Paper-Based Health Records: A Case Study on the Digitization of Handwritten Clinical Records

Vincenza Carchiolo[a], Michele Malgeri[b] and Lorenzo Spadaro Sapari
Dipartimento di Ingegneria Elettrica Elettronica e Informatica Università di Catania, Catania, Italy

Keywords: Health Management, OCR, Application.

Abstract: This paper presents a case study focused on the application of handwriting recognition to digitize historical clinical records containing significant handwritten content. The primary objective is to assess the feasibility of using commercial OCR technologies—in particular, Microsoft Azure's handwriting recognition API—for processing health documents. The study aims to determine whether these tools can support the extraction of meaningful clinical information, not only by recognizing individual characters but also by leveraging the structural layout of documents, such as forms, to infer semantic content.
Our methodology includes empirical evaluation of OCR output on real-world patient records, alongside a qualitative analysis of common recognition errors. In addition, we review relevant approaches from the literature, highlighting recent advances in deep learning for document understanding. The findings indicate that general-purpose OCR systems are currently insufficient for reliable clinical data extraction in such contexts, primarily due to the complexity and variability of handwritten medical records. However, the results also suggest that structural cues present in form-based documents could be harnessed—through tailored AI-based techniques—to significantly improve recognition and downstream information retrieval.

1 INTRODUCTION

In recent years, healthcare systems have undergone an accelerated digital transformation, with the goal of improving data accessibility, interoperability, and analytical capabilities. However, despite the proliferation of Electronic Health Record (EHR) systems, a significant portion of clinical information remains trapped in non-digitized formats. These include scanned paper records, printed reports, handwritten notes, and administrative forms. The presence of such unstructured data limits the potential of modern healthcare information systems to provide timely and data-driven insights. The problem is particularly severe in hospital settings, where documentation practices often vary across departments and time periods. Clinical records are typically long and detailed, encompassing a wide range of information from patient demographics to complex diagnostic descriptions, therapeutic plans, and procedural notes. These documents frequently include a combination of printed and handwritten text, non-standardized layouts, medical jargon, and institution-specific abbreviations. The lack of uniformity not only hinders human readability but also poses significant challenges for automated processing. Within the European Union, efforts are being made to harmonize the healthcare data landscape. Initiatives such as the European Health Data Space (EHDS), officially launched in 2025, promote secure cross-border data exchange and aim to facilitate the secondary use of health data for research and innovation. Standards such as HL7's Clinical Document Architecture (CDA) and Fast Healthcare Interoperability Resources (FHIR) are increasingly adopted to enable interoperability between systems. Nonetheless, a large portion of legacy documents predates these standards and exists only in paper or scanned format, making them inaccessible to modern digital workflows.

To bridge this gap, Optical Character Recognition (OCR) is still a foundational technology, also taking into account the chance to integrate AI algorithms to enhance the ability to recognize more contents. OCR enables the automatic conversion of scanned images or PDF documents into machine-readable text, allowing historical and unstructured records to become accessible for further analysis. However, applying OCR

[a] https://orcid.org/0000-0002-1671-840X
[b] https://orcid.org/0000-0002-9279-3129

Carchiolo, V., Malgeri, M. and Sapari, L. S.
Paper-Based Health Records: A Case Study on the Digitization of Handwritten Clinical Records.
DOI: 10.5220/0013853900003985

In *Proceedings of the 21st International Conference on Web Information Systems and Technologies (WEBIST 2025)*, pages 244-251
ISBN: 978-989-758-772-6; ISSN: 2184-3252

in the healthcare domain is far from trivial. Clinical documents differ significantly from standard printed text in terms of complexity, content density, and variability. Handwriting recognition remains particularly difficult, especially when combined with poor scan quality or domain-specific terminology. Additionally, the presence of tables, multiple columns, and mixed formatting adds another layer of complexity for OCR systems to handle.

In this work, we present a case study focused on the extraction of structured information from complex clinical documents using a pipeline based on Microsoft Azure Document Intelligence (Microsoft, 2025), part of the more general Microsoft Azure Cognitive Services (Microsoft, na). The proposed solution integrates multiple tools from the Microsoft ecosystem, including the Read OCR API for text extraction, the Layout API for document structure analysis, and custom modules for medical entity recognition and normalization. These tools are orchestrated in a modular workflow designed to cope with heterogeneous document types, allowing for preprocessing, layout-aware recognition, and post-OCR analysis. The pipeline was applied to a dataset of anonymized Italian clinical records collected from a hospital environment. These documents reflect the typical diversity of healthcare records: they include admission reports, discharge summaries, and intraoperative notes—many of which contain mixed handwritten and typed sections. The evaluation focused both on the quality of text recognition (e.g., character error rates, text segmentation) and on the ability to extract key information such as patient identifiers, diagnoses, and timestamps. Beyond digitization, a key contribution of this work lies in positioning OCR as a critical enabling step for advanced data analysis. Once clinical text has been extracted, it can serve as input for a wide range of artificial intelligence (AI) applications, such as natural language processing (NLP), named entity recognition (NER), temporal reasoning, and predictive modeling (Carchiolo and Malgeri, 2025). In particular, the ability to transform unstructured clinical narratives into structured data opens the door to more sophisticated tools for clinical decision support, cohort identification, risk stratification, and automated report summarization. Although OCR alone does not solve the full problem of semantic understanding, it provides the essential first layer of machine interpretability. The combination of OCR and AI-driven post-processing can help unlock the latent value stored in years of handwritten or non-standard documentation, contributing to the broader goal of modernizing healthcare information systems and improving data-driven patient care. In summary, this study offers a realistic and scalable approach to document digitization in clinical contexts using Microsoft-based OCR technologies. It illustrates both the current potential and the limitations of applying these tools in real-world hospital settings, and lays the groundwork for future integration with AI-powered healthcare analytics pipelines.

The remainder of this paper is organized as follows: Section 2 describes the health records and the related standards, if any. Section 3 details the OCR pipeline, including the tools and methods adopted for preprocessing, recognition, and post-processing. Section 4 presents our proposal giving details about architecture models and whatever has been studied and section 5 discusses the findings. Finally, Section 6 concludes the paper and outlines directions for future work.

2 ABOUT CLINICAL RECORDS

In the European Union, health records are central documents for both healthcare delivery and medico-legal accountability. While there is no binding European regulation that imposes a uniform structure or content for health records across all member states, multiple initiatives and technical standards have been introduced to promote interoperability, data quality, and security. The most significant initiative in this regard is the European Health Data Space (EHDS), proposed by the European Commission in 2022 and officially entered into force in March 2025 (European Commission, 2025b). Its goal is to facilitate the secure exchange and use of health data across the EU, including electronic health records (EHRs), while respecting patient privacy and national healthcare governance structures. Technical interoperability efforts are also supported by the eHealth Network, a voluntary collaboration between EU countries, which has produced common specifications for cross-border health data exchange, particularly in the form of “Patient Summaries” and “ePrescriptions” (European Commission, 2025a). At the technical level, several international standards developed by Health Level Seven International (HL7) have been increasingly adopted across Europe. These include the Clinical Document Architecture (CDA), used to define the structure of clinical documents such as discharge summaries, and in a more recent version (HL7 FHIR) includes Fast Healthcare Interoperability Resources (FHIR), the standard that facilitates the exchange of healthcare information across systems using modern web technologies (Bender and Sartipi, 2013). Among these, FHIR has emerged as the pre-

ferred standard for modern EHR systems due to its compatibility with RESTful APIs. The EHDS initiative aims to unify FHIR standards across member states, targeting 80% adoption by 2026 (Willis, 2025). This initiative underscores the EU's commitment to enhancing healthcare interoperability, improving patient care, and facilitating data-driven research and innovation.

Despite these harmonization efforts, each country retains significant autonomy in determining the mandatory content and structure of health records. In Italy, for example, the clinical record is recognized as both a medical and legal document that accompanies the entire hospitalization episode and documents the diagnostic and therapeutic process in a traceable and continuous manner. It is compiled and maintained primarily by the attending physician, in compliance with various legal, ethical, and procedural standards.

A legal foundation for the clinical record can be found in a combination of sources. The Decree of August 5, 1977 (Ministero della Sanità, 1977), requires that private healthcare institutions compile a medical record for each hospitalized patient, containing full personal data, initial and final diagnoses, family and personal medical history, objective examinations, laboratory and specialist tests, therapy, outcomes, and post-treatment status. These records must be signed by the treating physician and archived by the healthcare facility. The Ministry of Health Guidelines of June 17, 1992, concerning the management of hospital discharge forms (Scheda di Dimissione Ospedaliera), describe the clinical record as an individual information tool that documents all relevant demographic and clinical data related to a single episode of hospitalization, from admission to discharge, effectively representing the patient's entire stay in the hospital. Ethical obligations are further codified in the 2014 Italian Code of Medical Ethics, where Article 26 specifies that the clinical record must be compiled with completeness, clarity, diligence, and in a timely manner (Federazione Nazionale Ordine Medici Chirurghi ed Odontoiatri, 2014). It must record both objective and subjective clinical data, details of diagnostic and therapeutic procedures, informed consent or dissent—including for sensitive data processing—and any advance care planning, particularly for patients with progressive illnesses. The ethical code also mandates the traceability of all entries and corrections, underscoring the importance of documentation integrity. Italian jurisprudence reinforces this view by defining the clinical record as a diagnostic-therapeutic diary in which all information of medical and legal relevance must be accurately recorded. This includes the patient's personal and medical history, diagnostic evaluations, treatments administered, clinical evolution, outcomes, and any lasting consequences of the illness.

Structurally, a typical Italian hospital-based health record includes: Administrative and demographic information (such as patient ID, hospital unit, date and mode of admission), admission diagnosis and presenting complaints, medical and nursing notes (a chronological log of observations, decisions, and care delivered), diagnostic results (including laboratory tests and imaging reports), specialist consultations and interdisciplinary opinions, pharmacological and therapeutic prescriptions, surgical and anesthesiology documentation (when applicable), informed consent forms and ethical disclosures, *sheda di Dimissione Ospedaliera* (SDO)[1], which uses the ICD-9-CM (World Health Organization, 2025) standard for coding health conditions, mainly in specification of Chronic and/or relevant pathologies of the patient and coded representation of all known pathologies in progress at the time of filling out the document, and discharge summary (final diagnosis, summary of care, and follow-up recommendations).

The Fascicolo Sanitario Elettronico (FSE), Italy's national digital health record platform, has further standardized the collection and availability of this information. As part of the European interoperability effort, the FSE is being progressively integrated with HL7 FHIR standards to support secure, structured, and cross-provider data sharing, as reported in (Agenzia nazionale per i servizi sanitari regionali, 2023).

3 OCR AND INFORMATION EXTRACTION

In the healthcare domain, legacy documents often exist in scanned or handwritten formats, including printed reports, PDF files, and clinical notes, often handwritten. In (White-Dzuro et al., 2021) highlight the practical difficulties faced during the COVID-19 pandemic, when large volumes of handwritten forms and non-standardized clinical records had to be processed rapidly. Their findings show that even modern OCR systems struggle with the low quality of input scans, domain-specific abbreviations, and the lack of consistent formatting. In (Wang et al., 2023), the authors examine various deep learning-based algorithms for text detection and recognition, providing insights into their methodologies and applications.

[1]"Hospital Discharge Summary": This form is an official medical document issued by a hospital at the time of a patient's discharge.

Given these challenges, modern OCR platforms—such as Google Cloud Vision (Google Cloud Platform, 2016), Amazon Textract (Amazon Web Services, 2019), and Microsoft OCR's Azure Cognitive Services (Microsoft, na) have incorporated AI-based modules to enhance the detection of text regions, interpretation of complex layout structures, and recognition accuracy across diverse document types. Google Cloud Vision offers broad support for multilingual OCR and layout detection; Amazon Textract emphasizes structured data extraction, including tables and forms; while Azure OCR integrates seamlessly with other cognitive APIs for enhanced document analysis. While these platforms mark a significant advancement in general-purpose OCR, they often struggle with domain-specific applications—such as extracting clinically relevant content from electronic health records (EHRs), pathology reports, or discharge summaries—where specialized tools tailored to the medical domain can provide more accurate and context-aware results. One example of said specialized tools is the DEXTER system (Nandhinee et al., 2022) which presents a complete pipeline for extracting tabular content from electronic medical records. By combining deep learning-based table detection with conventional vision techniques for cell segmentation it achieves great results on real-world medical datasets. In a more recent study, the authors of (Li et al., 2024) proposed a deep learning-based OCR pipeline specifically designed for scanned laboratory reports. Their system integrates advanced models such as Detection Transformer (DETR) R18 for table detection and an encoder-dual-decoder (EDD) architecture for table recognition. The study also emphasizes the challenges posed by document noise, handwritten notes, and diverse table formats commonly found in medical records—issues that general-purpose OCR tools often fail to address.

4 WHAT WE DID

Retrieving medical records is often a complex task for patients due to fragmentation across systems and inconsistent formats. This section presents the core contribution of this work and it introduces the architecture and implementation details of the proposed system, which aims to extract structured information from medical reports containing both printed and handwritten text. The system leverages a custom-trained OCR model and a large language model(Carchiolo et al., 2026)to support accurate and efficient retrieval of documents based on user input, this enables patients to access their own medical records with minimal effort.

4.1 System Architecture

The proposed system adopts an OCR-based pipeline designed to process anonymized medical reports containing both printed and handwritten text. The goal is to extract structured information from these documents and enable retrieval through user interaction. The OCR component is implemented using Microsoft Azure Document Intelligence (Microsoft, 2025), a cloud-based service that allows for custom model training tailored to specific document layouts. The information gathered from the OCR component is then leveraged by an LLM, namely Mixtral 8x7B (Jiang et al., 2024), to guide the user in the retrieval of the medical report he's looking for. The interaction with the user is carried out through a web-based conversational interface, where the language model dynamically adapts its queries based on the user's previous answers. If multiple reports match the provided criteria, the model refines the search by asking additional, targeted questions. The high-level workflow of the system comprises the following steps:

(a) Medical reports are fetched from the relevant medical database and processed by the custom OCR model.
(b) Extracted key-value pairs are used to build a structured index for each report.
(c) The Mixtral 8x7B model interprets the user's natural language requests via a brief conversational exchange. This assists the user in filtering the desired report(s) from potentially many available documents.
(d) The system uses the criteria derived from the interpreted user request to search the structured indices and identify matching reports.
(e) Once some matches are found, the system provides direct links to the corresponding documents.

This process is designed to operate entirely in the cloud. User login and authentication are handled through SPID (Sistema Pubblico di Identità Digitale), the Italian digital identity system (AgID, 2020). This allows patients to securely access the system via authorized medical platforms, such as the ones provided by hospital companies. By leveraging SPID, the system obtains the necessary personal information to perform a precise query on the medical database, retrieve all the user's reports for indexing, and subsequently facilitate secure access.

4.2 Model Training

Recognizing the need to handle specific structural nuances present in the medical reports, a custom OCR model was developed rather than relying solely on a generic pre-trained solution. Although the reports generally adhered to a common template regarding the approximate placement of information, significant variations existed between documents. For instance, the same logical field (such as the hospital unit) might appear as printed text in one report and as handwritten text in another, albeit typically within the same region of the page. The resulting model is capable of identifying key clinical data fields across the page, effectively processing regions containing both printed and handwritten text, irrespective of these minor inconsistencies.

To train and evaluate this custom OCR model, a dedicated dataset was meticulously gathered and prepared. A corpus of sixteen medical reports, was collected from different units of the general hospital in Catania. Crucially, prior to any processing, all documents underwent a rigorous anonymization procedure to remove any patient's personal information (such as names, addresses, fiscal codes) and any other potentially identifying information. This step was performed in strict compliance with Europe's General Data Protection Regulation (GDPR) (Proton Technologies AG (GDPR.EU), 2018) requirements, ensuring patient privacy was paramount. After having anonymized the entire dataset, it was split into distinct sets for training and testing, a standard 60/40 split ratio was applied: 10 documents were allocated for the training set and the 6 remaining for the testing set. Then, a detailed annotation phase was undertaken, using the Microsoft Document Intelligence Studio UI, to label the regions of interest within each document. This phase consisted of defining precise bounding boxes around each target key field and naming such field.

From these documents, the following information can be extracted:

- Name and City of the Hospital that produced the report
- Patient's residence
- Date of admission and discharge from the hospital
- Diagnosis of admission and discharge of the patient.

The OCR system returns information in the form of key-value pairs, such as the ones represented in table 1. The output of the OCR phase is a structured dictionary that represents the essential metadata of each medical report.

Table 1: Output sample.

Key	Value
hospital	Azienda Ospedaliera ...
city	Catania
residence	Palermo
admission_date	01/01/2025
discharge_date	01/01/2025
admission_diagnosis	Pneumonia
discharge_diagnosis	Pneumonia

Unlike other systems that rely on separate parsing modules or NER (Named Entity Recognition) pipelines to interpret and extract data from raw OCR text such as the ones in (Rasmussen et al., 2012),(Tan et al., 2022) and (Karthikeyan et al., 2022), the proposed approach benefits from the native structured output of the custom OCR model. Since key-value data is extracted directly, no additional parsing or information extraction steps are required. This architecture reduces processing complexity and improves response time. Moreover, the structured format facilitates accurate comparison with user input, improving the overall effectiveness of the retrieval process.

4.3 Effect of OCR Limitations on the System

While the proposed system demonstrates high performance on documents following the trained layout, several limitations and assumptions constrain its generalization capabilities. The custom OCR model was trained on a set of medical reports with a fixed layout, however due to the absence of standardized formatting among medical institutions, supporting additional report types would likely require dedicated retraining.

The recognition of handwritten text remains highly dependent on the legibility of the handwriting itself. In favorable cases, the model achieved a Word Error Rate (WER) of 0% and a Character Error Rate (CER) as low as 2%. Importantly, no fields were consistently more error-prone than others, indicating uniform performance across the page. In cases where a field is missing or illegible, the system logs the corresponding key in the structured index with the placeholder value "not found". This mechanism ensures that downstream processes, namely comparison and retrieval, can proceed without exceptions or crashes due to missing fields. Scalability of the solution was not tested extensively due to resource constraints; the experimental evaluation was limited to a dataset of 16 reports. While the approach is expected to scale linearly with the number of documents, further experimentation on larger datasets is needed to validate this assumption.

Current latency for user interaction, including OCR inference and response generation via the Mixtral 8x7B model, ranges from 5 to 15 seconds, which is acceptable for interactive use. Finally, data privacy is a critical consideration, particularly in medical applications. Microsoft Azure's documentation specifies that files processed through Document Intelligence are temporarily stored on their servers for up to seven days, presumably for caching and performance optimization.

5 TESTING

The testing phase was designed to evaluate the performance of the OCR component in extracting structured information from medical documents containing both printed and handwritten text. The source material consisted of full medical records, each spanning several hundred pages and originating from the same healthcare institution, all anonymized to preserve patient confidentiality. For the purpose of both training and evaluation, only the first page of each report was used, as it consistently contains the key administrative and clinical fields targeted by the system (e.g., hospital name, admission date, diagnosis, admission unit). This approach allowed the creation of a representative and manageable dataset without compromising the diversity of layout and content necessary for robust model evaluation. The OCR engine employed was a custom-trained model developed using Microsoft Azure Document Intelligence. Its output is a hashmap of key-value pairs representing the structured fields extracted from the page, eliminating the need for additional post-processing.

To complement the quantitative evaluation, Figure 1 presents a visual example of the OCR system's output on one of the test documents. The figure shows the scanned page overlaid with bounding boxes corresponding to the detected key-value pairs extracted by the custom model. Each bounding box encloses either a field label or its associated value, indicating the model's interpretation of the document structure.

The image illustrates the mixed nature of the content, which includes both machine-printed and handwritten entries. Printed fields like the document header are generally recognized with high accuracy. Conversely, handwritten content, particularly in the fields "Diagnosi di ingresso" (admission diagnosis) and "Diagnosi di dimissione" (discharge diagnosis), presents more variability due to inconsistent handwriting styles and legibility. These challenges are manifested in the error rates discussed later in this section.

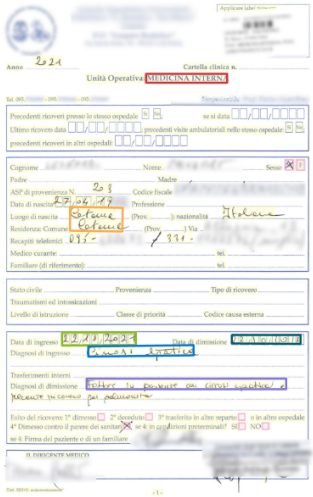

Figure 1: OCR result with bounding boxes showing extracted fields from a test medical report.

Evaluation focused on the accuracy of text recognition. Given the presence of both machine-printed and handwritten content, two standard metrics were adopted: Character Error Rate (CER) and Word Error Rate (WER). These were computed using the JiWer Python library for each test document and then averaged. This allowed quantification of both fine-grained errors (e.g., character-level substitutions) and larger semantic discrepancies (e.g., missing or misinterpreted words). The test set comprised 51 representative documents. Figure 2 illustrates the distribution of recognition errors across the document set. Specifically, it presents the percentage of documents grouped by error rate intervals of 10%, using both Character Error Rate (CER) and Word Error Rate (WER) as evaluation metrics. This binning approach enables a clearer understanding of how frequently different levels of error occur, highlighting the prevalence and severity of recognition issues within the dataset. The results show a wide range in recognition performance, primarily due to differences in handwriting legibility. In documents where printed text

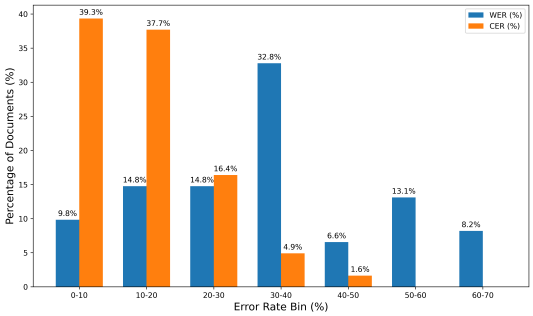

Figure 2: WER and CER. The figure highlight the error percentage vs the error classes.

dominated error rates yielded a WER of 10.00% and a CER 1.23%. Conversely, other documents saw WERs exceeding 60%. Nevertheless, the average values of WER, that is 37.43% and CER, 14.27%, reflect an acceptable baseline for mixed-content recognition in a real-world setting. Can be observed that the higher error values are primarily attributable to the variability in the handwriting of the document authors, which introduces significant inconsistencies in the graphical representation of characters. Furthermore, a minor portion of the discrepancies could stem from differences in document layout compared to those used during the training phase, suggesting that model generalization to previously unseen document structures remains an area for improvement.

This is exemplified in Figure 3a, which shows the OCR results for the worst-performing case among the 51 documents tested. Figure 3b exemplifies the challenges posed by poor handwriting in document recognition. In this case, the model failed to accurately extract the dimission diagnosis field (highlighted by the brown bounding box), resulting in a Word Error Rate (WER) of 67.57% and a Character Error Rate (CER) of 22.04%. Despite correctly locating all target fields, the model still produced a WER of 66.67% and a CER of 19.44%, underscoring the significant impact of handwriting legibility on recognition performance.

(a) (b)

Figure 3: OCR results for worst performing (left) and handwriting problems.

Figure 4 shows the document that achieved the best results, with a WER of 10.00% and a CER of 1.23%. The model successfully extracted all target fields, and due to the clarity of both the document and handwriting, only minor character-level misinterpretations were observed. Such calligraphic heterogeneity poses a substantial obstacle for automated text recognition systems, reducing model accuracy even with robust training.

In all test cases, the system successfully returned structured output. On average, the processing time for each document was under five seconds using the Azure cloud infrastructure. While no local engine was benchmarked for comparison, the cloud-based setup enabled rapid iteration and ensured scalability. Finally, end-to-end tests confirmed that the chatbot-assisted retrieval system correctly identified the intended medical record when the user input matched the fields extracted by the OCR module. Final validation was performed by the user through a binary confirmation ("yes" or "no"), reinforcing the system's effectiveness under realistic usage conditions.

Figure 4: OCR result of document #8.

6 CONCLUSIONS

This study investigated the effectiveness of a commercial OCR solution, specifically Microsoft Azure's handwriting recognition, in processing real-world clinical records that include substantial handwritten content. The goal was to evaluate whether existing general-purpose OCR technologies are suitable for extracting meaningful and structured information from historical patient documentation. The methodology combined both quantitative metrics and manual inspection to assess recognition quality and semantic coherence in the output.

The results indicate that current off-the-shelf OCR systems, while offering basic recognition capabilities, often fail to provide sufficient accuracy for downstream processing, particularly in highly domain-specific and variable contexts like handwritten medical forms. Issues such as fragmented recognition, loss of document structure, and confusion in domain-specific terminology were recurrent.

In parallel, the paper reviewed state-of-the-art approaches and recent research in form understanding and handwriting recognition, which suggest that hybrid and AI-enhanced techniques—such as the integration of contextual models, semantic parsing, or domain-specific post-processing—could offer significant improvements over current commercial tools.

Ultimately, this case study underscores the need for accurate and robust OCR technologies as a foun-

dational component in any pipeline that aims to leverage artificial intelligence for clinical document analysis. Without reliable text extraction, the potential of AI to derive insights from vast volumes of archived handwritten data remains largely untapped. Future work should explore the integration of domain-trained models, active learning strategies, and multimodal document representations to improve recognition accuracy and usability in medical and archival settings.

REFERENCES

Agenzia nazionale per i servizi sanitari regionali (2023). Piattaforma di telemedicina e FSE. https://www.agenas.gov.it/comunicazione/primo-piano/2090-piattaforma-telemedicina-fse. Accessed 8-May-2025].

AgID (2020). Sistema pubblico di identità digitale. https://www.spid.gov.it. Accessed 17-April-2025.

Amazon Web Services (2019). Amazon Textract – Fully Managed ML for Text and Data Extraction. https://docs.aws.amazon.com/textract/. General Availability announced November 28, 2018; service available from May 2019, accessed on 2025-05-15.

Bender, D. and Sartipi, K. (2013). HL7 FHIR: An agile and RESTful approach to healthcare information exchange. In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, pages 326–331.

Carchiolo, V. and Malgeri, M. (2025). Trends, challenges, and applications of large language models in healthcare: A bibliometric and scoping review. *Future Internet*, 17(2).

Carchiolo, V., Malgeri, M., and Sapari, L. S. (2026). A conversational agent for handling health report inquiries. *Communications in Computer and Information Science*, 2518 CCIS:202 – 211.

European Commission (2025a). eHealth network. https://health.ec.europa.eu/ehealth-digital-health-and-care/digital-health-and-care/eu-cooperation/ehealth-network. Accessed 07-May-2025.

European Commission (2025b). European health data space regulation (EHDS). https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds. Accessed: 2025-05-07.

Federazione Nazionale Ordine Medici Chirurghi ed Odontoiatri (2014). Nuovo codice di deontologia medica. https://www.health-management.it/codice_dentologico/cdm_03_25_26.htm.

Google Cloud Platform (2016). Google Cloud Vision API. https://cloud.google.com/vision. Accessed: 2025-05-15.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2024). Mixtral of experts.

Karthikeyan, S., de Herrera, A. G. S., Doctor, F., and Mirza, A. (2022). An OCR post-correction approach using deep learning for processing medical reports. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):2574–2581.

Li, Y., Wei, Q., Chen, X., Li, J., Tao, C., and Xu, H. (2024). Improving tabular data extraction in scanned laboratory reports using deep learning models. *Journal of Biomedical Informatics*, 159:104735.

Microsoft (2025). What is azure AI document intelligence? https://learn.microsoft.com/en-us/azure/ai-services/document-intelligence/overview?view=doc-intel-4.0.0". Accessed 1-April-2025.

Microsoft (n.a.). Azure cognitive services — computer vision ocr documentation. https://learn.microsoft.com/en-us/azure/cognitive-services/computer-vision/concept-recognizing-text. Accessed: 2025-05-15.

Ministero della Sanità (1977). Determinazione dei requisiti tecnici sulle case di cura private. http://architettura.it/notes/ns_nazionale/anno_70-79/D.M.5-8-77.html.

Nandhinee, P., Harinath, K., Koushik, S., Anil, G., and Sudarsun, S. (2022). DEXTER: An end-to-end system to extract table contents from electronic medical health documents. arXiv preprint arXiv:2207.06823. Available at: https://arxiv.org/abs/2207.06823.

Proton Technologies AG (GDPR.EU) (2018). General data protection regulation (GDPR). https://gdpr.eu/tag/gdpr. Accessed 17-April-2025.

Rasmussen, L. V., Peissig, P. L., McCarty, C. A., and Starren, J. (2012). Development of an optical character recognition pipeline for handwritten form fields from an electronic health record. *Journal of the American Medical Informatics Association*, 19(e1):e90–e95.

Tan, Y. F., Connie, T., Goh, M. K. O., and Teoh, A. B. J. (2022). A pipeline approach to context-aware handwritten text recognition. *Applied Sciences*, 12(4).

Wang, X.-F., He, Z.-H., Wang, K., Wang, Y.-F., Zou, L., and Wu, Z.-Z. (2023). A survey of text detection and recognition algorithms based on deep learning technology. *Neurocomputing*, 556:126702.

White-Dzuro, C. G., Schultz, J. D., Ye, C., Coco, J. R., Myers, J. M., Shackelford, C., Rosenbloom, S. T., and Fabbri, D. (2021). Extracting medical information from paper COVID-19 assessment forms. *Applied Clinical Informatics*, 12(1):170–178.

Willis, N. (2025). IFHIR adoption statistics in 2025: A global overview. https://www.linuxactionshow.com/fhir-adoption-statistics-in-2025-a-global-overview. Accessed 8-May-2025].

World Health Organization (2025). International statistical classification of diseases and related health problems (ICD). https://www.who.int/classifications/classification-of-diseases. Accessed 8-May-2025.