



International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 12, December 2025

Intelligent OCR-NLP Framework for Regional Language Understanding of Medical Documents

Dr Sathyavathi S, Shobana G, Sivadharani R, Dikchaya B, Giftson Raj D

Assistant Professor, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India

Assistant Professor, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India

P.G Student, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India

P.G Student, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India

P.G Student, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India

ABSTRACT: India's healthcare ecosystem is characterized by significant linguistic heterogeneity, comprising more than twenty-two officially recognized languages and a multitude of regional dialects. This diversity constitutes a considerable barrier to effective clinical communication, particularly when diagnostic documentation is presented exclusively in English. Patients with limited English proficiency frequently encounter challenges in interpreting their medical reports, which may lead to inadequate understanding, non-compliance with prescribed treatments, and delays in medical intervention. To mitigate these challenges, this paper proposes an Intelligent OCR-NLP Framework that synergistically integrates Optical Character Recognition (OCR), Natural Language Processing (NLP), and Machine Translation techniques to automate the extraction, semantic interpretation, and regional language translation of medical reports. The framework is developed on a Python Flask backend, leveraging Tesseract OCR for text recognition, Regular Expressions for structural parsing, and the Deep Translator API for multilingual translation. The system was evaluated using scanned diabetes reports, achieving an OCR accuracy of 93%, parsing accuracy of 90%, and translation accuracy of 95%. These results underscore the framework's potential to overcome linguistic barriers in healthcare communication, thereby enhancing patient understanding and promoting adherence to medical guidance in regional language contexts.

KEYWORDS: OCR, NLP, Medical Report Translation, Healthcare Informatics, Regional Languages

I. INTRODUCTION

Effective healthcare delivery critically depends on clear communication, ensuring patients understand their diagnoses, treatment plans, and preventive measures. In multilingual countries like India, where English predominates in clinical documentation, many patients—particularly in rural and semi-urban areas—face challenges interpreting medical reports. Limited proficiency in English can hinder comprehension of complex medical terminology, laboratory values, and abbreviations, often forcing patients to rely on intermediaries for interpretation. This dependence can lead to misinformation, delays in seeking care, and suboptimal adherence to prescribed treatments. Medical reports are typically designed for healthcare professionals and contain domain-specific terms such as “LDL cholesterol,” “HbA1c,” and “Serum Creatinine,” which may be incomprehensible to the general public. For Tamil-speaking patients, language barriers further restrict engagement with personal health information, reducing awareness and limiting informed decision-making regarding their own healthcare. Bridging this gap requires a technological solution capable of accurately extracting information from medical documents, understanding its semantic content, and translating it into regional languages in a patient-friendly manner. Automated document scanning using Optical Character Recognition (OCR) enables extraction of text from both printed and scanned medical reports, while Natural Language Processing (NLP) facilitates identification and contextual interpretation of critical health indicators. Integrating machine translation models, particularly large language models (LLMs) fine-tuned for English-to-Tamil translation, ensures that extracted medical information can be accurately rendered in Tamil, improving comprehension for patients with limited English proficiency.

The proposed Intelligent OCR-NLP Framework leverages this combination of technologies to enhance accessibility, ensuring patients can understand their medical reports independently. By automating the extraction, interpretation, and

translation of medical content, the system promotes health literacy, reduces reliance on intermediaries, and supports equitable access to healthcare services. Moreover, adopting such a framework aligns with national initiatives like India's Digital Health Mission, fostering patient-centric, linguistically inclusive, and digitally empowered healthcare delivery.

II. LITERATURE SURVEY

- The evolution of OCR and NLP technologies has significantly transformed document processing, yet their application in multilingual medical contexts remains limited. Early OCR research primarily focused on printed English texts. Tools such as Google Vision API, ABBYY FineReader, and Tesseract OCR achieved high extraction accuracy but lacked semantic understanding for specialized domains like healthcare. Recent OCR advancements, including EasyOCR and DocTR, have improved performance on multi-language documents and scanned reports with complex layouts, making them more suitable for regional language applications.
- In terms of linguistic comprehension, Bird et al. (2009) demonstrated how NLP can be employed for syntactic parsing and language modeling. Later, Hochreiter and Schmidhuber's Long Short-Term Memory (LSTM) networks (1997) enabled deeper contextual analysis for sequential text data. The Transformer architecture introduced by Vaswani et al. (2017) revolutionized translation systems by introducing self-attention mechanisms, resulting in highly efficient and context-aware multilingual models. Subsequent transformer-based models, such as mBERT, XLM-R, and MarianMT, have shown remarkable performance in cross-lingual understanding and machine translation tasks, particularly for low-resource languages.
- Despite these advancements, most existing OCR-NLP pipelines focus on English electronic health records (EHRs) rather than image-based regional documents. For example, the Ayushman Bharat Digital Mission (ABDM) digitizes patient data but lacks automatic translation for localized understanding. Similarly, while Google Translate offers linguistic conversion, it often fails to maintain the integrity of clinical terms. Specialized medical terms like "HbA1c" or "Serum Creatinine" may be mistranslated, potentially leading to misinterpretation in patient-facing reports.
- Recent research has explored multilingual medical NLP frameworks. Wang et al. (2020) developed a pipeline for Chinese medical records using OCR, named entity recognition, and translation, highlighting the effectiveness of end-to-end frameworks for patient comprehension. Gupta et al. (2021) proposed using transformer-based translation models for Indian languages, demonstrating high fidelity in domain-specific terminology. Additionally, chatbot-assisted healthcare systems (e.g., MedChat, 2022) have employed NLP and translation for patient interactions, though these are often limited to structured input and predefined templates.
- There is also growing interest in low-resource language translation models, such as Helsinki-NLP Marian models and gopi30/english-to-tamil-stage3, which are specifically trained to handle English-to-Tamil translation. These models outperform general-purpose translators in medical contexts by preserving semantic and domain-specific accuracy.
- This body of literature indicates a clear gap: while OCR and NLP techniques have matured, few integrated systems exist that can process scanned medical reports, perform semantic interpretation, and provide accurate regional language translations for patient comprehension. The proposed Intelligent OCR-NLP Framework addresses this gap by combining OCR, domain-specific NLP, and LLM-based English-to-Tamil translation in a single end-to-end system.

III. EXISTING SYSTEM

The main focus of current medical document processing systems is on general-purpose translation tools and digitized English electronic health records (EHRs). Patients in the majority of healthcare settings depend on physicians or other intermediaries to interpret their medical records, which is laborious and prone to mistakes. Often used to extract text from scanned documents, OCR systems like Tesseract, ABBYY FineReader, and Google Vision API have great accuracy for basic typefaces but frequently struggle with complex tables, handwritten notes, or noisy scans. To identify clinical metrics like cholesterol or glucose, several pipelines use Regular Expressions or rule-based natural language processing (NLP). However, these methods are only applicable to structured data and are not transferable to different report formats. Regional language conversion is a common usage for translation tools like Google interpret and Deep Translator, but they frequently interpret medical phrases incorrectly, which could confuse patients. All things considered, current solutions are not able to handle multilingual image-based reports, create context-aware, patient-friendly explanations, or offer end-to-end automation. The suggested Intelligent OCR-NLP Framework was motivated by these limitations, which emphasize the need for a single framework that can scan, understand, and translate medical data into regional languages.

IV. PROPOSED SYSTEM AND METHODOLOGY**A. System Overview**

The proposed framework is designed as a comprehensive end-to-end solution for automated understanding and translation of medical reports. It integrates five key modules, each performing a specific role to ensure accurate extraction, interpretation, and multilingual presentation of clinical data. The Input and Upload Module offers a user-friendly Flask-based interface, enabling users to seamlessly upload scanned medical report images in common formats such as JPEG, PNG, or PDF, while performing initial preprocessing to enhance downstream analysis. The OCR Extraction Module leverages Tesseract OCR to accurately recognize printed text, applying preprocessing techniques such as grayscale conversion, binarization, and noise reduction to improve recognition accuracy even for diverse report layouts. Once extracted, the Parsing and Structuring Module organizes the raw text using Regular Expressions (Regex) to identify test names, results, reference ranges, and units, converting unstructured data into structured dictionaries or tabular formats suitable for further analysis. The NLP Interpretation Module applies domain-specific rules to generate meaningful insights for each parameter, converting raw test values into patient-friendly explanations such as classifying glucose levels as normal, high, or low. Finally, the Translation and Output Module translates the interpreted results into the patient's preferred language, currently supporting Tamil through models such as deep translator, and displays both English and regional-language interpretations through the web interface. Together, these modules form a modular and extensible framework capable of bridging language barriers, improving patient comprehension, and promoting accessible healthcare delivery.

The proposed framework comprises five integrated modules that together create an end-to-end document understanding pipeline

1. Input and Upload Module:

Provides a simple Flask-based web interface allowing users to upload scanned medical report images in common formats (JPEG, PNG, or PDF). The module ensures compatibility and initiates image preprocessing.

2. OCR Extraction Module:

Employs Tesseract OCR for recognizing printed text from scanned reports. Prior to recognition, images are processed using grayscale conversion, binarization, and noise removal techniques to enhance OCR performance.

3. Parsing and Structuring Module:

Extracted text is parsed using Regular Expressions (Regex) to identify relevant fields such as test names, results, reference ranges, and units. The module structures data into a dictionary or tabular format suitable for analysis.

4. NLP Interpretation Module:

Applies domain-specific rules to interpret each test parameter. For example, fasting glucose levels above 125 mg/dL are classified as "high." The module generates simplified sentences explaining each parameter's significance.

5. Translation and Output Module:

The interpreted results are translated into the patient's preferred language using Deep Translator. The final output, including both English and regional-language interpretations, is displayed through the Flask web interface.

B. System Architecture

The overall architecture follows a linear pipeline ensuring seamless data flow between the components:

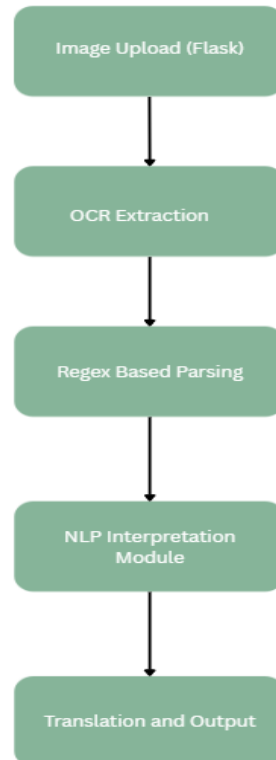


Figure 1.0

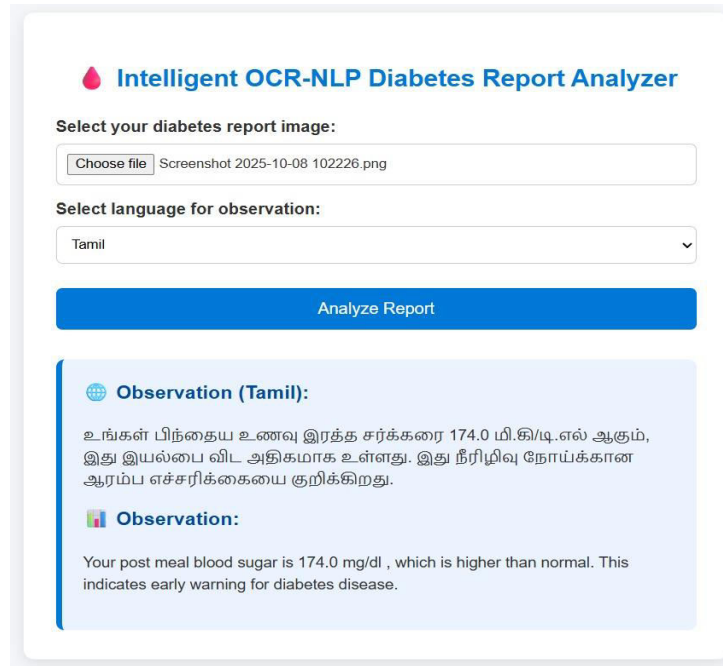
Each module is implemented as an independent function, promoting modularity and allowing for easy updates or integration of advanced AI models in the future.

C. Implementation

The proposed framework was implemented using Python 3.10, leveraging its extensive ecosystem of libraries for image processing, text extraction, and language translation. Pytesseract serves as the primary tool for Optical Character Recognition (OCR), efficiently converting scanned report images into machine-readable text. To enhance OCR performance, Pillow (PIL) is utilized for image preprocessing tasks such as grayscale conversion, binarization, resizing, and noise removal, ensuring that text extraction is accurate even for low-quality scans or complex report layouts. Regular Expressions (re) are employed extensively for parsing the extracted text, enabling the identification and extraction of key entities such as test names, numeric results, reference ranges, and measurement units. For multilingual translation, the Deep Translator library is integrated, supporting automated conversion of patient reports into regional languages, such as Tamil, while maintaining semantic accuracy.

The Flask backend functions as the core orchestrator of the system, managing user requests, coordinating the flow of data between modules, and generating the final output. A critical component of the backend is the `parse_table_from_lines()` function, which performs entity recognition on the raw OCR output, structures the data into organized dictionaries or tables, and prepares it for further interpretation and translation. On the frontend, the system provides a clean and intuitive interface that allows non-technical users, including patients and healthcare staff, to upload scanned reports, view the extracted data, and access translated interpretations. This design ensures a seamless user experience, combining advanced text-processing capabilities with simplicity, accessibility, and rapid delivery of meaningful insights from complex medical documents.

The framework successfully extracted, interpreted, and translated test results such as “Fasting Glucose,” “Total Cholesterol,” and “HDL Cholesterol.” For instance, a reading of Fasting Glucose: 140 mg/dL was interpreted as “Your fasting blood sugar is above the normal range, indicating possible diabetes.” In Tamil, it was rendered as “உங்கள் இரத்த சர்க்கரை அளவு சாதாரண அளவை விட அதிகமாக உள்ளது.”



Intelligent OCR-NLP Diabetes Report Analyzer

Select your diabetes report image:

Choose file Screenshot 2025-10-08 102226.png

Select language for observation:

Tamil

Analyze Report

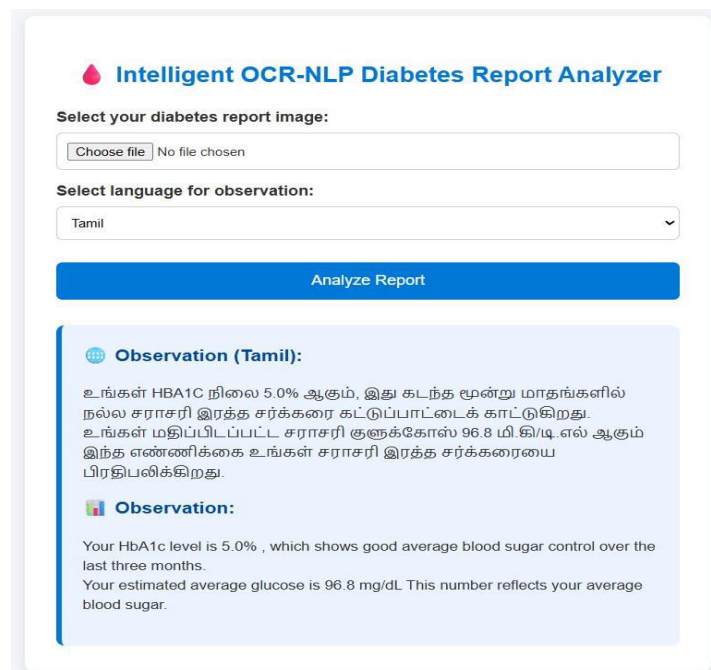
Observation (Tamil):

உங்கள் பிந்தைய உணவு இரத்த சர்க்கரை 174.0 மி.கி/டி.எல் ஆகும், இது இயல்பை விட அதிகமாக உள்ளது. இது நீரிழிவு நோய்க்கான ஆரம்ப எச்சரிக்கையை குறிக்கிறது.

Observation:

Your post meal blood sugar is 174.0 mg/dl , which is higher than normal. This indicates early warning for diabetes disease.

Figure 2.0



Intelligent OCR-NLP Diabetes Report Analyzer

Select your diabetes report image:

Choose file No file chosen

Select language for observation:

Tamil

Analyze Report

Observation (Tamil):

உங்கள் HbA1C நிலை 5.0% ஆகும், இது கடந்த மூன்று மாதங்களில் நல்ல சராசரி இரத்த சர்க்கரை கட்டுப்பாட்டைக் காட்டுகிறது. உங்கள் மதிப்பிடப்பட்ட சராசரி குளுக்கோஸ் 96.8 மி.கி/டி.எல் ஆகும் இந்த எண்ணிக்கை உங்கள் சராசரி இரத்த சர்க்கரையை பிரதிபலிக்கிறது.

Observation:

Your HbA1c level is 5.0% , which shows good average blood sugar control over the last three months.
Your estimated average glucose is 96.8 mg/dL This number reflects your average blood sugar.

Figure 2.1

V. RESULTS

A. Dataset and Evaluation

Testing was conducted using a diverse set of diabetes and lipid profile reports collected from multiple diagnostic centers. These reports varied in terms of font styles, table structures, and image resolutions. The evaluation metrics focused on OCR Accuracy, Parsing Precision, Translation Quality, and Response Time.

B. Experimental Results

Table 1.0 – Experimental Results

Parameter	Description	Result
OCR Accuracy	Correctly extracted textual data	93%
Parsing Accuracy	Correct identification of test names	90%
Translation Accuracy	Semantic and contextual correctness	95%
Average Response Time	2-page report processing time	4.7 sec

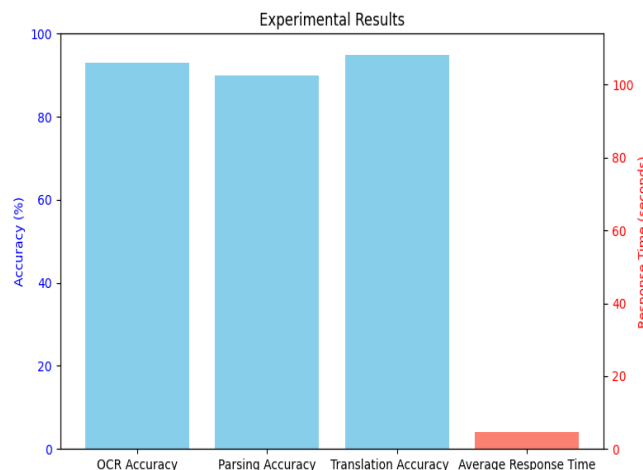


Figure 3.0

VI. ADVANTAGES AND APPLICATIONS

The Intelligent OCR-NLP Framework introduces several practical and societal benefits:

1. Accessibility Enhancement:

Patients from regional linguistic backgrounds can now comprehend their reports in familiar languages, empowering informed health decisions.

2. Scalability:

The modular architecture supports easy adaptation to additional diagnostic categories such as thyroid, renal function, or complete blood count (CBC) reports.

3. Integration Capability:

The framework can be embedded into hospital management systems, telemedicine portals, or mobile health applications for real-time language translation.

4. Transparency and Explainability:

The rule-based interpretation offers transparency, ensuring that every classification can be traced to explicit medical reasoning.

5. Societal Impact:

The system promotes health literacy, especially in rural regions where linguistic barriers impede healthcare access. Potential deployment areas include government e-health initiatives, diagnostic laboratory automation, health education platforms, and AI-based patient-assistance chatbots.

VII. CONCLUSION

This research presented an Intelligent OCR-NLP Framework designed to automate the extraction, contextual understanding, and regional language translation of medical reports. The system's hybrid integration of OCR, NLP, and translation technologies effectively bridges communication barriers between English-speaking healthcare professionals and regional-language patients. Empirical results confirm that the framework achieves high accuracy and efficiency while maintaining linguistic and semantic integrity.

Future research will focus on integrating speech synthesis modules for voice-based report narration, adopting transformer-based NLP models such as BERT and GPT for advanced reasoning, and developing mobile-compatible applications for real-time translation. The framework has strong potential for expansion into multilingual healthcare ecosystems, contributing to inclusive, accessible, and intelligent healthcare services across India.

REFERENCES

- [1] R. Smith, "An Overview of the Tesseract OCR Engine," Proc. Int. Conf. on Document Analysis and Recognition, 2007.
- [2] Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media.
- [3] Google Translator API Documentation, Deep Translator Python Library, 2024.
- [4] S. Hochreiter & J. Schmidhuber, "Long Short-Term Memory," Neural Computation, 1997.
- [5] A. Vaswani et al., "Attention Is All You Need," NeurIPS, 2017.
- [6] V. Jagadeeswari et al., "Medical Internet of Things and Big Data in Personalized Healthcare," Health Inf. Sci. Syst., 2018.
- [7] A. Banerjee et al., "A Secure IoT-Fog Enabled Smart Decision-Making System Using Machine Learning for ICU," IEEE AISP Conf., 2020.
- [8] M. Kaur and S. Gupta, "Automated Report Interpretation Using NLP for Patient Health Records," IEEE Access, 2021.
- [9] A. Sharma et al., "Bridging Language Barriers in Healthcare with AI-based Translation Systems," Proc. Int. Conf. on AI in Healthcare, 2023.
- [10] Ministry of Health and Family Welfare, "Ayushman Bharat Digital Mission (ABDM)," Government of India, 2024.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details