DOI: 10.20998/2079-0023.2026.01.14
UDC 004.89:61

M. A. GRINCHENKO, Candidate of Technical Sciences (PhD), Docent, National Technical University "Kharkiv Polytechnic Institute", Head of the Department of Project Management In Information Technologies, Kharkiv, Ukraine, e-mail: marina.grynchenko@khpi.edu.ua, ORCID: https://orcid.org/0000-0002-8383-2675
D. O. KUTSENKO, graduate student of the Department of Project Management In Information Technologies, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine,
e-mail: Dmytro.O.Kutsenko@cs.khpi.edu.ua, ORCID: http://orcid.org/0009-0005-6359-3143

COMPARATIVE STUDY OF TRANSFORMER-BASED AND INTELLIGENT DOCUMENT ANALYSIS METHODS FOR AUTOMATED EXTRACTION OF MEDICAL DATA FROM PDF DOCUMENTS

This paper presents a study on automated processing of medical laboratory reports in PDF format, with a focus on text recognition and structured information extraction. The research investigates the effectiveness of different approaches to optical character recognition (OCR), including classical methods and transformer-based models, as well as techniques for extracting key medical data from unstructured and semi-structured text. A comparative experimental analysis was conducted using medical documents with different structural characteristics, including tabular and text-based formats. The study evaluates the performance of OCR methods and extraction pipelines using a set of quantitative metrics, including Character Error Rate (CER), Word Error Rate (WER), Exact Match (EM), Precision, Recall, and F1-score. The obtained results demonstrate that OCR accuracy alone does not guarantee high-quality structured data extraction, as recognition errors significantly affect downstream processing and reduce the reliability of extracted information. Special attention is given to layout-aware approaches that utilize the structural properties of PDF documents. The proposed method based on direct text extraction using pdfplumber shows superior performance by preserving spatial relationships between document elements and eliminating the need for OCR in documents with an embedded text layer. This approach ensures higher stability and accuracy when processing structured medical data. The findings highlight that the main challenge in processing medical documents lies in the extraction stage rather than in text recognition. The study demonstrates the importance of integrating layout-aware and intelligent extraction methods for improving the reliability, robustness, and scalability of automated data processing systems. The proposed approach can be used as a foundation for developing medical information systems and decision support tools aimed at efficient and accurate clinical data management.

Keywords: optical character recognition, information extraction, medical documents, medical data processing, layout-aware methods, data extraction, document analysis, decision support systems, artificial intelligence tools.

Introduction. In the context of the rapid digitalization of healthcare systems, the efficient processing of medical documentation has become a critical task for modern medical institutions. Laboratory test results represent an essential component of clinical information and are widely used for diagnosis, treatment planning, and monitoring of patient conditions. In practice, such results are often distributed in the form of PDF documents, which are convenient for human reading but difficult to process automatically due to their unstructured or semi-structured nature.

The increasing volume of medical data and the need for its integration into clinical information systems and decision support tools significantly intensify the demand for automated document analysis methods. Manual data entry remains a common practice; however, it is time-consuming, error-prone, and does not scale efficiently in real-world healthcare environments. Additionally, the absence of standardized formats for laboratory reports leads to substantial variability in document structure, layout, and content representation, which further complicates automated processing.

Recent advances in the field of intelligent document processing, including optical character recognition (OCR) and machine learning-based information extraction methods, have opened new opportunities for addressing these challenges. In particular, transformer-based models have demonstrated high performance in text recognition tasks, while layout-aware approaches enable the incorporation of spatial information into document understanding. At the same time, traditional rule-based extraction methods remain widely used due to their simplicity and interpretability, despite their limited ability to generalize across heterogeneous document formats.

In this context, the problem of automated extraction of structured data from medical PDF documents requires a comprehensive analysis of different methodological approaches. It is important to evaluate not only the quality of text recognition but also the impact of OCR output on subsequent information extraction stages. Therefore, the present study focuses on the comparative analysis of transformer-based OCR methods, classical recognition approaches, and layout-aware document analysis techniques for the automated processing of medical laboratory documents.

Analysis of research and publications. Recent studies indicate that automated processing of medical documentation is increasingly approached as an integration of natural language processing, optical character recognition, and intelligent information extraction methods. In the review by Patil and Golbhavi [1], healthcare NLP is described as a rapidly evolving field that has moved from symbolic and statistical methods toward deep learning and transformer-based architectures. Similar conclusions are presented in [14], where transformer-based NLP techniques are shown to improve the extraction of clinically relevant entities and the analysis of unstructured medical text. The strategic role of NLP in medical documentation is also emphasized in [19].

A substantial body of research focuses on Named Entity Recognition as a core mechanism for transforming unstructured medical text into structured data. In [4], the

authors show that transformer-based NER models consistently outperform traditional rule-based and statistical approaches in healthcare text processing. In [5], a BERT-based end-to-end framework for NER and relation extraction in electronic medical records is proposed, while [6] presents a hybrid BiLSTM-BERT-RAG architecture for medical NER and classification. Additional evidence of the effectiveness of transformer-based models for heterogeneous medical records is reported in [25]. Together, these studies confirm the strong potential of neural approaches for extracting semantically meaningful entities from medical documentation.

Another important research direction concerns multilingual and low-resource clinical information extraction. In [3], prompt-based large language models are applied to multilingual biomedical NER and entity linking, demonstrating strong performance even with limited annotated resources. This issue is particularly relevant for non-English medical documentation, including Ukrainian-language healthcare texts. At the same time, studies such as [2] indicate that even open large language models with limited computational requirements can provide practically usable medical text generation and documentation support.

Significant attention is also given to direct extraction of information from medical reports containing complex layouts, tables, numerical values, units, and reference ranges. In [7], an end-to-end OCR and information extraction pipeline is proposed for laboratory reports, combining line normalization, NER, and multicolumn analysis. In [8], OCR-induced noise in medical records is analyzed in detail, and hybrid approaches combining deterministic rules with contextual neural models are presented as a more robust alternative to purely rule-based systems. Similar applied solutions that combine OCR with image preprocessing and structured extraction are reported in [18] and [20], confirming the practical importance of OCR-based digitization in healthcare workflows.

Recent research trends demonstrate a shift toward integrated and multimodal document analysis systems. In [9], a multi-engine OCR framework is proposed for heterogeneous medical documents, while [11] presents a multimodal architecture integrating OCR, formatting, and AI-based interpretation. The study in [12] investigates multimodal large language models for digitizing and interpreting handwritten and printed medical reports, and [16] explores hybrid vision-language models for surgical documentation. Related developments in automatic summarization and document generation for healthcare applications are described in [17] and [21]. These works suggest that medical document processing is increasingly moving beyond isolated OCR toward end-to-end intelligent documentation systems.

Another growing line of research concerns the use of large language models for extracting, validating, and structuring clinical information. In [23], LLMs are applied to extract key parameters and detect inconsistencies in clinical trial documentation, including Ukrainian-language data. In [24], a scalable LLM-based framework is proposed for validated extraction of structured data from electronic health records, combining preprocessing, parsing, embedding-based retrieval, entity extraction, and mapping to healthcare interoperability standards. These results demonstrate the practical potential of LLM-based pipelines for structured medical data generation and validation.

In parallel, studies on OCR post-correction and robustness provide further insight into the limitations of purely text-based pipelines. In [15], the authors demonstrate that neural models for OCR error correction can significantly improve digitized text quality when trained on domain-relevant patterns and error distributions. Although the study is not focused on medical documentation, it highlights an important methodological point that is also relevant in healthcare document processing: OCR quality alone does not guarantee reliable structured data extraction, especially when downstream methods are sensitive to formatting variations and token-level distortions

Overall, the analysis of existing studies shows that automated medical document processing is a multidimensional problem requiring the integration of OCR, information extraction, semantic modeling, and document layout analysis. Existing research demonstrates significant progress in transformer-based NLP and NER [1], [4]–[6], multilingual biomedical processing [3], [23], OCR-based and hybrid extraction approaches [7], [8], [18], [20], multimodal document analysis [9], [11], [12], [16], and large language model-based methods for structured data extraction and validation [2], [13], [17], [21], [24], [25]. Additional studies highlight the importance of NLP in healthcare documentation and the impact of deep learning on medical data processing [14], [19], [22]. At the same time, there remains a practical and methodological gap in comparative studies focused specifically on medical laboratory PDF documents, particularly in evaluating OCR-based pipelines against layout-aware approaches for extracting structured numerical results, units, and reference ranges. This gap determines the relevance of the present study.

Aim and tasks of the study. The aim of this study is to develop and evaluate an approach for improving the automated extraction of structured information from medical PDF documents by applying transformer-based OCR models and intelligent document analysis methods under conditions of heterogeneous document formats. In the authors previous work, a conceptual and functional model of medical center business processes was developed, focusing on the digitalization of laboratory test results and the formalization of service delivery processes for the design of an intelligent data analysis information system [26]. This prior research provided a methodological foundation for identifying key processes requiring automation and highlighted the importance of structured data extraction as a critical component of the overall system.

To achieve this aim, the following tasks are defined:

- to analyze the characteristics of medical laboratory PDF documents and identify key challenges in their automated processing;
- to implement and compare different text recognition approaches, including classical OCR and transformer-based models;
- to develop and apply baseline rule-based methods for structured data extraction;

- to implement a layout-aware extraction approach using positional information from PDF documents;
- to evaluate the quality of the proposed approaches using a set of metrics for both text recognition and information extraction;
- to analyze the impact of OCR output on the accuracy of structured data extraction and determine the most effective approach for practical application.

Materials and model. Document Processing Pipeline. The processing of medical laboratory reports is implemented as a multi-stage pipeline that reflects the sequential transformation of unstructured document data into structured, machine-readable information. The general workflow of the proposed pipeline is illustrated in Fig. 1.

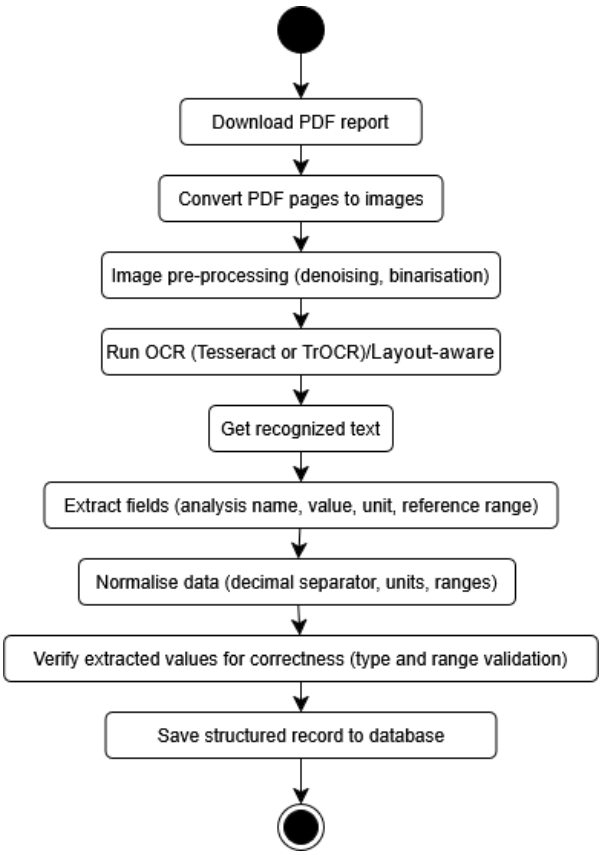

Fig. 1. General workflow of medical document processing pipeline

The proposed approach is based on the decomposition of the overall task into independent functional stages, which allows for flexible modification, replacement, and optimization of individual components without affecting the entire system. At the initial stage, PDF documents containing laboratory test results are converted into raster images to ensure compatibility with image-based text recognition methods. Given the variability in document quality, including noise, compression artifacts, and differences in scanning resolution, an image preprocessing step is applied. This stage includes noise reduction and adaptive binarization, which improve the contrast between text and background and enhance the performance of subsequent recognition algorithms.

Following preprocessing, optical character recognition is performed using selected OCR models. In this study, both classical and transformer-based approaches are considered, enabling a comparative analysis of their effectiveness in processing medical documents. The output of this stage is unstructured textual data, which may contain recognition errors, formatting inconsistencies, and disrupted logical relationships between document elements.

The next stage focuses on the extraction of structured information from the recognized text. This involves identifying key fields such as the test name, measured value, measurement units, and reference ranges. The extraction process is particularly challenging due to the tabular nature of laboratory reports and the close spatial relationship between textual and numerical data.

To ensure consistency and correctness, the extracted data undergo normalization, including standardization of decimal separators, measurement units, and value ranges. Additionally, validation procedures are applied to verify the correctness of extracted values based on expected data types and medically plausible ranges.

The final stage of the pipeline involves storing the validated structured data in a database, making it available for further analysis and integration into clinical information systems.

It is important to emphasize that optical character recognition represents only an intermediate stage of the overall process and does not guarantee the correctness of the final structured output. The experimental results demonstrate that the quality of extracted information depends not only on the accuracy of text recognition but also, to a greater extent, on the effectiveness of the extraction method and the ability to account for the structural characteristics of medical documents.

Evaluation Metrics. To evaluate the performance of the proposed approaches, a set of metrics was used to assess both the quality of text recognition and the accuracy of structured data extraction. The evaluation process considers errors at different levels, including character-level discrepancies, word-level distortions, and the correctness of extracted structured fields.

The quality of optical character recognition is measured using the Character Error Rate (CER), which reflects the proportion of incorrectly recognized characters relative to the reference text. This metric is defined as the normalized Levenshtein distance between the predicted and reference strings:

$$CER = D(pred, ref) / |ref|, \quad (1)$$

where $D(pred, ref)$ denotes the Levenshtein distance, and $|ref|$ is the length of the reference text.

In addition to CER, the Word Error Rate (WER) is used to evaluate recognition quality at the word level. This metric captures structural distortions in the text and is particularly sensitive to word order and segmentation errors:

$$WER = (S + D + I) / N, \quad (2)$$

where S represents substitutions, D denotes deletions, I corresponds to insertions, and N is the total number of words in the reference text.

To assess the correctness of structured data extraction, the Exact Match (EM) metric is applied. This metric

evaluates whether all required fields are extracted without any errors and is defined as a binary indicator:

$$EM = [\text{if all fields are correctly extracted}], \quad (3)$$

Furthermore, the evaluation includes standard classification metrics such as Precision and Recall, which measure the correctness and completeness of extracted fields. Precision reflects the proportion of correctly extracted fields among all predicted fields:

$$Precision = TP / (TP + FP), \quad (4)$$

where TP denotes true positives and FP represents false positives.

Recall, in turn, measures the proportion of correctly identified fields relative to all relevant fields in the reference data:

$$Recall = TP / (TP + FN), \quad (5)$$

where FN corresponds to false negatives.

To provide a balanced assessment of extraction performance, the F_1-score is used as a harmonic mean of *Precision* and *Recall*:

$$F_1 = 2 \cdot Precision \cdot Recall / (Precision + Recall). \quad (6)$$

Finally, for numerical fields such as measured values and reference ranges, the relative error is calculated to quantify the deviation between predicted and true values:

$$Error = |x_{pred} - x_{true}| / |x_{true}|. \quad (7)$$

This metric is particularly important in the medical domain, where even small numerical inaccuracies may significantly affect the interpretation of laboratory results.

The combination of these metrics enables a comprehensive evaluation of both recognition quality and extraction accuracy, providing a detailed understanding of the strengths and limitations of the analyzed approaches.

Experimental Setup. The experimental study was designed to evaluate the effectiveness of different approaches to text recognition and structured information extraction from medical laboratory PDF documents. The experimental workflow includes the preparation of input data, application of alternative OCR methods, extraction of structured fields, and evaluation of results using a set of predefined metrics.

The experimental study was conducted to evaluate the effectiveness of different approaches to text recognition and structured data extraction from medical laboratory PDF documents. A set of documents was selected, and reference data were manually created to serve as ground truth for objective comparison.

Two pipelines were implemented using classical (Tesseract) and transformer-based (TrOCR) OCR methods. The recognized text was processed with a rule-based extraction module to identify key fields, including test names, values, units, and reference ranges.

The performance was evaluated using CER and WER for text recognition, and Exact Match, Precision, Recall, F1-score, and relative error for structured data extraction. The results were aggregated into a comparative table to enable systematic analysis of the approaches.

The interaction between system components during document processing is represented using a sequence diagram, shown in Fig. 2.

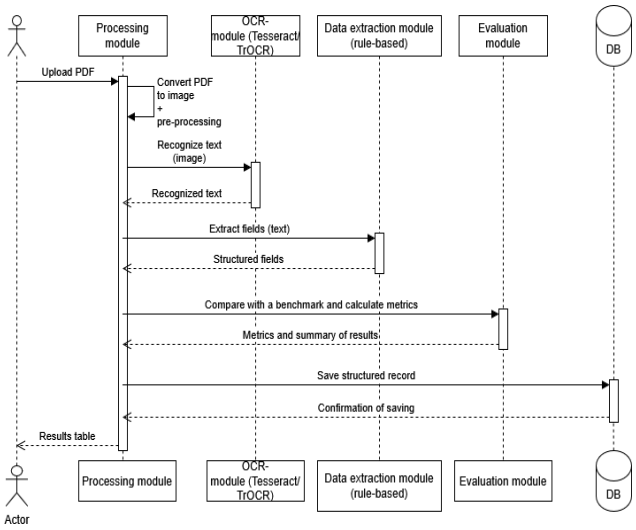

Fig. 2. Sequence diagram of document processing and evaluation pipeline

This diagram provides a detailed view of the data flow between the processing module, OCR module, extraction module, evaluation module, and database. It demonstrates how the input document is sequentially transformed into structured data and how evaluation results are generated and stored.

Such a representation allows for a clearer understanding of the system architecture and emphasizes the modular nature of the proposed approach. In particular, it illustrates that each component of the pipeline operates as an independent functional unit, which facilitates scalability, maintainability, and the integration of alternative methods for text recognition or information extraction.

Results and discussion. The results of the comparative evaluation are presented in Table 1, which summarizes the performance of three different pipelines combining text recognition and structured data extraction approaches.

Table 1 – The results of the comparative evaluation

Pipeline	Metrics					
	CER	*WER*	*EM*	Precision	Recall	F_1
TrOCR + Rules	~0.99	1.00	0	0.0	0.0	0.00
Tesseract + Rules	0.69	0.81	0	0.8	1.0	0.89
pdfplumber + Layout-aware	0.00	0.00	1	1.0	1.0	1.00

The obtained results demonstrate a clear difference in performance between the evaluated pipelines, particularly when comparing OCR-based approaches with layout-aware extraction. The transformer-based TrOCR model showed significantly lower performance, with high CER and WER values indicating substantial distortions in the recognized text. As a result, the extraction stage failed

completely, since rule-based methods rely on correctly recognized patterns.

The Tesseract-based pipeline achieved better text recognition quality, which enabled partial extraction of structured data. The results indicate that all relevant fields were detected, as reflected by high Recall, while lower Precision values reveal the presence of incorrectly extracted elements caused by OCR errors. This confirms that, although classical OCR methods provide more stable results, recognition errors still propagate to the extraction stage and reduce overall reliability.

In contrast, the layout-aware approach based on pdfplumber demonstrated the best performance across all evaluation metrics. By directly utilizing the textual layer of the PDF and preserving spatial relationships between document elements, this method eliminates recognition errors and ensures accurate extraction of structured data.

Overall, the results indicate that the effectiveness of medical document processing is determined primarily by the extraction stage and the ability to account for document structure, rather than by OCR accuracy alone.

Conclusion and future work. The conducted study addressed the problem of automated processing of medical laboratory reports with a focus on comparing different approaches to text recognition and structured data extraction. The experimental evaluation demonstrated that the overall effectiveness of document processing pipelines depends not only on OCR quality but, to a greater extent, on the methods used for extracting structured information.

The results of the comparative analysis showed that transformer-based OCR models, such as TrOCR, may produce unsatisfactory results without domain-specific adaptation, leading to a complete failure of the extraction stage. Classical OCR methods, represented by Tesseract, provide more stable recognition quality and enable partial extraction of relevant fields; however, recognition errors still negatively affect the final structured output. In contrast, the layout-aware approach based on pdfplumber achieved perfect performance across all evaluation metrics by leveraging the inherent structure of PDF documents and bypassing OCR altogether.

These findings confirm that the key challenge in processing medical documents lies in the correct interpretation of document structure and the reliable extraction of semantically meaningful data. The study highlights the importance of integrating layout-aware techniques into document processing pipelines and demonstrates that combining different approaches may be necessary to achieve robust performance across various document types.

Future work will focus on extending the experimental study to a larger and more diverse dataset of medical documents, including cases without an embedded text layer where OCR remains unavoidable. Additionally, it is planned to investigate advanced extraction methods, such as Named Entity Recognition (NER) models and layout-aware deep learning approaches (e.g., LayoutLM), which can better capture the relationship between textual content and document structure. Further research will also consider the development of hybrid pipelines that dynamically select the optimal processing strategy depending on document characteristics, as well as the integration of the proposed system into real-world medical information systems for practical validation and decision support.

Declaration on the use of generative AI. During the preparation of this work, the authors used ChatGPT and Grammarly for grammar and spell checking, as well as for rephrasing and reformulating the text. After using these tools, the authors reviewed and edited the content as necessary and take full responsibility for the content of this publication.

References

1. Patil S., Golbhavi S. *Advances and Applications of Natural Language Processing in Healthcare.* Available at: https://www.ijraset.com/best-journal/advances-and-applications-of-natural-language-processing-in-healthcare (accessed: 28.02.2026). DOI: https://doi.org/10.22214/ijraset.2025.73380.
2. Heilmeyer F., Böhringer D., Reinhard T., et al. *Viability of Open Large Language Models for Clinical Documentation in German Health Care: Real-World Model Evaluation Study.* Available at https://medinform.jmir.org/2024/1/e59617: (accessed: 28.02.2026). DOI: https://doi.org/10.2196/59617.
3. Mazzucato E., Seinen M. *Advancements in Multilingual Biomedical Natural Language Processing: exploring Large Language Models for Named Entity Recognition and Linking.* Available at: https://www.medrxiv.org/content/10.64898/2026.01.22.26344605v1 (accessed: 28.02.2026). DOI: https://doi.org/10.64898/2026.01.22.26344605.
4. Almeida S. S., Fontes R. S., Alves L. et al. *Artificial intelligence in healthcare text processing: a review applied to named entity recognition.* Available at: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1584203/full (accessed: 28.02.2026). DOI: https://doi.org/10.3389/frai.2025.1584203.
5. Loor-Torres R., Duran M., Toro-Tobon D., et al. *A systematic review of natural language processing methods and applications in thyroidology.* Available at: https://www.mcpdigitalhealth.org/article/S2949-7612(24)00027-0/fulltext (accessed: 28.02.2026). DOI: 10.1016/j.mcpdig.2024.03.007.
6. Guo B., Liu H., Niu L. *Integration of natural and deep artificial cognitive models in medical images: BERT-based NER and relation extraction for electronic medical records.* Available at: https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2023.1266771/full (accessed: 28.02.2026). DOI: https://doi.org/10.3389/fnins.2023.1266771.
7. Jothi G., Virgeniya S. C. *Medical data NER and classification using hybridized BERT model.* Available at: https://wjaets.com/content/medical-data-ner-and-classification-using-hybridized-bert-model (accessed: 28.02.2026). DOI: https://doi.org/10.30574/wjaets.2024.13.1.0376.
8. Ma M.-W., Gao X.-S., Zong H. et al. *Extracting laboratory test information from paper-based reports.* Available at: https://link.springer.com/article/10.1186/s12911-023-02346-6 (accessed: 28.02.2026). DOI: https://doi.org/10.1186/s12911-023-02346-6.
9. Abhijeet F. S. *Hybrid approaches for NER in noisy OCR medical records.* Available at: https://ijsra.net/content/hybrid-approaches-ner-noisy-ocr-medical-records (accessed: 28.02.2026). DOI: https://doi.org/10.30574/ijsra.2025.16.3.2499.
10. Zhang Q. *Improving classification accuracy for unstructured medical documents via multi-engine OCR and deep learning collaboration*. Available at: https://scipublication.com/index.php/JACS/article/view/292 (accessed: 28.02.2026). DOI: https://doi.org/10.69987/jacs.2026.60201.
11. Frei J., Kramer F. *GERNERMED – an open German medical NER model.* Available at: https://www.sciencedirect.com/science/article/pii/S2665963821000944 (accessed: 28.02.2026). DOI: https://doi.org/10.1016/j.simpa.2021.100212.
12. Wei Q., Chen X., Cao C. *A technical framework for recognizing and interpreting complex medical records: based on multimodal large language model.* Available at: https://dl.acm.org/doi/10.1145/

3702386.3702396 (accessed: 28.02.2026). DOI: https://doi.org/10.1145/3702386.3702396.
13. Negm M., Mourad A., Fawzi S. et al. *Leveraging large language models for digitization and clinical interpretation of handwritten psoriasis reports.* Available at: https://ieeexplore.ieee.org/document/11418814 (accessed: 28.02.2026). DOI: https://doi.org/10.1109/MELECON64486.2026.11418814.
14. Bai E., Luo X., Kutzin J. M. et al. *Assessment and integration of large language models for automated electronic health record documentation in emergency medical services.* Available at: https://link.springer.com/article/10.1007/s10916-025-02197-w (accessed: 28.02.2026). DOI: https://doi.org/10.1007/s10916-025-02197-w.
15. Shrivastava D., Malathi H., Bansal S. et al. *Integrating natural language processing in medical information science for clinical text analysis.* Available at: https://mw.ageditor.ar/index.php/mw/article/view/513 (accessed: 28.02.2026). DOI: https://doi.org/10.56294/mw2024513.
16. Dhote M. G., Deore M. P., Jadhav T. et al. *Hybrid Vision-Language Models for Real-Time Surgical Report Generation and Documentation.* Available at: https://jneonatalsurg.com/index.php/jns/article/view/2752 (accessed: 28.02.2026). DOI: https://doi.org/10.52783/jns.v14.2752.
17. Tang Y. *Research on NLP-Based Automatic Summarization for Medical Records.* Available at: https://www.clausiuspress.com/article/9613.html (accessed: 28.02.2026). DOI: https://doi.org/10.23977/acss.2023.070903.
18. Thorat Aditya S. *Automated Data Entry Through Image Processing for Medical Records.* Available at: https://ijsrem.com/download/automated-data-entry-through-image-processing-for-advanced-medical-inventory/ (accessed: 28.02.2026). DOI: https://doi.org/10.55041/ijsrem50196.
19. N. Bina. *The Role of Natural Language Processing in Medical Documentation.* Available at: https://rojournals.org/wp-content/uploads/2025/02/ROJBAS-51-2025-P3.pdf (accessed: 28.02.2026). DOI: https://doi.org/10.59298/rojbas%2F2025%2F511114.
20. Raja M. S., Aarthi C. R., Gayathri P., Pavithra J. P. *Automated Prescription Analysis and Alternative Suggestion Using OCR and NLP.* Available at: https://internationaljournalssrp.org/index.php/ijmst/article/view/92 (accessed: 28.02.2026). DOI: https://doi.org/10.64137/31079911%2Fijmst-v1i2p101.
21. Paithankar S., Patil S., Shivgan A., Mahabudhe M., Dharmadhikari P. A. *Medical Prescription Generator using Natural Language Processing.* Available at: https://ieeexplore.ieee.org/document/10969167 (accessed: 28.02.2026). DOI: https://doi.org/10.1109/ISACC65211.2025.10969167.
22. Gomathi S., Roopa Chandrika R. *Advancing Medical Image Processing with Deep Learning: Innovations and Impact* Available at: https://ictactjournals.in/ArticleDetails?id=bpkwui (accessed: 28.02.2026). DOI: https://doi.org/10.21917/ijivp.2025.0494.
23. Horlatch V., Pasichnyk V. *Automated Extraction of Key Parameters and Detection of Inconsistencies in Clinical Documentation Using Large Language Models.* Available at: https://journals.uran.ua/eejet/article/view/337915 (accessed: 28.02.2026). DOI: https://doi.org/10.15587/1729-4061.2025.337915.
24. Stuhlmiller T. J., Rabe A. J., Rapp J. et al. *A Scalable Method for Validated Data Extraction from Electronic Health Records with Large Language Models.* Available at: https://www.medrxiv.org/content/10.1101/2025.02.25.25322898v1 (accessed: 28.02.2026). DOI: https://doi.org/10.1101/2025.02.25.25322898.
25. Senkadi Kh., Belmiloud M., Benslimane S. M., Dif N. *Multi-Label Classification of Digitized Clinical Records Using Transformer-Based Models.* Available at: https://ieeexplore.ieee.org/document/11298127 (accessed: 28.02.2026) DOI: https://doi.org/10.1109/ICNAS68168.2025.11298127.
26. Grinchenko M., Kutsenko D. *Models of medical center business processes to improve decision-making efficiency.* Available at: https://journals.uran.ua/itssi/article/view/348508. (accessed: 28.02.2026) DOI: https://doi.org/10.30837/2522-9818.2025.4.005.

Received 02.03.2026
Accepted 26.03.2026
Published 20.05.2026

УДК 004.89:61

М. А. ГРИНЧЕНКО, кандидат технічних наук (PhD), доцент, Національний технічний університет «Харківський політехнічний інститут», завідувачка кафедри управління проєктами в інформаційних технологіях, м. Харків, Україна, e-mail: marina.grynchenko@khpi.edu.ua , ORCID: https://orcid.org/0000-0002-8383-2675
Д. О. КУЦЕНКО, здобувач освіти PhD, Національний технічний університет «Харківський політехнічний інститут», аспірант кафедри управління проєктами в інформаційних технологіях, м. Харків, Україна, e-mail: Dmytro.O.Kutsenko@cs.khpi.edu.ua, ORCID: http://orcid.org/0009-0005-6359-3143

ПОРІВНЯЛЬНИЙ АНАЛІЗ ТРАНСФОРМЕРНИХ ТА ІНТЕЛЕКТУАЛЬНИХ МЕТОДІВ ОБРОБКИ ДОКУМЕНТІВ ДЛЯ АВТОМАТИЗОВАНОГО ВИЛУЧЕННЯ МЕДИЧНИХ ДАНИХ З PDF

У статті розглянуто задачу автоматизованої обробки медичних лабораторних звітів у форматі PDF з акцентом на розпізнавання тексту та вилучення структурованої інформації. Досліджено ефективність різних підходів до оптичного розпізнавання символів (OCR), зокрема класичних методів і трансформерних моделей, а також методів вилучення ключових медичних даних із неструктурованого та напівструктурованого тексту. Проведено порівняльний експериментальний аналіз із використанням медичних документів різного типу, зокрема таких, що мають табличну та текстову структуру. Оцінювання якості здійснювалося за допомогою набору кількісних метрик, серед яких Character Error Rate (CER), Word Error Rate (WER), Exact Match (EM), Precision, Recall та F1-score. Отримані результати показали, що висока точність OCR не гарантує якісного вилучення структурованих даних, оскільки помилки розпізнавання суттєво впливають на подальші етапи обробки. Особливу увагу приділено підходам, що враховують структуру документа. Метод на основі прямого вилучення тексту з PDF із використанням pdfplumber продемонстрував найкращі результати завдяки збереженню просторових зв'язків між елементами документа та відсутності необхідності застосування OCR за наявності текстового шару. Результати дослідження підтверджують, що основна складність полягає у вилученні даних, а не лише у розпізнаванні тексту. Показано доцільність інтеграції структурно-орієнтованих та інтелектуальних методів для підвищення ефективності автоматизованих систем обробки медичних даних.

Ключові слова: оптичне розпізнавання символів, вилучення інформації, медичні документи, обробка медичних дани,, методи з урахуванням структури документа, вилучення даних, аналіз документів, системи підтримки прийняття рішень, інструменти штучного інтелекту.

Повні імена авторів / Author's full names

Автор 1 / Author 1: Гринченко Марина Анатоліївна / Grinchenko Marina Anatoliiovyvna
Автор 2 / Author 2: Куценко Дмитро Олександрович / Kutsenko Dmytro Oleksandrovich