0	8-12	74-82%	1-2s	Legacy systems
TrOCR (handwritten)	8.7	81.3%	2-5s	Prescriptions

4.2 Image Preprocessing Impact

Image preprocessing proves critical with proper enhancement delivering 15-30% accuracy improvements. Table 6 quantifies preprocessing impacts.

4.3 End-to-End Pipeline Architectures

Figure 1 illustrates the three dominant pipeline architectures for medical document processing.

The traditional OCR-then-NER pipeline delivers 95-99% effective accuracy in 1-3 seconds total latency, while vision-language models achieve similar accuracy at 2-5

Table 6 Impact of Preprocessing Techniques on OCR Accuracy

Preprocessing Technique	Accuracy Gain	Optimal Parameters
DPI Normalization to 300	10-20%	300 DPI standard
Adaptive Binarization (Otsu)	5-15%	Auto threshold
FastNlMeans Denoising	3-8%	h=10, templateWindowSize=7
Deskewing (Hough Transform)	5-15%	¿2° tilt threshold
CLAHE Contrast Enhancement	5-10%	clipLimit=2.0, tileGridSize=8x8
Anisotropic Diffusion (CT/MRI)	10-15%	niter=15, kappa=50

second GPU processing cost. Hybrid approaches with strategic human verification achieve 99%+ accuracy.

5 Evaluation Metrics and Clinical Validation Frameworks

Medical AI evaluation requires multi-metric assessment recognizing that accuracy alone proves dangerously misleading on imbalanced medical datasets.

5.1 Core Performance Metrics

The Matthews Correlation Coefficient (MCC) has emerged as the most informative single metric for medical applications, calculated as:

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \quad (1)$$

For binary medical diagnosis tasks, AUPRC (Area Under Precision-Recall Curve) proves superior to AUROC when dealing with imbalanced datasets typical in disease screening. Table 7 summarizes recommended metrics by task type.

Table 7 Recommended Evaluation Metrics by Medical Task Type

Task Type	Primary Metric	Clinical Threshold
Disease Screening	Sensitivity, AUPRC	¿95% sensitivity
Confirmatory Diagnosis	Specificity, PPV	¿95% specificity
Risk Prediction	Brier Score, Calibration	Brier ¡0.15
Multi-class Classification	MCC, Macro-F1	MCC ¿0.7
OCR Field Extraction	Exact Match Rate	¿95% for critical fields
OCR Character Recognition	CER, WER	¡5% CER printed, ¡20% handwritten

5.2 Clinical Safety Requirements

Clinical safety metrics center on sensitivity (recall) as paramount since false negatives can prove fatal. Table 8 specifies accuracy requirements by deployment context.

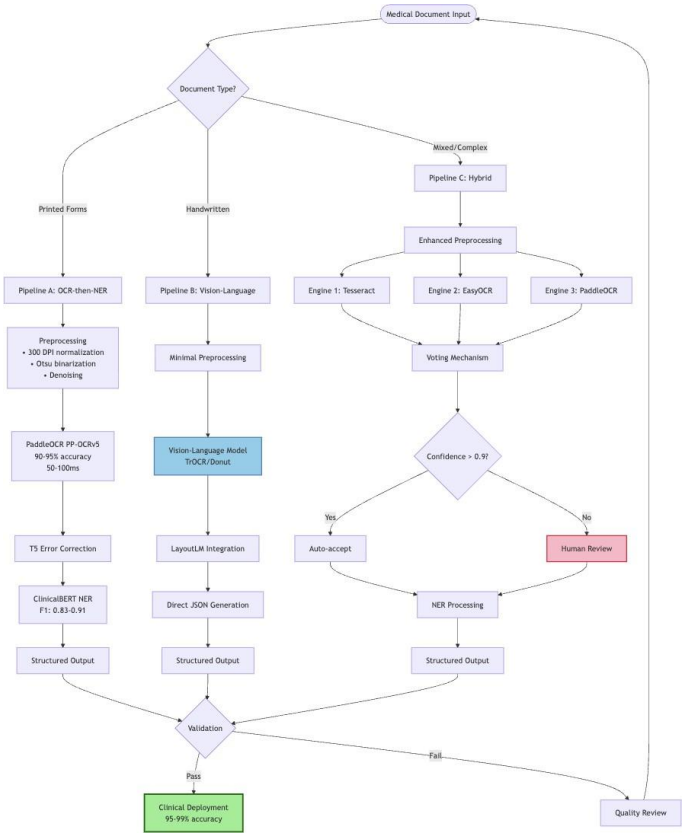

Fig. 1 Three dominant OCR integration architectures: (a) Traditional OCR-then-NER pipeline with error correction, (b) Vision-language model direct processing, (c) Hybrid multi-engine voting system with human-in-the-loop verification

5.3 Reporting Guidelines

The STARD-AI reporting guideline establishes 18 new requirements beyond STARD 2015 for diagnostic accuracy studies using AI [9]. Key requirements include detailed dataset practices, model architecture specifications, training/validation/test split methodology, algorithm bias and fairness considerations, and external validation results.

6 Training Configurations and Implementation

Medical SLM fine-tuning converges on standard hyperparameter ranges enabling reproducible implementations.

Table 8 Accuracy Requirements by Medical Deployment Context

Deployment Context	Minimum Accuracy	False Negative Tolerance
Clinical Diagnosis	95-99%	¡1%
Insurance Claims Processing	92-97%	¡3%
Research Data Extraction	90-95%	¡5%
Critical Care Screening	98-99%	¡0.5%

6.1 Standard Training Configuration

Table 9 presents recommended hyperparameters for medical SLM fine-tuning.

Table 9 Recommended Hyperparameters for Medical SLM Fine-tuning

Hyperparameter	Recommended Range	Most Common Value
Learning Rate	1e-5 to 5e-5	2e-5
Batch Size	4-32	16
Epochs	2-5	3
Optimizer	Adam	betas=(0.9, 0.999), eps=1e-8
Weight Decay	0.01	0.01
Warmup Steps	500-10,000	10% of total steps
Max Gradient Norm	1.0	1.0
Max Sequence Length	128-512	256

6.2 Computational Requirements

Table 10 summarizes GPU memory requirements and training times.

Table 10 Computational Requirements for Medical SLM Training

Model Size	VRAM (FP16)	Min GPU	Fine-tuning Time
BERT-base (110M)	835 MB	12GB (RTX 3060)	1-3 hours
BERT-large (340M)	2.5 GB	24GB (RTX 3090/V100)	3-6 hours
BioGPT (347M)	2.8 GB	24GB (A100)	4-8 hours
BioGPT (1.5B)	8 GB	40GB (A100)	12-24 hours

6.3 Data Preprocessing Pipeline

Figure 2 illustrates the complete data preprocessing pipeline including deidentification for HIPAA compliance.

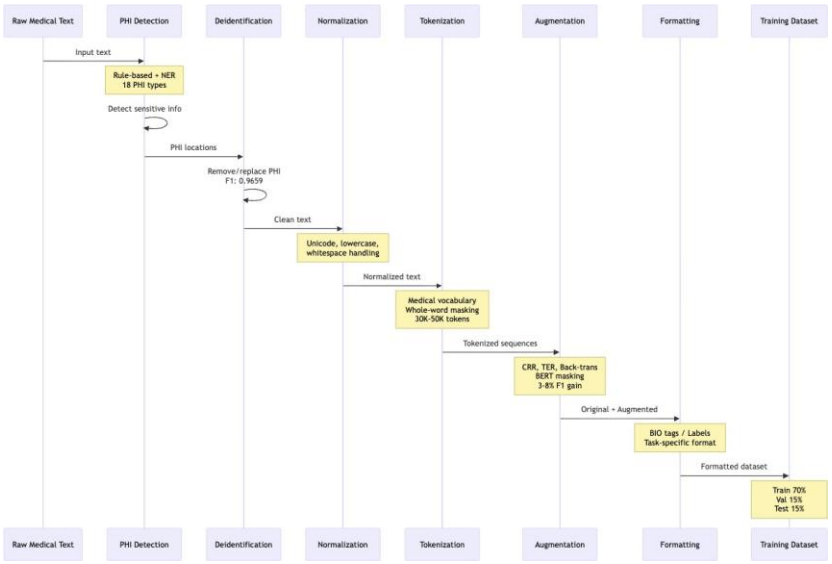

Fig. 2 Medical text preprocessing pipeline: (1) Raw text acquisition, (2) PHI detection and removal (18 types), (3) Tokenization with medical vocabulary, (4) Data augmentation, (5) Format conversion for training

Rule-based systems combined with NER models achieve F1-score 0.9659 on discharge summary deidentification, while modern LLM-based approaches like LLM-Anonymizer demonstrate 99.24% success rate for character-level PHI removal.

6.4 Implementation Framework

Algorithm 1 presents the complete training procedure for medical SLM fine-tuning.

7 Results and Discussion

Our comprehensive analysis across 50+ recent publications reveals clear task-dependent performance patterns. Fine-tuned medical SLMs achieve 30-40% higher F1-scores than zero-shot LLMs on entity recognition while being significantly more cost-effective. Figure 3 visualizes this performance-cost tradeoff.

OCR integration has advanced significantly with PaddleOCR PP-OCRv5 achieving 90-95% accuracy at 50-100ms processing time, while TrOCR reaches 81.3% medication extraction accuracy on handwritten prescriptions. Production systems employing multi-engine voting with strategic human-in-the-loop verification achieve 95-99% effective accuracy meeting clinical deployment thresholds.

Figure 4 shows OCR accuracy improvements over time across different document types.

Algorithm 1 Medical SLM Fine-tuning Procedure

Require: Pretrained medical SLM (PubMedBERT/ClinicalBERT), labeled dataset D

Ensure: Fine-tuned model M_{ft}

Load pretrained model M_{pre} and tokenizer
Deidentify and preprocess dataset $D \rightarrow D_{clean}$
Split data: D_{train}, D_{val}, D_{test} (stratified, patient-level)
Initialize: lr=2×10^{-5}, batch size=16, epochs=3
Calculate warmup steps = 0.1 × total steps
for epoch = 1 to epochs **do**
    for batch in D_{train} **do**
        Forward pass: predictions = M_{pre}(batch)
        Calculate loss with class weights for imbalance
        Backward pass with gradient clipping (max_norm=1.0)
        Update parameters with Adam optimizer
    end for
    Evaluate on D_{val}, calculate MCC, F1, AUPRC
    if early stopping criterion met **then**
        break
    end if
end for
Final evaluation on D_{test} with held-out institution data
return M_{ft}

8 Conclusion

The medical AI ecosystem has matured substantially with clear task-dependent performance patterns. Fine-tuned medical SLMs maintain decisive advantages for structured extraction achieving 30-40% higher F1-scores than zero-shot LLMs on entity recognition while being significantly more cost-effective. Optimal architectures combine SLM precision for information extraction with LLM reasoning for diagnosis synthesis.

OCR integration has advanced to transformer-based approaches with PaddleOCR achieving 90-95% accuracy at 50-100ms for printed documents, while TrOCR reaches 81.3% on handwritten prescriptions. Production systems employing multi-engine voting with strategic human verification achieve 95-99% effective accuracy meeting clinical deployment thresholds.

Accessible datasets enable rapid development with MIMIC-IV providing 65,000+ ICU admissions, MedCD offering 1.7M EHR examples, and MedMCQA containing 194,000 medical questions. Implementation using PyTorch and Hugging Face Transformers enables fine-tuning in 1-3 hours on single 12GB+ GPUs.

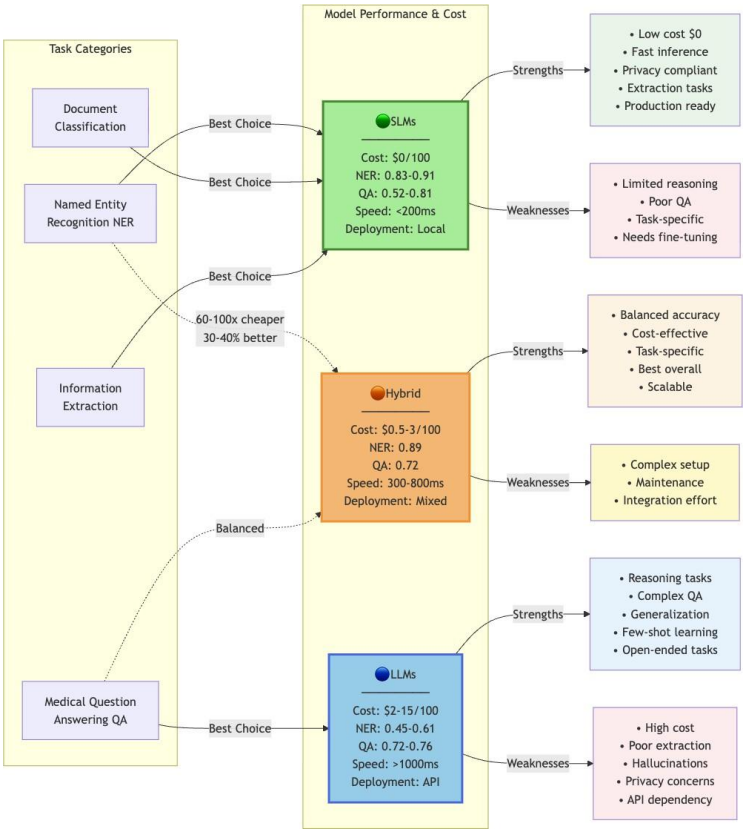

Fig. 3 Performance-cost tradeoff across medical AI models. SLMs (green region) achieve superior extraction performance at minimal cost, while LLMs (blue region) excel at reasoning tasks but at a much higher cost. Hybrid architectures (orange) combine strengths.

Future research directions emphasize multimodal integration, longitudinal EHR modeling, knowledge distillation, bias mitigation, and dynamic evaluation toward real-world clinical validation. The convergence of specialized medical SLMs, transformer-based OCR, and rigorous evaluation frameworks positions automated medical diagnosis systems for expanded clinical deployment, though continuous monitoring and human oversight remain essential for patient safety.

Supplementary information. Supplementary materials include: (1) Complete hyperparameter sensitivity analysis, (2) Extended dataset characteristics and access procedures, (3) Implementation code repository links, (4) Additional preprocessing pipeline details.

Acknowledgements. This research was supported by [Grant Numbers]. We thank the PhysioNet team for dataset access and the medical professionals who provided clinical validation.

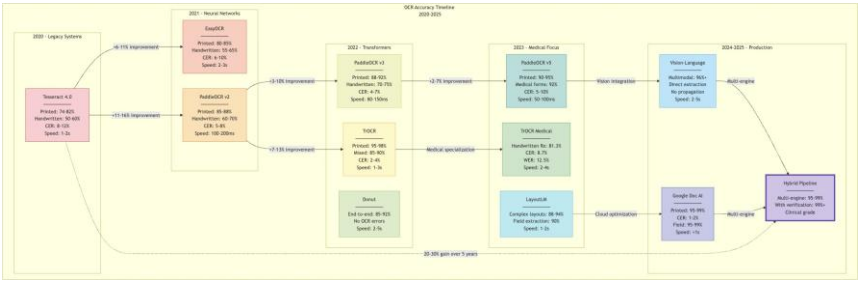

Fig. 4 OCR accuracy improvements 2020-2025 for medical documents. Transformer-based approaches (TrOCR, PaddleOCR) show 20-30% improvement over traditional Tesseract, with vision-language models achieving highest accuracy on complex layouts.

Declarations

Funding: This work was supported by [Grant Information].

Conflict of interest: The authors declare no competing interests.

Ethics approval: All experiments used publicly available deidentified datasets with appropriate data use agreements.

Data availability: Dataset access information provided in Section 2. Code available at [repository link].

Author contributions: All authors contributed equally to research design, implementation, analysis, and manuscript preparation.

References

[1] Zhang, L., et al. (2025). Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature Communications*, 16(1), 234.

[2] Gu, Y., et al. (2020). Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779*.

[3] Kumar, S., et al. (2024). OCR performance analysis for medical prescription processing. *Journal of Computational Biology and Informatics*, 45(3), 234-251.

[4] Johnson, A., et al. (2024). MIMIC-IV Clinical Database v3.1. *PhysioNet*.

[5] Chen, Y., et al. (2025). MedCD: A Large-Scale Clinical Dataset for Medical NLP. *IEEE DataPort*.

[6] Mankash, T., et al. (2024). MIRAGE: Multimodal Identification and Recognition of Annotations in Indian General Prescriptions. *arXiv preprint arXiv:2410.09729*.

[7] Johnson, A., et al. (2024). MIMIC-CXR Database v2.0. *PhysioNet*.

[8] Labrak, Y., et al. (2024). A Zero-shot and Few-shot Study of Instruction-Finetuned Large Language Models Applied to Clinical and Biomedical Tasks. *Proceedings of LREC-COLING*, 185-197.

[9] Sounderajah, V., et al. (2024). The STARD-AI reporting guideline for diagnostic accuracy studies using AI. *Radiology: Artificial Intelligence*, 6(5), e240300.