



Conversion of Physical Medical Records to Electronic Medical Records (EMRs)

Prof. Rushikesh Padaki,
Assistant Professor,
Department of AI & ML,
RV College of Engineering



Introduction

Go, change the world®

- Physical patient records (e.g., discharge summaries, prescriptions, lab reports) are often paper-based, unstructured, and fragmented.
- Electronic Medical Records (EMRs) provide standardized, searchable, and interoperable digital storage of patient data.
- Converting legacy physical records to EMRs ensures:
 - Improved **continuity of care**
 - Enhanced **data analytics and research capabilities**
 - Compliance with **national digital health initiatives** (e.g., NDHM, HIPAA, NHS Digital)
- **Challenges:** handwriting variability, document layout diversity, OCR errors, and interoperability.
- **Objective:** To design a robust, AI-driven methodology for digitizing and structuring physical medical records into interoperable EMR formats.



No.	Author(s) / Year	Publisher / Source	Focus / Contribution	Key Findings / Techniques Used
1	Bouh M.M. et al. (2024)	IEEE IECBES	Post-OCR correction in EMR digitization	Domain-specific spell correction and medical dictionary normalization improved OCR accuracy
2	Hsu E. et al. (2021)	JAMIA / Oxford	Deep learning-based NLP pipeline for scanned docs	Image preprocessing + layout-aware OCR + Transformer NER improved information extraction
3	Landolsi M.Y. (2023)	Springer	Survey on IE from EMRs	Hybrid ML + rule-based pipelines achieve best clinical accuracy
4	Murray L. (2021)	ACM	MedKnowts – unified EMR capture	Clinician-in-loop design improves data fidelity and usability
5	Mahadevkar S.V. (2024)	Springer (J. Big Data)	AI for unstructured medical documents	Hybrid Vision + NLP architectures handle diverse layouts effectively
6	Li Y. (2024)	Elsevier	Tabular data extraction from lab reports	Specialized table detectors outperform general OCR engines



No.	Author(s) / Year	Publisher / Source	Focus / Contribution	Key Findings / Techniques Used
7	Ren X. (2025)	PMC	Serialization of scanned lab reports	Layout recovery + privacy-preserving structured export improves archival quality
8	Ganzinger M. (2025)	PMC	LLM-based discharge summary generation	LLMs useful for summary generation once structured input is clean
9	Miake-Lye I.M. (2023)	Springer	Transition between EHR systems	Emphasizes governance, validation, and stakeholder training
10	Derecho K.C. (2024)	BMC	EMR adoption in developing nations	Implementation success tied to training and workflow redesign
11	Wu H. (2022)	Nature npj Digital Medicine	Survey of clinical NLP	Highlights domain adaptation and normalization to ontologies



Technical Gaps

- Existing OCR tools have **limited accuracy** for handwritten or poorly scanned medical records.
- Inconsistent document **layouts and formats** reduce model generalizability.
- Limited research on **multi-lingual EMR digitization** (especially for regional languages).
- Lack of **integrated pipelines** combining image, text, and structure recognition

Clinical and Semantic Gaps

- Insufficient use of **medical ontologies** (e.g., SNOMED CT, ICD-10, LOINC) for semantic normalization.
- Limited **error propagation studies** — few papers evaluate downstream impact of OCR/NLP errors on patient care.

Implementation Gaps

- **Human-in-loop validation** mechanisms are underdeveloped.
- Absence of **standard evaluation metrics** and **benchmark datasets** for scanned document conversion.
- **Interoperability gaps** — many systems do not conform to HL7 FHIR or openEHR standards.
- Lack of **change management frameworks** for large-scale deployment in hospitals.



Proposed Methodology

AI-Driven Methodology for Converting Physical Medical Records to EMRs

1 Project Planning & Data Governance

- Define scope, data types, and compliance standards (HIPAA / DHM)
- Establish data governance board and privacy policies

2 Data Collection & Scanning

- High-resolution scanning (>300 DPI) with proper metadata tagging
- Image preprocessing: de-skew, denoise, binarization using OpenCV/ImageMagick

3 Document Layout Analysis–OCR

- Hybrid OCR (Tesseract/Google Vision + domain dictionaries)
- Detect templates for common forms (e.g., TrOCR)

5 Post-OCR Handwritten Text Recognition

- Apply spell correction and medical dictionary normalization
- Align structured fields (demographics, vitals, lab)

6 NLP-Based Information Extraction

- ClinicalBERT / BioBERT for NER and Entity Recognition Extraction
- Map extracted entities to medical terminologies

7 Integration to EMR Standards

- Convert structured data into FHIR (Patient, Observation, Condition) or openEHR archetypes' via FHIR validators before ingest

8 Human-in-the-Loop Validation

- Clinicians verify high-risk fields (diagnosis, medication)
- Corrections fed back to retrain models (active learning loop)

9 Security & Deployment

- AES-256 Data encryption, role-based, audit-logging

Outcome Accurate, interoperable, and secure EMR conversion pipeline ready for large-scale deployment



RV College of
Engineering®

Go, change the world®

Thank You

A large, irregularly shaped teal watercolor wash serves as the background for the text 'Thank You'. The watercolor has a soft, blended appearance with lighter shades at the edges and darker, more saturated areas in the center. The text 'Thank You' is written in a black, cursive, sans-serif font, positioned centrally within the watercolor shape.