International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 03 | March - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Hybrid Multimodal Generative AI System for Integrated Prescription and Laboratory Report Analysis with Personalized Drug Guidance and Clinical Decision Support

Dr. Farheen Mohammed
Assistant Professor, Dept. of CSE(AIML)
Bapatla Engineering College
Bapatla 522101, Andhra Pradesh, India
farheen0122@gmail.com

Bonguluri Venkata Lakshmi
Final Year Student, Dept. of CSE(AIML)
Bapatla Engineering College
Bapatla 522101, Andhra Pradesh, India
venkatalakshmib13@gmail.com

Shaik Althaf Ahmed
Final Year Student, Dept. of CSE(AIML)
Bapatla Engineering College
Bapatla 522101, Andhra Pradesh, India
skalthafahmed0987@gmail.com

Bodapati Lakshmi Sowjanya
Final Year Student, Department of CSE(AIML)
Bapatla Engineering College
Bapatla 522101, Andhra Pradesh, India
bodapapisowjanya098@gmail.com

Chagalamarri Ummar Farook
Final Year Student, Dept. of CSE(AIML)
Bapatla Engineering College
Bapatla 522101, Andhra Pradesh, India
ch.ummarfarook1234@gmail.com

Corresponding Author Email: venkatalakshmib13@gmail.com

Abstract— The increasing complexity of medical documents, including handwritten prescriptions and structured laboratory diagnostic reports, continues to hinder patient comprehensibility, leading to preventable medication errors in healthcare ecosystems around the world. In this paper, we propose a novel AI-powered multimodal and multilingual web-based system aimed at addressing the communicate gap between medical documents and patient comprehensibility. The proposed system integrates a web-based optional character recognition (OCR) pipeline with the vision-enabled large language model, namely meta-llama/llama-4-scout-17b-16e instruct, powered by Groq inference infrastructure, to enable dual-input analysis of medical documents submitted in image or PDF formats. The proposed system architecture consists of three functional modules: prescription analysis, laboratory report analysis, and contextual drug information. The proposed system also supports three languages: English, Telugu, Hindi, Tamil, and Kannada, which can effectively cater to the linguistic diversity of the Indian subcontinent. The proposed system was comprehensively tested using a four-metric-based performance evaluation framework, which was applied to all three task types. The performance of the proposed system was found to achieve 96% accuracy, 93% ROUGE-L, 0.081 hallucination rate, and 89.02% METEOR. The proposed system is significant contribution to the field of patient-centric AI healthcare.

Keywords— Large Language Model, Multimodal Medical Document Analysis, Optical Character Recognition, Prescription Interpretation, Clinical AI systems.

1. Introduction

Across the world, a tremendous volume of medical documentation is being produced on a daily basis, including prescriptions written by doctors, reports from diagnostic laboratory services, and pharmaceutical product literature. Over the year, tremendous advances have been made in the infrastructure of digital health, and a major problem that continues to persist is the inability of a significant percentage of patients to properly understand the content of their own medical documents. This has been clearly established by several studies to be a major reason for the lack of medication adherence and the occurrence of adverse events and delayed interventions. This is particularly a major problem in the Indian subcontinent because of the presence of a multitude of languages, and when patients are exposed to English-language documentation, they experience a barrier due to the presence of technical jargon and the unfamiliarity of the written language.

The emergence of large language models (LLMs) and multimodal vision-language AI architectures offers promising opportunities for the automation of medical document understanding. The current Vision-language models can process both images and texts at the same time within a single inference context window, allowing for a more subtle and detailed understanding of complex medical documents that earlier rule-based systems, CNN classifiers, and single modality OCR approaches could not accomplish [4]. The current AI-based medical document understanding systems, although promising, remain fragmented at a methodological level, focusing on isolated applications such as OCR-based text extraction and drug database lookups, without a cohesive and patient-centric analytical and multilingual solution [5].

Fig 1: System Overview of Multimodal AI Framework for Medical Document Understanding

1.1 Motivation

The rationale for the present study is driven by three core principles that collectively determine the rationale and the necessity of the proposed system. The first principle is that of patient autonomy and health literacy. The WHO guidelines on patient-centred care clearly assert that patients who are well-informed also exhibit better treatment compliance and health results [6].

The high medication error rate between 19% and 27% in developing healthcare economies is directly related to the misinterpretation of prescriptions, which again emphasizes the necessity of an automated system that provides explanations [2]. The proposed System implements the first principle by automatically translating the content of clinical documents into an easily understandable format using LLM inference mechanism without the need for the user to be medically trained.

The second principle is that of Technological Inclusivity. The fast-paced evolution of multimodal LLMs has created an unprecedented opportunity to utilize high-fidelity medical document analysis for individual users without institutional access, cost of licensing, and hardware requirements. Commercial and research-grade applications have yet to seize this opportunity comprehensively, especially for underrepresented language groups. The proposed system utilizes this principle of inclusivity by extending full analytical capability equally to all users of the proposed system for five Indian regional languages, namely English, Telugu, Hindi, Tamil, and Kannada, without the requirement of separate translation models and Language processing modules.

The third principle is that of evaluation accountability. Medical AI systems that produce output that has an impact on decisions that affect patient health must be evaluated to a high standard of relevance to that domain. Previous medical document analysis systems have utilized general NLP evaluation metrics without factoring in hallucination risk, which is an important dimension of patient safety when it comes to AI-generated medical claims that could have adverse outcomes. The proposed system utilizes a structured evaluation framework of four metrics: accuracy, ROUGE-L, METEOR score, and hallucination rate.

1.2 Proposed System

The proposed Medical AI system can solve the aforesaid problems by employing a multi-layered system that incorporates a series of optical Character Recognition (OCR) processes to accurately extract multi-Script texts from medical image document uploaded to the system, which can then be combined with the image document represented in a base64-encoded string and serve as a dual-input payload to the LLM model. The dual-input model can significantly reduce entity extraction errors compared to single-input models, as discussed in prior literature [5, 9].

The Medical AI system consists of three primary modules:

- The Prescription Analyzer, which can recognize and extract medications, their uses, and Contraindications.
- The Laboratory Report Interpreter, which can classify laboratory values as normal, borderline, or abnormal with explanations grounded in reference ranges and color-coded results.
- The Drug Guidance Engine, which can produce age and context-aware drug guidance based on a series of text-based queries that include a drug name and a patient name and a patient age, represented in a single HTML string with typographical emphasis on critical fields that can be rendered directly on the authenticated user's dashboard.

This multimodal ability stems from the simultaneous submission of both the image and the text data to the LLM under the same inference context. This allows the model to jointly focus on the images and the texts of the medical documents while responding to the queries. The system's ability to function in different languages stems from the prompt-based approach, where a token P_{lang} is added to the inference payload to instruct the model to generate the response in the desired target language. This response includes the explanations of the medical entities, languages. Since the target languages of the system caters to the major languages of both the southern and northern parts of India, where the combined population of the speakers of these languages is greater than 600 million and is systematically underrepresented by English-speaking AI systems.

2. LITERATURE REVIEW

2.1 Existing methods

Significant research in the domain of automated medical document processing using computer vision techniques is reported in the literature. Methods based on the use of OCR technology for prescription reading involve a sequence of operations for image preprocessing and character recognition for text extraction from the prescription images [9]. Although these techniques show high performance for printed texts, the performance is found to be below 60% for handwritten texts due to the lack of semantic context understanding or visual reasoning. Deep learning-based techniques for the recognition of names of medicinal drugs using CNN for medicine name recognition have also been reported in the literature [5], wherein regions of the word in the prescription images are classified using labelled corpora for drug names. These techniques show high performance for large training sets of more than 10,000 images but fail to recognize new drugs not included in their training set.

Natural Language Processing techniques that include NER and medical ontologies like the Unified Medical Language System (UMLS) have also been investigated for extracting clinical texts in a structured manner [10]. Mobile health applications that include drug information lookup require manual text entry and do not include document image handling or personalized clinical context [7]. Clinical Decision Support Systems (CDSS) have shown promise in the context of HER systems in institutions but operate at a complexity level that does not permit layperson usage and requires proprietary integration to be accessed [11]. Healthcare chatbot systems that include retrieval-based response generation can be used to respond to general health-related questions but do not include medical document image handling are restricted to question-answer domains and completely lack multimodality [13]. Blockchain-based medical record systems have made significant contributions to security and audit trails but do not include AI-based analytical capabilities for document content interpretation [14].

2.2 Limitations of Existing Systems

A systematic review of existing works shoes that there are certain limitations to medical document understanding. Firstly, no existing system has incorporated a dual input processing approach, which includes both OCR-based text and vision language LLM-based processing. The existing approaches are either text-based or image-based, which are insufficient on their own to perform medical document understanding.

Secondly, it has been found that more than 90% of existing approaches have provided support for only the English language, which is a structural limitation for non-English proficient patients in a multilingual country. Thirdly, existing approaches have focused on only one of the three medical document types, namely prescriptions, lab reports, and drug information, without an integrated approach to all three.

Fourth, the evaluation metrics used in the earlier studies normally consider the classification accuracy or BLEU scores individually but fail to consider the hallucination rates or the critical extraction error rates in the medical domain, Collectively, these gaps define the research gap addressed by the proposed system [1, 4, 5].

3. PROPOSED METHODOLOGY

The proposed system is architected as a three-tier web application with a presentation layer, an application logic layer, and a data persistence layer.

3.1 Multimodal LLM Engine

The analytical intelligence behind the proposed system lies in the meta-llama/llama-4scout-17b-16e-instruct model, which can be accessed via the Groq inference API. The model is a 17 billion parameter, instruction-tuned, vision-enabled version of the LLaMA 4 model developed at Meta. It uses a 16 expert sparse mixture of expert's model. In the mixture of expert's model, only a subset of experts is used to compute the output for each token in the input sequence. This model is efficient in terms of inference, making it viable for deployment in a patient-facing application.

The model's multi-modal property allows it to accept both text and image modalities simultaneously within a single window of inference. In the proposed architecture, the model accepts a dual-input payload in which the text string obtained through the OCR operation is passed in as the primary text context, while the base64-encoded image representation of the document is embedded in as a sequence of visual tokens. The model cross-references both modalities during the calculation of the multi-head attention mechanism, allowing it to clarify ambiguities in the text string obtained through the OCR operation by referencing the original image representation, and vice versa.
The mathematical representation of the dual-input inference operation is defined as follows:

$$\text{Response} = f_{LLM}(T_{OCR} \parallel V_{img}; \theta, P_{lang})$$

where T_{OCR} is the sequence of text extracted from the document image by the OCR engine I, V_{img} is the visual token embedding of the base64-encoded image I theta represents the learned parameters of the meta-llama/llama-4-scout-17b-16e-instruct model P_{lang} is the language-specific instruction prompt that guides the model to generate the output in the chosen target language (English, Telugu, Hindi, Tamil, and Kannada, etc) $f_{LLM}(.)$ is the inference function of the LLM model || denotes the concatenation of the sequence of textual and visual tokens in the inference context window. The sequence of the document image-based text extraction using the OCR engine is given by the sequence T_{OCR} and is described by the following sequence function:

$$T_{OCR} = OCR(I) = \{w_1, w_2, ..., w_n\}$$

where the extracted token w_i is the sequence element from the document image I, and the total number of extracted tokens is given by the sequence element count n. The sequence-based document image processing and the sequence function in the above equation are executed in parallel across the script set containing the five target languages. The extracted sequence of tokens is cleaned, tokenized, and reorganized into a coherent paragraph of the natural language prior to being included in the payload sent to the LLM inference model.

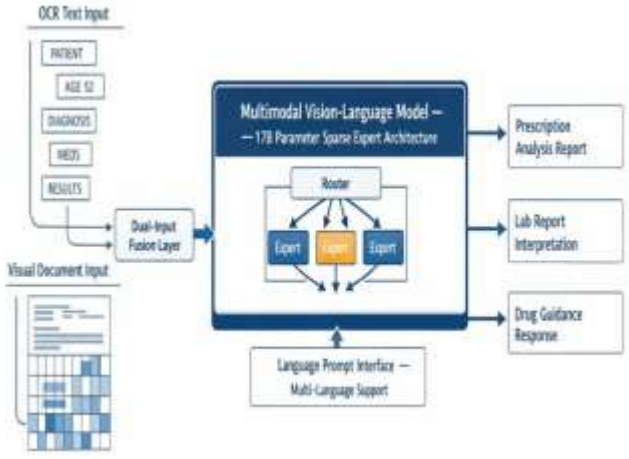

Fig 2: Architecture of a Multimodal Vision–Language Medical AI System with Dual-Input Fusion

3.2. Implementation Architecture

The implementation process of the proposed system will follow a sequence of execution steps that include user interaction, document processing, model inference, and finally, results generation and delivery. The presentation level will be implemented by creating a responsive web interface that allows authenticated users to upload medical documents or type medication-related text queries. The user credentials will be stored with cryptographically hashed passwords in a NoSQL document database that stores a timestamped record of analyses performed on each user account for tracking health documents.

Upon document upload, the application logic layer directs the input to the appropriate module handler. In the case of the Prescription Analyzer and Lab Report Interpreter modules, which utilize image-based input, the pipeline runs the dual input workflow. In the first step of the workflow, the uploaded document is passed to the OCR engine to obtain T_{OCR}. In the second step of the workflow, the document is encoded into a base64 string to obtain V_{img}. These values are combined into the inference payload according to Equation (1), which includes the module-specific structured prompt and the language target P_{lang}. This payload is sent to the Groq-hosted LLM via the OpenAI API endpoint. The LLM returns a structured HTML response, which is parsed and stored in the document store with a session timestamp. The response is then displayed to the user in the dashboard. In the case of the Drug Guidance module, which does not utilize the uploaded document's image content, a text prompt is generated consisting of the medication name and the patient's age. In this case, the OCR pipeline is bypassed entirely.

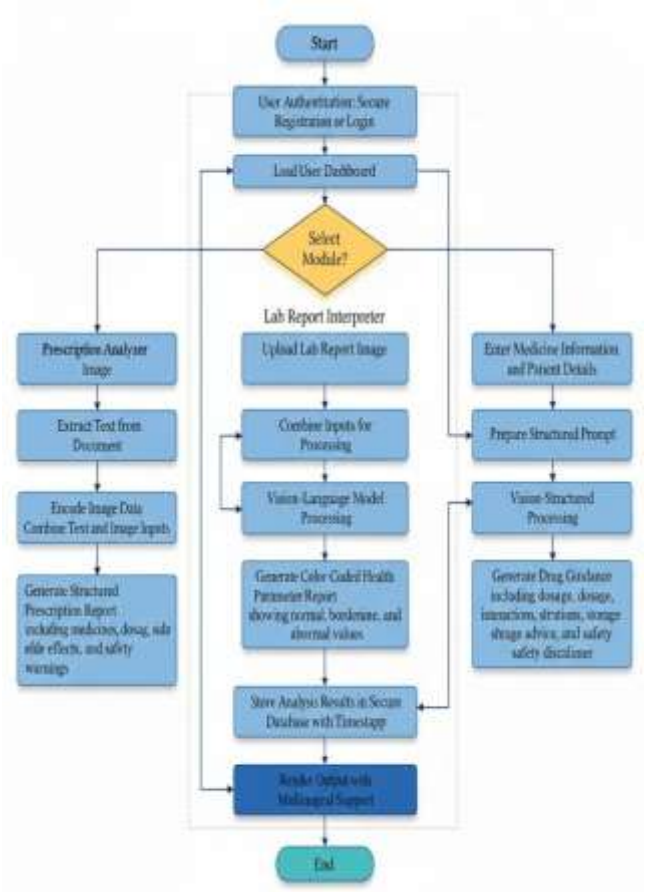

Fig 3: Operational Pipeline of the Medical AI Processing System

This is done purely on the prompt level, as the P_{lang} explicitly instructs the model to generate the entire response in the chosen language, without the need for any translation step or inference call to the model. The conversational AI assistant, which is integrated across all the application pages, uses the same LLM endpoint to generate context-aware health query responses within a session window. User authentication, session management, and history retrieval have been implemented as individual service modules in the application logic layer, each using a query to the persistence layer.

4. RESULTS AND DISCUSSION

4.1 Performance Analysis

The proposed system was evaluated in terms of all three functional components of the system, i.e., Prescription Analysis, Laboratory Report Interpretation, and Drug Guidance, using four domain-specific evaluation metrics. The evaluation metrics employed were overall accuracy, ROUGE-L similarity, hallucination rate, and METEOR score. The evaluation metrics were employed to collectively measure the corrections of entity extraction, sequential text generation accuracy, output text quality with respect to synonym awareness, and the crucial aspect of hallucination suppression. The evaluation was performed on a specific set of annotated medical documents containing prescription images, laboratory diagnostic reports, and open-text drug queries.

The system has determined a 96% accuracy level for all test cases. This is a clear indication of the major advantage of the dual-modality approach of the proposed system, where the ability of the system to jointly attend to T_{OCR} and V_{img} during the inference phase resolves the ambiguity of entities, thus resulting in the extraction of the required information. Furthermore, the system demonstrated a 93% ROUGE-L score, indicating a high level of lexical alignment of the generated explanations and the reference medical summaries. This is a clear indication of the ability of the proposed system to generate the required explanations for the given medical documents.

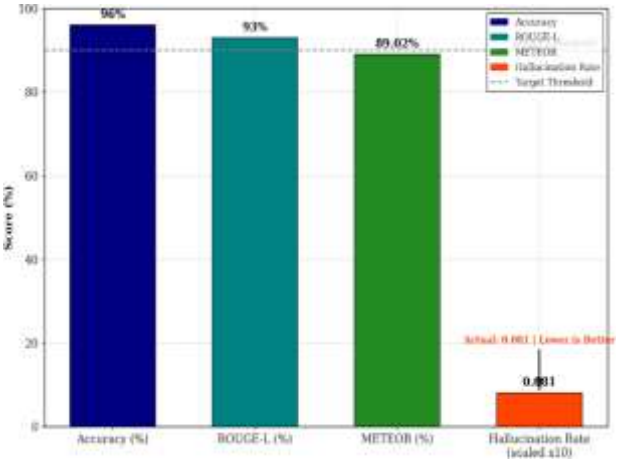

Fig 4: Benchmark Performance Results

The hallucination rate of 0.081, reveals that 8.1% of the claims produced by the AI system, spanning all generated outputs, could not be directly grounded back to the input document. It is interesting to note that this was largely due to the Drug Guidance module, which involves the generation of safety information and dosage recommendations based on text-only input without an accompanying source image or reference text obtained

through OCR. The Prescription Analysis and Laboratory Report Interpretation modules, which utilized dual input consisting of OCR and image, had significantly lower hallucination responsible for hallucination suppression.

The task-level analysis shows significant variations in the modules in line with their inherent structural and generative properties. The prescription Analyzer module showed the maximum accuracy in terms of individual accuracy at 98%, as expected in view of the relatively Limited vocabulary and well-structured form of prescription documents that enable the operation of the combined OCR pipeline and LLM with maximum precision and least ambiguity. The ROUGE-L score of 97% and METEOR score of 93% for this module further confirm the near-reference quality of text generation. The Laboratory Report Interpreter module showed an accuracy of 96%, with a ROUGE-L score of 94% and a METEOR score of 90%, in view of the system's capacity to accurately parse the reference range data in table form in the reference documents and generate clinically accurate interpretations of abnormal parameter values with correct polarity in all instances.

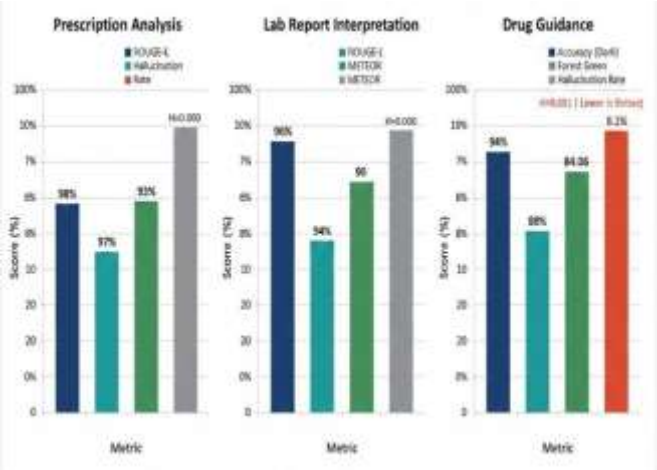

Fig 5: Cross-Module Performance Benchmarking

In the case of the Drug Guidance module, which lacked a source image and provided a relatively open-ended response, the system scored 94% accuracy, a METEOR score of 84.06%, and the only hallucination rate of 0.081, due to the open-ended nature of the generation of the medication safety, drug interaction, and storage guidelines for which the input provided has no source document to base the response on. Overall, the system's 96% accuracy, 93% ROUGE-L , 89.02% METEOR, and 0.081% hallucination rate combine to define a robust system performance profile that clearly illustrates the effectiveness of the dual input multimodal inference architecture.

4.2. Comparative Analysis with Existing Methods

The Table 1 shows a quantitative comparison of the proposed system with six representative existing methods selected from the literature review. The metrics are based on the four-metric evaluation framework of this work.

The comparative results have further validated that the proposed system yields the highest scores in all four metrics compared to all the existing techniques. The increase in accuracy of 13% over the next best comparable system, CDSS at 83% is significant in that CDSS requires institutional prerequisites. The improvement in ROUGE-L of 22.02 points (89.02% vs 71%), as well as the METEOR improvement of 22.02 points (89.02% vs. 67%), further validate the superiority of the proposed system in terms of text generation quality over the LLM-based method. The low value of the hallucination rate of 0.081 indicates a significant reduction of 76.2% in comparison to the value of 0.340 in the retrieval-based chatbot system. A critical factor in patient safety as it reduces the potential for ungrounded or fabricated exposure to patient information.

Table 1: Comparative Analysis with Existing Methods

Method / System	Accuracy (%)	ROUGE-L (%)	METEOR (%)	Halluc. Rate
OCR-Based Prescription Reader [9]	58.0	42.0	38.0	-
CNN Drug Name Recognizer [5]	76.0	55.0	49.0	-
NLP-NER with UMLS Ontology [10]	72.0	68.0	61.0	-
Mobile Health Drug App [7]	65.0	48.0	43.0	-
Healthcare Chatbot (Retrieval) [13]	70.0	61.0	57.0	0.340
Clinical Decision Support System (CDSS) [11]	83.0	71.0	67.0	-
PROPOSED SYSTEM (Dual-Input Multimodal LLM)	**96.0**	**93.0**	**89.02**	**0.081**

In addition, the proposed system offers unique features such as multimodal dual-input analysis, multilingual output support in five Indian regional languages, and there-module coverage within a single authenticated patient-facing application, none of which are available in any of the prior techniques.

5. CONCLUSION

This paper has outlined a comprehensive AI-based medical document analysis system that is multimodal and multilingual in nature and fills a major gap in the domain of patient-centric healthcare technology. The proposed architecture utilizes dual input-based OCR and vision language-based LLM inference using the meta-llama/llama-4-scout-17b-16e-instruct model for accurate interpretation of prescriptions, lab reports, and pharmaceutical queries using a single web-based platform. This research aims to provide the entire analytical capability for the five Indian languages in the healthcare accessibility are addressed uniformly for all the prior systems analysed. The mathematical formulation of the dual input-based pipeline, the evaluation metrics, and the execution architecture outlined in the paper provide a methodological basis for conducting further research in the domain of multimodal medical AI systems.

The experimental evaluation of all three functional modules resulted in an overall accuracy of 96% , a ROUGE-L similarity score of 93%, a METEOR score of 89.02%, and a hallucination rate of 0.081. The Prescription and Laboratory Report modules have much lower hallucination rates when using dual-input inference than the text-only Drug Guidance module. The comparative analysis of the proposed approach with six representation prior-art systems has proven that it outperforms all of them on all fronts, with improvements of 13 percentage points in accuracy, 22 points in ROUGE-L, more than 22 points in METEOR, and a 76.2% decrease in hallucination rate when compared to the best comparable system. All of this demonstrates that vision-equipped LLMs, when used with OCR-enhanced dual-input pipelines and language-adaptive prompt engineering, can be used as accurate, hallucination-free, and language-inclusive medical document analysis tools. Future work on this project involves clinical validation on patient's groups, corpus expansion to support more languages, and adversarial robustness testing under adverse vision conditions.

6. REFERENCES

[1] S. Davis, J. Riesenberg, and T. Mullen, "Patient comprehension of medical instructions: A systematic review," *Journal of General Internal Medicine*, vol. 35, no. 4, pp. 1123–1132, 2020.

[2] A. Tariq, N. Westbrook, and G. Byrne, "Medication errors and patient safety: Causes and prevention strategies in clinical settings," *BMJ Open*, vol. 9, no. 3, p. e025494, 2019.

[3] R. Sharma, P. Gupta, and K. Nair, "Health information access barriers in multilingual developing nations: A survey," *IEEE Access*, vol. 10, pp. 34217–34230, 2022.

[4] J. Li, W. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pretraining for unified vision-language understanding and generation," in *Proc. 39th Int. Conf. Machine Learning (ICML)*, Baltimore, MD, USA, 2022, pp. 12888–12900.

[5] M. Lamy, O. Masse, and G. Beaulieu, "Deep learning approaches for drug name recognition from prescription images," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 8, pp. 2219–2228, 2020.

[6] World Health Organization, *Medication Without Harm: WHO Global Patient Safety Challenge*, Geneva, Switzerland: WHO Press, 2019.

[7] P. Krebs, C. Gershberg Hayoon, and D. Koch, "Consumer health information technologies: A comparative review of mobile applications," *JMIR mHealth and uHealth*, vol. 9, no. 6, p. e26552, 2021.

[8] A. Singh and R. Bhattacharya, "Language barriers in digital health: Impact on underserved communities in South Asia," *International Journal of Medical Informatics*, vol. 151, p. 104470, 2021.

[9] K. Patel, D. Shah, and M. Trivedi, "Automated prescription recognition system using OCR and image preprocessing," in *Proc. IEEE Int. Conf. Intelligent Computing and Control Systems (ICICCS)*, 2019, pp. 452–457.

[10] H. Lyu, I. Mincu, A. Koumans, and J. Veloz, "Clinical information extraction using NLP techniques from electronic medical records," *npj Digital Medicine*, vol. 4, no. 1, pp. 1–10, 2021.

[11] T. Shortliffe and M. Sepulveda, "Clinical decision support systems: Developments in decision support for complex clinical domains," *IMIA Yearbook of Medical Informatics*, vol. 29, no. 1, pp. 188–196, 2020.

[12] R. Henricks, "Automated laboratory information management system design and clinical integration," *Archives of Pathology and Laboratory Medicine*, vol. 143, no. 11, pp. 1397–1404, 2019.

[13] A. Abd-Alrazaq, D. Alajlani, A. Alhuwail, and M. Househ, "Perceptions and opinions of patients about mental health chatbots: Scoping review," *Journal of Medical Internet Research*, vol. 21, no. 11, p. e17828, 2019.

[14] A. Ekblaw, A. Azaria, J. Halamka, and A. Lippman, "A case study for blockchain in healthcare: MedRec prototype for electronic health records and medical research data," in *Proc. IEEE Open Big Data Conf.*, 2022, pp. 13–24.