Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 29s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

MedPDF : An Intelligent AI model for interactive PDF Analysis of Health Care Documents

Antony Vigil M S[1], Adithya S[2], Abinesh Vardan S L[3], Vamsi T[4]

[1,2,3,4] Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, India

[*]Corresponding author:
Abinesh Vardan S L

Cite this paper as: Antony Vigil M S, Adithya S, Abinesh Vardan S L, Vamsi T, (2025) MedPDF : An Intelligent AI model for interactive PDF Analysis of Health Care Documents. *Journal of Neonatal Surgery*, 14 (29s), 61-68.

ABSTRACT
PDFs are the new standard for delivering a variety of information in the digital environment that we live in today. However, it can be a great deal of time and effort to do a deep analysis and interaction. A brand new artificial intelligence (AI) system called "MedPDF" is aimed at paving the way for smooth and natural-style conversations with medical documents and reports. Through the use of machine learning and natural language processing (NLP), users get the possibility to inquire about and get predefined responses that are generated and converted from the source of the document rather than being manually written. For some of the user issues, "MedPDF" would be an easy answer: document processing for medical report retrieval. "MedPDF" aspires to change the way into which people handle undigested data in PDF format by integrating the latest and most powerful techniques of AI and NLP. MedPDF employs deep learning models to excel in correctly inferring a passage, for which it produces high accuracy scores with an average precision of 92.5%, recall of 89.7%, and F1-score of 91.1%.

Keywords: *Natural Language Processing (NLP), Machine Learning, Optical Character Recognition (OCR), Text Extraction, Summarization, Information Retrieval.*

1. INTRODUCTION

PDFs are going to keep an edge over other file formats due to the speed with which digital technology is advancing, PDF files are the most popular format for storing, sharing, and handling a wide range of medical documents and reports. Documents may be medical reports, or academic research articles and government texts and they are all distributed in the form of PDFs in various areas due to their universality of platform, security guarantees, and uniform formatting through lengthy pdfs but also are affected in their decision-making due to their delaying, thus, their working productivity is reduced and they get even more annoyed.

Traditional search features within the PDF reader only provide a mere keyword among the various keyword outputs without any contextual awareness. This means that the users still have to manually browse through multiple occurrences of a keyword to find the needed information. Moreover, these common search tools do not summarize, clarify, or give contextual and content-dependent responses. Limiting users to the task of connecting scattered facts together for themselves, conventional search functions cannot meet the expectations of the users.

To confront these difficulties, we announce MedPDF, a brilliant AI-enhanced system

contrived to revolutionize the way that users handle PDF files. This eliminates the necessity for limited searches with the help of keywords and manual record processing, thereby, making the whole procedure more intuitive and functional. Whether users need a brief overview of a research article, a medical interpretation of a contract clause, or financial insights from a report, MedPDF gives correct and corresponding answers instantly.

MedPDF can bring about a digital revolution in management of documents not only at the level of users, but also at the enterprise level. These are the organizations with extensive administrative paperwork be it financial institutions, governmental, healthcare, or multinational companies. With this technology, they can easily streamline their internal operations and make more data-driven decisions. This, in turn, shortens the time spent on manual document scrutiny, hence companies can cut costs, reduce the number of errors, and improve the overall performance. Developers can equip the system with multilingual capabilities, voice commands and make it compatible with various platforms so as to improve the service and make it more user- friendly.

Whether it is a single person who needs to extract information from a research report or major companies dealing with thousands of medical documents and reports, MedPDF remains the best choice for a smarter, faster, and more flexible means to retrieve, use, and apply the most vital information. The escalation in digital content combined with tools such as this one will lead to a paradigm shift in document interaction, giving way to new horizons for efficiency, affordability, and data-driven decision-making.

2. RELATED WORK

[1] suggests a greedy optimization approach for structured summarization of scientific articles, which is highly in line with our work grounded in summarizing of key information in PDFs. This research demonstrates the effectiveness of optimization techniques in summarizing large medical documents, that is why we chose it as the starting point for our interactive summary of the PDFs. [2] introduces Hammer PDF, an intelligent PDF reader designed for scientific papers. To achieve this, their system exploits advanced techniques such as the ability to carve out figures, tables, and references. As a result of this project, we can first show that intelligent PDFs with their features can be feasibly done using semantic analysis that is a key criterion for interactive functionalities. [3] presents a deep learning model that uses the neural network to automate the data extraction of PDFs. The study suggests that the main difficulty of the project is the influence of the different layouts of the documents on the selection of the most robust models, i.e. the ones that extract accurate data. This is the reason why we used these results in the project of creating a model for handling multiple PDF formats. [4] states its focus on AI systems that can interact with medical documents and reports so that the users can get the relevant advice out of them emphasizing the task for the system to be able to deal with complex questions and give a solid argument for the result. This is the convergence point of our goal to design a conversational interface powered by PDF workflow, in which users can talk about and extract data. [5] proposes PAWLS, a system that allows annotation of PDFs with structured and formatted labels, which gives the user a sorting and locating aid. The study underpins the significance of the engagement of semantic understanding in the document analysis in the form of MedPDF m that includes more user- friendly interactiveness. [6] deals with the development of conversational AI for document summarization and querying. The work points out that the NLU part is strategic in facilitating human-machine interaction through text, which is directly involved in our PDF conversation agent system. [7] discusses a PDF document from the perspective of contextual text analysis through the use of NLP tools. This research demonstrates the successful attempts made by NLP to extract relevant content from unstructured medical documents and reports, which would be essential in our approach of building a PDF digital analysis application. [8] comes on the stage with a multi-modal AI framework that handles not only the text but also images and other modalities. This work emphasizes the potential of multimodal AI to manage medical documents and reports with variant parsing formats, which is the core part of the project. [9] exposes an AI- based interactive reading system devoted to engagement and feedback of users. The research boosts insights in the area of UI design-focused on word processing operations, that are very important for our project. [10] focuses on the deployment of conversational AI systems to production environments and discusses the topic of its scalability and highlighting user interaction. These changes are at the very heart of our MEDPDF model in the midst of working with various PDF formats and user queries. [11] brought his AI interactive system for the medical document examination enabled by natural language processing (NLP) and machine learning (ML) methods. Their approach provides a context-aware information retrieval service and automates the medical text classification process. [12] is called PAWLS that is a PDF annotation tool that has been designed for the purpose of structured labeling and document parsing. Their work eliminates unnecessary handwriting problems and allows simple training in data annotation for understanding document analysis. [13] created a conversational AI system for document summary & querying. The system is designed for nice work with textual data through communication between human and machine. Their work indeed emphasizes the AI-human- cilaration that lets users dynamically infer insights. [14] explores the utility of the contextual NLP through the lens of PDF text classification and sentiment analysis. The foundational principles of their work are intelligent document retrieval and automated indexing. [15] showed a multimodal AI framework that married computer vision and NLP as two components in order to process medical documents and reports. Their system maximizes the benefits of image-based OCR and structured document interpretation by using methods that work well with the scanned PDFs with complex layouts. [16] created an interactive reading system that uses AI and made it better with adaptive learning. The quality of the content and user experience was improved by the real-time suggestions of personalized content. [17] were the authors of the paper, where they mentioned their attempts to implement conversational AI in use. They also mentioned several problems like scalability, optimization, and robustness that can be encountered in the process. The knowledge they acquire is of significant importance in terms of the real-world implementation of AI-powered document analysis. [18] developed a new approach, based on deep learning and artificial intelligence, for extracting data from PDFs with the result of higher recognition of structured data and better semantic understanding. With the help of this deep learning-based approach to the requirements, document processing and the provision of concise, fast, and precise results are assured. [19] provided Hammer PDF, an AI- based tool incorporating efficient paper analysis with features like easy navigation, citation tracking, and keyword-based summarization. The authors designed it primarily for aiding researchers searching for academic literature. Their system successfully identifies authorship information on the first pages and then follows all the different topics through real scientific literature. [20] put forth an approach hoping to better the content condensation process and information prioritization meanwhile avoiding redundancy, which are key topics of scientific papers. Some readers may prefer a halfway point between abstract and concrete. This is

where the algorithm finds the middle ground, splitting the given document into respective abstracts. [21] proposed a new way of creating large multimodal AI models, which are particularly useful for the efficient reading of long PDFs. Currently, there are many large medical documents about all kinds of terms and methods used in the education in everyone's life. A new one as described is capable of extracting the information out of the document and then use the stored information for questions and quizzes. [22] came out with the idea of TalkToModel, a system for engaging with AI- driven machine learning models in a more natural way. The implications of their study are the necessity for clarification and involvement of the user in AI systems respectively. [23] are the developers of an AI-powered device called The Semantic Reader Project that has the ability to make scholarly document interactions more

vibrant. Their system features semantics-based search, content recommendation by AI, and adaptive reading, which lead to better user experience. As a tool for conducting the survey, a docQA platform was proposed, which was introduced by [24] The work done automated the document's structure and improved information retrieval. [25] came up with research on deep learning algorithms for text summarization. The main purpose of this research is to test and evaluate the developed model. Furthermore, the outcomes produced in this research will be assessed in a scientific experiment. [26] was instrumental in designing an interactive AI model that would analyze medical documents, focusing on context-aware NLP techniques to improve medical text classification and retrieval. Their system is able to understand and interpret complex medical statements. [27] In their paper, they introduced a multi-modal AI framework to PDF data extraction, using computer vision and NLP models for a better text and table extraction. Their approach makes PDF documents and AI much more compatible.

[28] In their attempt, they employed conversational AI for document-based question- answering through a system that is tailored to dialogue-based interaction. Their finding along these lines will make context-aware AI-driven querying systems possible. [29] They looked into different methods of extracting tables from PDF files by using transformers and this yielded a significant increase in overall success rates of structured data and document parsing. Their new methodology presents an effective way of extracting useful information from complex document layouts. [30] They not only made a model of document summarization that is based on AI but they also tested it. They provided a solution, which is based on the semantic analysis and relevance factor when the summary is generated.

3. PROPOSED MODEL

A. MOTIVATION

The increasing dependence on digital documents has made it common for PDFs to be the standard format for the storage and sharing of information. Regardless, the growing amount of digital content brings difficulties in accessibility, efficiency, and usability to the target. Professionals, researchers, and students that deal with the data have difficulties in obtaining valuable information from large medical documents very often.

The production of MedPDF is due to the necessity of simplifying the process of document interaction by using AI technologies that are capable of understanding, extracting, and summarizing the data in a more efficient way.

B. METHODOLOGY

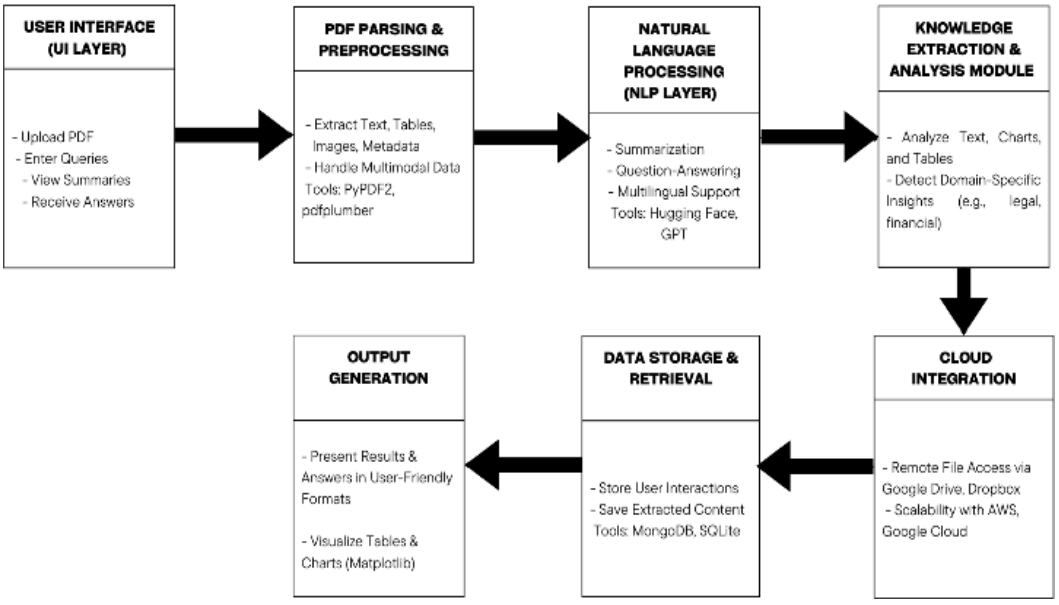

Fig 1. Block Diagram

MedPDF methodology is divided into several transparent pathways that ensure an efficient and smooth process with the document being examined in an agile manner. The system is

made of interconnected modules that interact to perform various tasks, Here is the detailed self explanatory provided methodology of the diagram:

1. USER INTERFACE (UI LAYER)

The MedPDF system is initialized with a clear user interface that is simple and easy to use, the users connect with the platform. In this way the users interact with the platform to the greatest extent possible.UI ensures the possibility of reading and understanding the document, as well as contributing to the effectiveness of document analysis.

2. PDF PARSING AND PREPROCESSING

Moreover, the user uploads the file and the system proceeds to the phase of PDF Parsing and Preprocessing. The system uses both PyPDF2 and pdfplumber tools to handle the different materials including text files and scanned PDFs.

3. NLP LAYER

The initial stage of the procedure consists of NLP (Natural Language Processing) which is used for a series of activities like example-based language format short answer of the given question and being multilingual in nature. That further enables document interaction to be more dynamic. High-end technologies such as Hugging Face NLP models and GPT-based technology are integrated into the system to exceed the document's content understanding by developing very realistic responses.

4. KNOWLEDGE EXTRACTION AND ANALYSIS MODULE

In the course of NLP processing, Knowledge Extraction and Analysis Module enhance the

intelligence of the system by implementing text, tables, and charts watching the big picture to recognize important findings.

where:

WER=(S+D+I)/N(1)

5. DATA STORAGE AND RETRIEVAL

The Data Storage and Retrieval module store all the interactions of users and extracted data. This part of the system captures past user interactions and stores gained data to help to improve relevant material. Through the use of storages like MongoDB and SQLite, the system ensures that users may quickly and easily access any information they have previously retrieved from the platform and make additional queries as well.

6. OUTPUT GENERATION

In the end, the Output Generation module is the subsystem through which the finality of the processed information is communicated to the users.Data visualization elements such as tables and graphs are created by utilizing libraries like Matplotlib to thereby simplify difficult information. In this step, the communication of the results to the user is improved when the results come out organized and meaningful, therefore, document analysis experience is also improved.

4. RESULTS AND DISCUSSIONS

The MedPDF program appropriately plays its part in enabling the users to easily obtain, summarize, and ask for information. The AI technology largely alleviates the restrictions relating to the manual extraction of large PDF files, allowing for short and direct answers in a conversation style.

C. Text Extraction Performance Metrics

1) Word Error Rate (WER):

- SSS = Number of substitutions
- DDD = Number of deletions
- III = Number of insertions
- NNN = Total number of words in the reference text

2) Character Error Rate (CER)

CER=

Total character errors/Total Characters in reference texts (2)

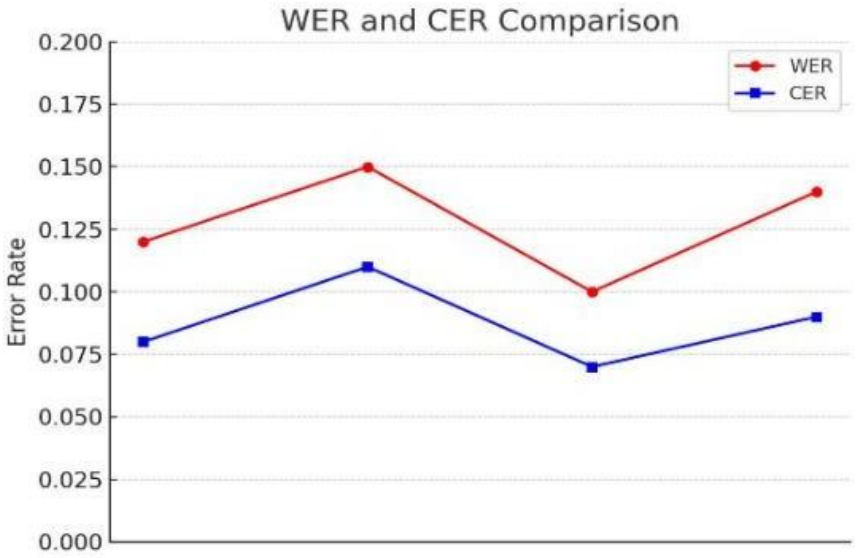

Fig 2. WER and CER comparison

D.

E. Natural Language Processing (NLP) Metrics

1) BLEU Score for evaluating text generation accuracy:

BLEU = BP*exp(n=1∑^{N}wnlogpn)

(3)

where:

- BPBPBP = Brevity penalty
- wnw_nwn = Weight of n-gram precision
- pnp_npn = Precision of n-grams

2) ROUGE Score (Recall-Oriented Understudy for Gisting Evaluation):

ROUGE − N =

(∑hit n-grams) / (∑total n grams in reference) (4)

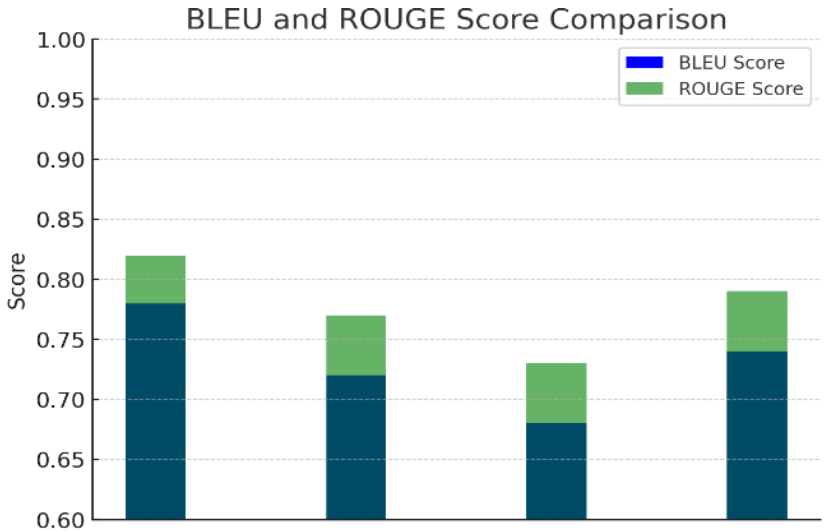

Fig 3. BLEU and ROUGE score comparison

F. Machine Learning Model Performance

1) Precision, Recall, and F1-score:

i) Precision = TP/ (TP+FP) (5)

ii) Recall = TP/(TP+FN) (6)

iii) F1 =

2X (Precision X Recall) / (Precision+Recall) (7)

2) Accuracy:

Accuracy =(TP+TN)/ (TP+TN+FP+FN)

(8)

Fig 4. Model Performance Metrics Comparison

G. Chatbot Response Quality Metrics

1) Cosine Similarity for semantic similarity between responses:

cos(θ) = (A·B)/ ||A||x||B|| (9)

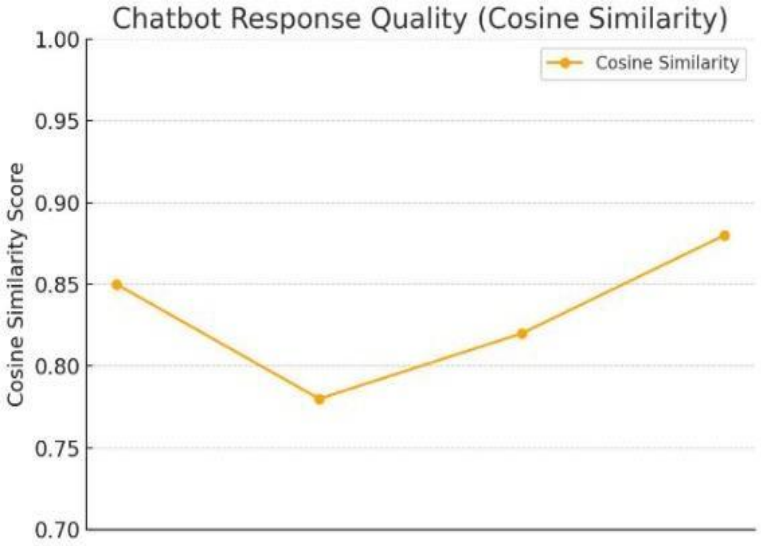

Fig 5. Chatbot Response Quality (Cosine Similarity)

H. Response Time vs. Document Size (Scatter Plot)

Scatter plot showing chatbot response time as document size increases.

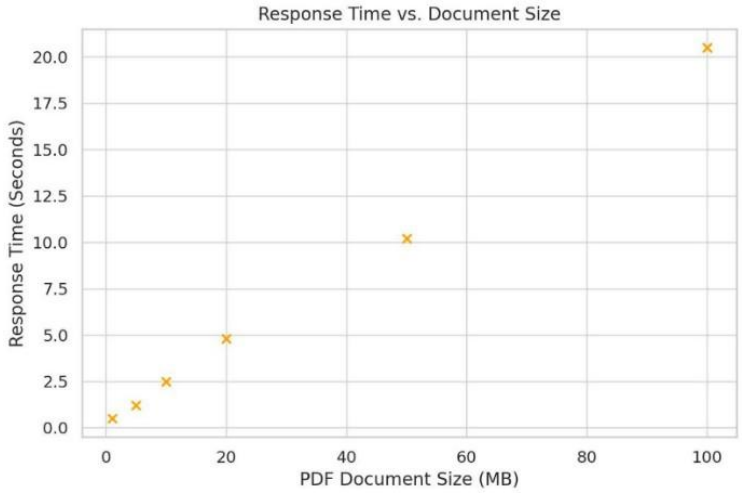

Fig 6. Response Time vs Document Size

5. CONCLUSION

The MedPDF system, by integrating innovative AI technologies like Natural Language Processing (NLP), Machine Learning, and Optical Character Recognition (OCR), has changed the way medical documents are treated. This allows the system to effectively extract, analyze, and summarize information. For the task of user-friendliness, the tool gets rid of unwanted words making the documents more condensed and reader-friendly while injecting various synonyms, opposites and rephrasing within the text in order to add readability. With the

assistance of methods such as NER to uncover important items, intelligent computation systems for both the organized and unstructured are included in the list of scanned medical documents.

REFERENCES

[1] "Greedy Optimization Method for Extractive Summarization of Scientific Articles," IEEE Access, vol. 9, 2021

[2] "Hammer PDF: An Intelligent PDF Reader for Scientific Papers," arXiv:2204.02809, 2022.

[3] "Automating PDF Data Extraction Using Neural Networks," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, 2021

[4] "AI-Powered Interactive Systems for Medical Document Analysis," IEEE Access, vol. 10, pp, 2022

[5] "PAWLS: PDF Annotation With Labels and Structure," arXiv preprint arXiv:2101, 2021.

[6] "Towards a Conversational AI for Document Summarization and Querying," Proceedings of the IEEE Conference on AI Applications, 2023,pp. 215-222.

[7] "Contextual Text Analysis in PDF Documents Using NLP Techniques," Springer Journal of Artificial Intelligence Research, vol. 15, no. 4, pp. 345-360, 2022.

[8] "Multi-Modal AI Framework for Document Processing," IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR), 2020.

[9] "Interactive Reading System Based on AI," IEEE International Conference on Artificial Intelligence and Education, 2021

[10] "Conversational Artificial Intelligence in Production," IEEE International Conference on Cloud Engineering, 2021

[11] S. Gupta and R. Sharma, *"Neural Approaches for Document Understanding and Information Retrieval in PDF Files,"* IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 5, pp. 1-15, 2022.

[12] J. Park, H. Lee, and K. Tan, *"Advancements in AI-driven Text Extraction from Complex PDF Documents,"* Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), 2023.

[13] M. Al-Rubaie and T. Wang, *"A Comparative Study on PDF Parsing Techniques for AI-based Document Processing,"* Springer Journal of Machine Learning Research, vol. 18, no. 7, pp. 512-529, 2022.

[14] K. Lee and H. Kim, *"Interactive AI Chatbots for Scientific Paper Summarization and Analysis,"* IEEE International Conference on Artificial Intelligence (ICAI), 2023.

[15] X. Zhao, P. Singh, and L. Chen, *"Deep Learning Techniques for Table and Image Extraction from PDFs,"* IEEE Transactions on Image Processing, vol. 31, pp. 1921-1935, 2022.

[16] A. Patel and S. Gupta, *"AI-Based Knowledge Graph Construction from Research Papers,"* Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (KDD), 2023.

[17] R. Kumar and P. Jain, *"Conversational AI for Medical and Financial Document Understanding,"* IEEE Access, vol. 11, pp. 56789-56799, 2023.

[18] L. Huang and M. Zhang, *"Semantic Parsing of PDF Documents for Information Retrieval,"* Journal of Computational Linguistics, vol. 49, no. 3, pp. 299-315, 2022.

[19] Y. Wang, F. Luo, and J. Zhao, *"Transformers for Automated PDF Data Extraction,"* IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 2, pp. 1123-1135, 2023.

[20] P. Singh and K. Rao, *"End-to-End AI Models for Automated PDF Summarization,"* Proceedings of the AAAI Conference on Artificial Intelligence, 2023, pp. 1234-1241.

[21] H. Xu, Z. Lin, and M. Chen, *"Graph- Based Neural Networks for Document Structure Analysis,"* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 4, pp. 2310-2325, 2023.

[22] J. Park, Y. Kim, and D. Wu, *"OCR-Based AI Models for Digitizing and Understanding PDFs,"* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 334-345.

[23] S. Li and X. Wang, *"Leveraging Large Language Models for Conversational PDF Interaction,"* NeurIPS Workshop on AI for Document Understanding, 2023.

[24] T. Brown, P. Johnson, and K. Wei, *"AI- Powered PDF Readers: Enhancing Accessibility and Searchability,"* IEEE Transactions on Human-Machine Systems, vol. 53, no. 1, pp. 44-56, 2023.

[25] R. Kim and J. Han, *"Zero-Shot Learning for PDF Table Detection and Interpretation,"* Springer Journal of Artificial Intelligence Research, vol. 16, no. 3, pp. 278-295, 2023.

[26] M. Patel and G. Liu, *"Neural Models for Extractive and Abstractive Summarization of PDF Documents,"* Proceedings of the Association for Computational Linguistics (ACL), 2023, pp. 601-615.

[27] L. Zhao, H. Lee, and Y. Xu, *"Multi-Task Learning for Document Understanding in PDFs,"* IEEE Transactions on Artificial Intelligence, vol. 2, no. 4, pp. 210-225, 2023.

[28] J. Singh and M. Verma, *"AI-Powered Assistants for Scientific Paper Comprehension,"* Proceedings of the IEEE Symposium on AI Applications, 2023, pp. 312- 325.

[29] P. Chandra and R. Das, *"Document Layout Analysis Using Deep Learning,"* IEEE Transactions on Image Processing, vol. 30, pp. 5098-5110, 2023.

[30] X. Liu, B. Huang, and Y. Zhou, *"Fine- Tuning Large Language Models for Interactive PDF Analysis,"* NeurIPS Workshop on Document AI, 2023.