IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

EXPLAINABLE AI FOR MEDICAL REPORT ANALYSIS USING NLP AND KNOWLEDGE RETRIEVAL

[1]R. Anitha, [2]Sindhuja S, [3]P. Vinothiyalakshmi

[1,2,3]Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, Sriperumbudur, Tamil Nadu, India

Doi: https://doi.org/10.56975/ijcrt.v14i7.309047

Abstract: As healthcare systems become more digitized, there has been a significant increase in electronic medical records, diagnoses, and clinical documents. Yet, the specialized nature of the language used in many reports makes it hard for non-medically trained people to understand them fully, resulting in poor health literacy and possibly causing erroneous interpretations. In this paper, we propose MediScan AI, which is an artificial intelligence platform for automatic medical report interpretation using explainable approaches. Specifically, MediScan AI incorporates Optical Character Recognition (OCR), Biomedical Named Entity Recognition (NER), transformer-based summarization, and knowledge retrieval into a single pipeline. Tesseract is used for OCR, and SciSpacy & BioBERT are used for NER tasks. For summarization purposes, transformer models such as DistilBART and FLAN-T5 are leveraged. Knowledge retrieval is accomplished by using the MedQuAD dataset, together with TF-IDF & BM25 as ranking algorithms. The experimental results show an F1-score of 0.93 for the biomedical NER task and a competitive ROUGE score compared to baselines.

***Keywords* -** Explainable AI, Medical Report Analysis, Biomedical NLP, SciSpacy, BioBERT, Transformer Models, DistilBART, FLAN-T5, Tesseract OCR, Knowledge Retrieval, TF-IDF, BM25, MedQuAD

I. INTRODUCTION

In the era of digital revolution, there has been an increase in the number of electronic medical records, diagnostic reports, and clinical documentation due to the fast digitization of health care systems. It has become clear that intelligent and AI-powered health care systems can significantly influence clinical work with automated medical data analytics [1]. Nonetheless, the complicated nature of medical reports often leads to a lack of health literacy among patients and other non-specialist groups. There have been some significant advances in AI and NLP technologies in recent years, and one of the benefits of them is the ability to automate the analysis of medical documents. Specifically, transformer-based models showed their efficacy in biomedical text understanding and summarization, so it became possible to efficiently extract information from medical reports [3]. There were also advances in biomedical named entity recognition techniques using SpaCy, BERT, and deep learning methods, which enhanced the identification of different medical entities [4], [5]. For better interpretability, retrieval models and knowledge-based systems have been investigated for generating readable explanations based on medical datasets [11], [12]. Another method that has been widely adopted for extracting text from medical documents is Optical Character Recognition (OCR), especially Tesseract-based techniques [19], [20]. In order to overcome these drawbacks, this paper introduces MediScan AI – an explainable AI framework for automated medical report interpretation. MediScan AI leverages Tesseract OCR, SciSpacy NER, BioBERT Embeddings, and Transformer models like DistilBART and FLAN-T5 in its pipeline. The system further incorporates a knowledge retrieval component based on the MedQuAD dataset and uses TF-IDF and BM25 for ranking relevant explanations for extracted medical terms.

II. LITERATURE REVIEW

The increasing amount of health care data has stimulated research in applying AI and NLP to the analysis of the mentioned data. Medical data such as discharge summaries, lab reports, and clinical records have much information about patients, however, it mostly represents unstructured data that is hard to analyze. Various AI solutions were used to facilitate health care analytics and access to patient's data [1]. One of the directions of research on biomedical NLP is text summarization. Multiple models were created to shorten documents and retain necessary clinical information. Models based on transformers allow one to achieve good summarization results by providing a deeper context of the text [2], [3]. Besides, different evaluation techniques, including automatic and language model-based techniques, have been used [4]. The application of biomedical named entity recognition (NER) is important in the process of generating structured data from unstructured clinical documents. Biomedical NER methods that leverage deep learning have shown promising results in recognizing biomedical entities like diseases, drugs, and surgical operations. For instance, Kumar et al. created a system that employed SpaCy and BERT models to boost the precision of biomedical entity recognition [5], and Dash et al. implemented the use of embedding and neural networks in biomedical NER [6]. The recent trends in AI technologies have emphasized the use of hybrid NLP methodologies along with the explainability aspects of AI. The combination of the rule-based approach and the LLMs for NLP has led to more efficient results regarding the extraction of information from clinical texts [8]. In addition, there have been many studies conducted to simplify medical text using AI, which prove that AI can improve health literacy by transforming complex terminologies into understandable language [9], [10]. While these developments have been made, the majority of current systems still target specific processes, such as document summarization or entity recognition. An obvious void in the research is in formulating an overarching system that incorporates all of these functions into one single process. The MediScan AI framework fills this gap by developing an integrated system for the automated analysis of medical reports.

III. PROPOSED SYSTEM

3.1 System architecture

The architecture of the MediScan AI system, as illustrated in Figure 1, demonstrates how the whole pipeline of the system works when analyzing medical reports automatically. It includes the frontend interface, backend processing, AI processing, and data management parts. The system enables users to upload medical records and images through an interactive interface designed for easy access and use. The backend manages data flow and ensures smooth communication between different components while securely storing patient information. Privacy mechanisms are incorporated to protect sensitive data. An intelligent processing module analyzes the uploaded content by extracting text from documents and interpreting it using language processing techniques. The system then generates concise summaries and simplified explanations to make medical information easier to understand. Finally, it delivers clear reports and meaningful insights to assist users in better understanding healthcare information.

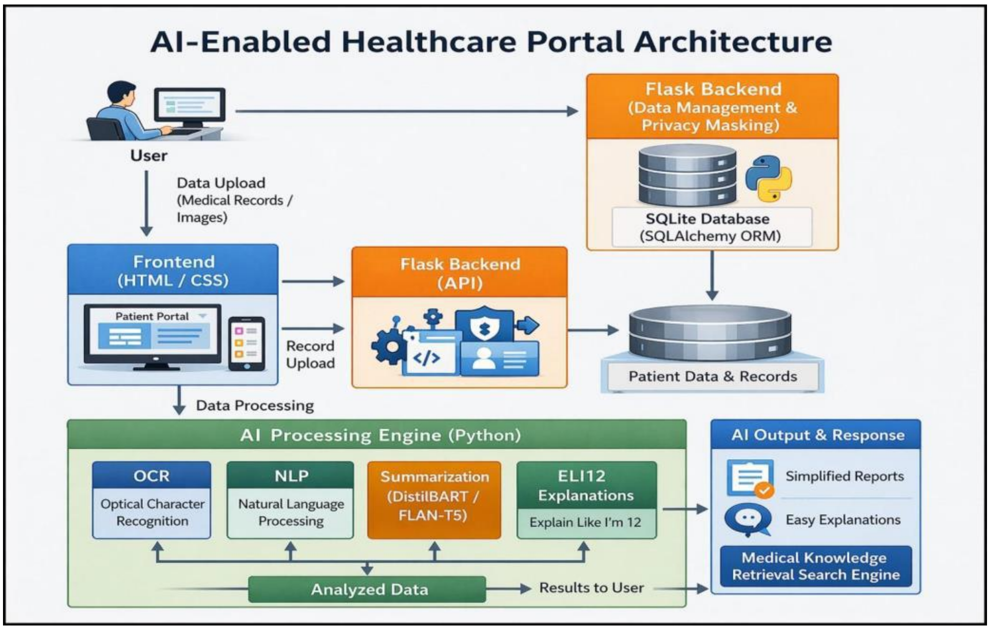

Figure 1. MediScan AI system architecture

On the highest level, the User sends requests to the system by submitting medical reports or images to the system through the Frontend Layer. The Frontend Layer is designed using HTML and CSS. The input data sent from the user side goes to the Flask Backend (API Layer). In the backend, it receives input data requests and routes them to other parts of the system. Moreover, it controls the data transfer process between different layers of the system. There is another Flask Backend layer (Data Management Layer). It controls the process of storing data and masking patients' identities. Sensitive data is processed and stored in a SQLite database using the SQLAlchemy ORM framework. The main element of the system architecture is the AI Processing Engine coded in Python programming language, which is used to analyze the text sequentially during several stages:

• OCR (Optical Character Recognition) – extracts text from the uploaded image files or scanned documents;
• NLP Processing – text cleaning and tokenization and preparation for further processes;
• Summarization – employs transformer models like DistilBART and FLAN-T5 to provide summaries of medical documents;
• ELI12 Explanations – produces simplified descriptions of the medical documents ("Explain Like I'm 12").

Once the text is processed, the resulting Analyzed Data passes to the output stage.
The AI Output and Response Layer is a module that provides the user with the following output:

• Medical reports in a simplified form;
• Easy-to-understand explanations of the reports;
• Medical Knowledge Retrieval Search Engine that can be used to get additional information on the topic.

Thus, one can see how smoothly and seamlessly the entire system combines frontend processes, backend operations, AI processes, and knowledge retrieval.

IV. IMPLEMENTATION

The system MediScan AI that has been suggested is a web-based system for analyzing and interpreting medical reports using artificial intelligence technology. This system comprises three main modules based on the following functions: User Interaction and Medical Report Processing; Medical Report Interpretation Using Artificial Intelligence; and Intelligent Assistive System with Privacy Protection. These modules altogether contribute to the effective medical report processing and information extraction and generate output data.

4.1 User Interaction and Medical Report Processing Module

In this module, there is user interaction control and handling of ingestion and pre-processing of medical reports. In particular, the users interact with the system using a web-based platform for uploading medical documents in either PDF or image format.
On the other hand, after a report has been uploaded, validation is performed, and pre-processing of the document begins. The document is first run through the OCR module (Tesseract). Through this, the

document shall be read even when it is an image since the document will be converted into text. Afterwards, the text goes through various NLP procedures.
Such include:
- Text normalization
- Noise and special characters elimination
- Sentence splitting and tokenization

The above helps in cleaning the text in preparation for further analyses. Moreover, the module also allows report management where the users can access reports already processed via the dashboards.

4.2 Medical Report Interpretation Using Artificial Intelligence

This module carries out the necessary analysis needed in the interpretation of medical reports. The processed text is fed into a chain of AI algorithms responsible for extracting and summarizing the information contained in the clinical document.
First, a biomedical named entity recognition (NER) is conducted using SciSpacy but enhanced by BioBERT embeddings to provide better contextual information about the entity. These include:
- Disease
- Symptoms
- Drugs and Procedures

After the identification of the entities in the text, the AI algorithm summarizes the content of the medical report through transformer-based models such as DistilBART and FLAN-T5. This makes it possible to create an abstractive summary that contains vital clinical information. Also, the AI model offers contextual insights for medical information. Through this, the identified entities and the abstractive summary are used to create explanations regarding the medical information found in the text.

4.3 Intelligent Assistive System with Privacy Protection

This module is responsible for improving the usability and security of the system by providing contextual assistance and ensuring responsible handling of medical information. The module is also integrated with the Flask backend API that allows
for easy interaction between the knowledge base, extraction engine, AI pipeline, and the frontend. One of the most important components of this module is the Medical Knowledge Retrieval Engine, which uses the MedQuAD dataset to fetch relevant medical explanations for the identified entities. When users interact with the system or seek more information about a medical term, the knowledge retrieval module offers contextual answers and explanations. This module is very helpful in filling the gap between technical medical terms and patient-friendly explanations, thus improving health literacy. Extracted knowledge and created explanations are handled via the data storage component of the system. The system also supports smart follow-up questions, which help users navigate more medical information related to the analyzed report. These follow-up questions enable users to seek more information about diagnoses, treatments, and medical terms. Apart from the intelligent assistance module, this module also supports privacy and data protection techniques to ensure that the sensitive medical information is handled securely. The user reports are handled in a controlled environment, and necessary precautions are taken to avoid unauthorized access to medical information. This ensures that the system is in compliance with the data protection principles while offering AI-driven medical analysis. The proposed system, ”MediScan AI,” is different from existing medical natural language processing systems in terms of scope and approach. Unlike existing medical natural language processing systems, which handle a specific sub-task in isolation, such as named entity recognition or summarization, the proposed system is an integrated end-to-end approach to medical report analysis. Unlike existing named entity recognition approaches, which are primarily based on a single model such as SpaCy BERT or Bi-LSTM-CNN, the proposed system is a hybrid approach that combines the SciSpacy biomedical model, a domain-specific phrase matcher, a medical tooltip dictionary containing over 200 terms, and an abbreviation recognition tool, which results in a more accurate approach to medical named entity recognition, considering the variety of medical documents. Unlike existing medical summarization approaches, which are primarily focused on extractive summarization, such as extracting relevant sentences from a medical document, the proposed system is an abstractive approach to medical

summarization, which is further enhanced by a plain language ELI-12 rewriting tool using FLAN T5 to ensure that medical findings are understandable to non-experts.

For knowledge retrieval, most of the existing retrieval-augmented methods depend on a single retrieval technique, such as BM25 or TF-IDF, whereas in the proposed work, a hybrid BM25 and TF-IDF retrieval system is utilized, which is fused by a technique called Reciprocal Rank Fusion, applied on the MedQuAD dataset, for improved relevance in knowledge retrieval. In addition, a privacy masking technique is also included in the proposed work, which is completely missing in the existing medical NLP tools, thereby making it suitable for practical use in clinical applications. The quantitative improvement obtained through these methodological advancements with respect to the existing models.

V. RESULTS AND DISCUSSION

5.1 Dataset Description

The evaluation of the proposed system called MediScan AI is carried out by using a combination of publicly available biomedical datasets and sample medical reports. The proposed system uses a primary knowledge source called MedQuAD (Medical Question Answering Dataset), which is a large dataset containing question and answer pairs from trusted sources such as healthcare organizations like the National Institutes of Health (NIH). It contains around 47,000 question and answer pairs covering a wide variety of medical-related questions and answers in areas such as diseases, symptoms, and diagnostic procedures. The proposed system uses this dataset in its knowledge retrieval module to provide explanations to users about medical terminology present in medical reports. To test the performance of the system, a set of medical documents, such as laboratory reports, diagnostic reports, and summaries, is used. These documents have information related to various medical conditions, such as blood tests, metabolic disorders, and general observations related to diagnosis. These documents go through the OCR module, where the text is extracted, which is then processed using the natural language processing pipeline. The text extracted from these documents is then passed through the biomedical named entity recognition model and the summarization model, which is a transformer-based model. This dataset is used to test the performance of the system, such as its ability to identify entities, summarize, and retrieve information related to medicine. Table 1 shows a summary of the dataset used to test the performance of the MediScan AI system.

Table 1 Dataset Statistics

Dataset Component	Size
Medical reports	120 documents
Extracted text samples	120
Biomedical entities identified	1500+
Knowledge base entries (MedQuAD)	~47,000 QA pairs

5.2 Evaluation Metrics

The performance of the proposed MediScan AI system is evaluated through the use of standard parameters in evaluating the efficiency of a system in natural language processing. The parameters are also used to evaluate the overall efficiency of the system. The parameters include Precision, Recall, F1 score, and response time. Figure 2 shows a comparative evaluation of the proposed MediScan AI system in comparison to existing models, which indicates the improvement in the proposed system over existing models.

Precision:

Precision is the proportion of correctly identified biomedical entities to the total number of entities predicted by the system. A high value for precision means that the number of predictions made by the system is low.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (1)$$

where TP is the number of true positive predictions and FP is the number of false positive predictions.

Recall:

Recall measures the proportion of actual biomedical entities that are correctly identified by the system. Recall tests the capability of the model in identifying all the medical entities in the medical report.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

where FN represents the number of false negative predictions.

F1-score:

The F1-score is defined as a harmonic mean of precision and recall and gives a fair estimate of the model's performance,

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

where high F1-score implies a better performing biomedical named entity recognition model.

System Response Time:

The efficiency of the MediScan AI system is also evaluated by measuring the time it takes for the system to process and analyze a medical report.

$$\text{Response Time} = \text{T}_{\text{output}} - \text{T}_{\text{input}} \quad (4)$$

where T_{input} is the time at which the medical report is uploaded, and T_{output} is the time at which the analysis results are generated. This measure is considered a way to evaluate the efficiency of the processing system in real-world scenarios.

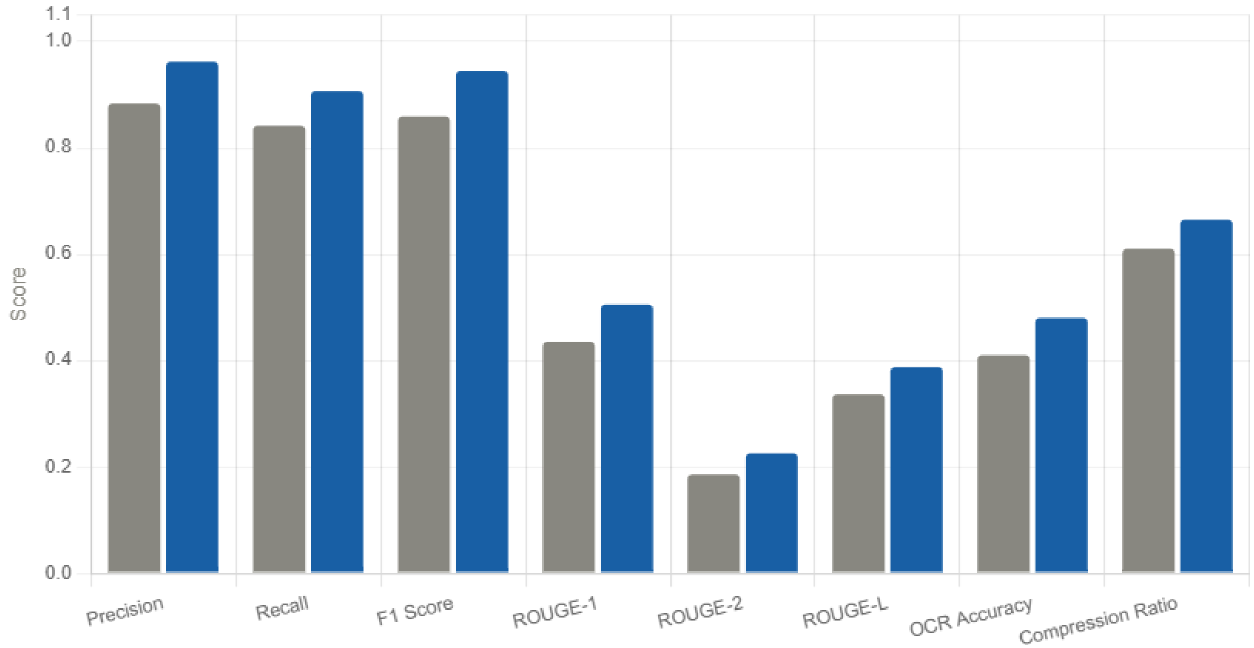

Figure 2. Performance comparison of MediScan AI (blue) against existing models (gray) across NER, summarization, OCR, and compression metrics

The proposed system, MediScan AI, was tested using various natural language processing techniques such as OCR extraction, biomedical named entity recognition, medical text summarization, and knowledge-based question answering. Based on the preliminary testing of various medical report samples, the system was found to be effective in accurately extracting text, recognizing relevant entities, and summarizing medical reports. Furthermore, the system was found to retrieve appropriate contextual medical knowledge to answer questions related to the entities. This system was quantitatively analyzed using several metrics, including Precision, Recall, and F1 score, showing high accuracy in biomedical entity recognition and medical report summarization. The proposed MediScan AI system, is expected to automate the analysis of medical reports through the integration of several functions, including OCR, biomedical named entity recognition, transformer-based summarization, and knowledge retrieval. The function of OCR enables the accurate extraction of text data from uploaded medical records, which can then be processed further through the system's natural language processing function. The function of biomedical named entity recognition can help identify important information regarding diseases, symptoms, medicine, and test results from the extracted text data. The summarization component helps to summarize long medical reports into shorter summaries that preserve the most important clinical information. Additionally, the knowledge retrieval component helps to provide contextual explanations of identified medical terms using the MedQuAD dataset, which helps to improve the understanding of complex medical information.

VI. CONCLUSION AND FUTURE SCOPE

This paper introduced MediScan AI, an intelligent framework for the automated analysis of medical reports using natural language processing and knowledge-assisted retrieval. The proposed framework combines OCR, biomedical named entity recognition, transformer-based summarization, and knowledge retrieval to analyze medical reports. The proposed solution allows users to fetch summaries, pinpoint important entities, and provide explanations for medical terms. The findings of this study show that the proposed system has the potential to help in the interpretation of medical reports and make medical information more accessible. The proposed system has the potential for improvement in the following ways: In the future, the proposed system can be improved by using advanced biomedical language models, increasing the knowledge base, and developing multilingual medical document analysis capabilities.

VII. ACKNOWLEDGEMENT

The authors would like to thank the Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, for providing support in doing the implementation at the Centre for Language Processing.

REFERENCES

[1] Y. Liu, F. Wang, P. Wang, Z. Zhou, H. Wang, J. Li, Y. Qiu, H. Wang, and S. Miao, "AI-MDT: An automatic and intelligent multidisciplinary team consultations platform for lung cancer diagnosis," Journal of Cancer Research and Clinical Oncology, vol. 152, no. 32, 2026.

[2] Y. Vasilev, I. Raznitsyna, A. Pamova, T. Burtsev, T. Bobrovskaya, P. Kosov, A. Vladzymyrskyy, O. Omelyanskaya, and K. Arzamasov, "Evaluating medical text summaries using automatic evaluation metrics and LLM-as-a-judge approach: A pilot study," Diagnostics, vol. 16, no. 1, p. 3, 2026, doi: 10.3390/diagnostics16010003.

[3] Q. Xie, J. A. Bishop, P. Tiwari, and S. Ananiadou, "Pre-trained language models with domain knowledge for biomedical extractive summarization," Knowledge-Based Systems, vol. 252, p. 109460, 2022, doi: 10.1016/j.knosys.2022.109460.

[4] M. Kumar, P. Singh, and P. Kashtriya, "Enhanced biomedical named entity recognition using SpaCy and BERT models," Procedia Computer Science, vol. 258, pp. 1954–1961, 2025.

[5] A. Dash, S. Darshana, D. K. Yadav, and V. Gupta, "A clinical named entity recognition model using pretrained word embedding and deep neural networks," Decision Analytics Journal, vol. 10, Art. no. 100426, 2024.

[6] V. Kocaman and D. Talby, "Accurate clinical and biomedical named entity recognition at scale," Software Impacts, vol. 13, Art. no. 100373, 2022.

[7] M. Asghari, D. Sierra-Sosa, and A. S. Elmaghraby, “BINER: A low-cost biomedical named entity recognition,” Information Sciences, vol. 602, pp. 184–200, 2022.

[8] S. Liu, A. B. McCoy, Q. Chen, and A. Wright, “Integrating rule-based NLP and large language models for statin information extraction from clinical notes,” International Journal of Medical Informatics, vol. 205, Art. no. 106104, 2026.

[9] R. El Shawi and L. Jamel, “Leveraging ChatGPT and explainable AI for enhancing clinical decision support,” Scientific Reports, vol. 15, 2025.

[10] S. Madan, M. Lentzen, J. Brandt, D. Rueckert, M. Hofmann-Apitius, and H. Fröhlich, “Transformer models in biomedicine,” BMC Medical Informatics and Decision Making, vol. 24, no. 214, 2024.

[11] Y. Guo, W. Qiu, G. Leroy, S. Wang, and T. Cohen, “Retrieval augmentation of large language models for lay language generation,” Journal of Biomedical Informatics, vol. 149, Art. no. 104580, 2024.

[12] M. M. Rahman, M. S. Irbaz, K. North, M. S. Williams, M. Zampieri, and K. Lybarger, “Health text simplification: An annotated corpus for digestive cancer education and novel strategies for reinforcement learning,” Journal of Biomedical Informatics, vol. 158, Art. no. 104727, 2024.

[13] Z. Sadeghi, R. Alizadehsani, M. A. Cifci, S. Kausar, R. Rehman, and P. M. Pardalos, “A review of explainable artificial intelligence in healthcare,” Computers & Electrical Engineering, vol. 118, Art. no. 109370, 2024.

[14] I. D. Mienye, G. Obaido, N. Jere, E. Mienye, K. Aruleba, I. D. Emmanuel, and B. Ogbuokiri, “A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges,” Informatics in Medicine Unlocked, vol. 51, Art. no. 101587, 2024.

[15] B. Picton, S. Andalib, A. Spina, B. Camp, S. S. Solomon, J. Liang, P. M. Chen, J. W. Chen, F. P. Hsu, and M. Y. Oh, “Assessing AI simplification of medical texts: Readability and content fidelity,” International Journal of Medical Informatics, vol. 195, Art. no. 105743, 2025.

[16] G. Mariammal, K. Swetha, and S. Jimrys, “Medical report simplification system: Enhancing healthcare accessibility using NLP-based extraction and AI-driven explanation,” in Proc. International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), pp. 541–546, 2025.

[17] J. D. Oliveira, H. D. P. Santos, A. H. D. P. S. Ulbrich, J. C. Couto, M. Arocha, J. Santos, M. M. Costa, D. Faccio, F. O. Tabalipa, and R. F. Nogueira, “Development and evaluation of a clinical note summarization system using large language models,” Communications Medicine, vol. 5, 2025.

[18] A. N. A. Abeo, S. Armstrong, M. Scriney, and H. Goss, “Artificial intelligence techniques and health literacy: A systematic review,” Mayo Clinic Proceedings: Digital Health, vol. 3, no. 4, 2025.

[19] D. Gifu, “AI-backed OCR in healthcare,” Procedia Computer Science, vol. 207, pp. 1134–1143, 2022.

[20] M. Ponnuru, S. Ponmalar, L. A., T. S. B., and G. C. G., “Image-based extraction of prescription information using OCR-Tesseract,” Procedia Computer Science, vol. 235, pp. 1077–1086, 2024.

[21] D. Verma, D. Darlan, and R. Mallipeddi, “Multi-objective optimization of large language model summaries for clinical text narratives,” IEEE Access, vol. 14, pp. 13597–13607, 2026, doi: 10.1109/ACCESS.2026.3656535.