IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, April 2025
Impact Factor: 7.67

Digitization and Advanced Analytics of Medical Records for Enhanced Healthcare Delivery

Prof. Poonam Narkhede[1], Manas Parab[2], Bhakti Patil[3], Kareena Darade[4]
Assistant Professor, Department of Computer Engineering[1]
Student, Department of Computer Engineering[2, 3, 4]
Shivajirao S Jondhale College of Engineering, Dombivli (E), Thane, Maharashtra, India

Abstract: *Managing and digitizing medical data in the healthcare industry presents significant challenges, particularly when dealing with physical prescriptions and patient records. MedSync is an advanced platform designed to streamline this process by leveraging Optical Character Recognition (OCR) technologies, such as Amazon Textract, to accurately extract text from medical documents. By integrating machine learning techniques like TF-IDF and Cosine Similarity, the system enhances data organization and accessibility while providing users with essential details about medications, including their composition, usage guidelines, potential side effects, and user ratings. Additionally, MedSync features an AI-powered chatbot to assist patients with their inquiries, improving access to healthcare information. Its predictive analysis component allows for the early detection of potential health risks by analyzing historical medical records. By incorporating these technologies, MedSync aims to enhance the efficiency, accuracy, and accessibility of healthcare data management. This paper explores the methodology, outcomes, and impact of MedSync in digitizing and standardizing medical records, ultimately improving patient care and supporting data-driven decision-making.*

Keywords: AI-powered Chatbot, Health Risk Detection, Data-Driven Decision-Making, Efficiency in Healthcare Data, Healthcare Information Accessibility, Historical Medical Records

I. INTRODUCTION

1.1 Overview

In recent years, data science has become a transformative force across various sectors, including IT, education, business, and healthcare. The healthcare industry, in particular, requires precise and secure data management due to the sensitive nature of patient information. Managing large volumes of paper-based medical records poses significant challenges, such as storage issues and the risk of data loss.

With advancements in Artificial Intelligence (AI) and Machine Learning (ML), healthcare systems are increasingly adopting predictive analytics and automation. These technologies help in forecasting medical outcomes, improving diagnosis, and enhancing patient care. However, the continued use of physical prescriptions and reports makes digitization a necessary step toward efficient data handling.

To address this, we propose **MedSync**, an integrated healthcare platform that digitizes handwritten prescriptions, provides detailed medicine information, and includes a conversational AI chatbot for user queries. This system ensures better record-keeping, improves access to medical information, and supports patient engagement. This paper outlines the development, methodology, and outcomes of the MedSync project.

1.2 Objectives

The system is designed to convert medical prescriptions from both handwritten and printed formats into digital text, ensuring precise extraction of health-related information for efficient data handling.

When a user searches for a specific medication, the platform provides comprehensive details, including its ingredients, intended uses, potential side effects, and patient feedback to enhance understanding.

DOI: 10.48175/IJARSCT-25255

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, April 2025
Impact Factor: 7.67

Additionally, the application recommends appropriate doctors by evaluating the user's medical issues and chosen location, making it easier to connect with nearby specialists.

Through the integration of machine learning, the system also examines previous health records to identify patterns and predict possible health concerns, supporting early detection and preventative measures.

To further assist users, an AI-driven chatbot is incorporated to answer medical questions, offer general health guidance, and help interpret symptoms, making basic healthcare information more accessible.

II. LITERATURE REVIEW

2.1 Literature Review & Model Evaluation

Rathod and Sari (2020) introduced a strategy for digitizing healthcare records, as outlined in [1]. They implemented Handwritten Character Recognition (HCR) to convert physical medical documents into digital formats. Their primary goal was to automate the digitization process for handwritten records, ensuring seamless storage and retrieval in databases. The system effectively demonstrated its intended purpose, improving both accuracy and efficiency in managing health data. However, a key limitation of HCR is its struggle to accurately recognize cursive handwriting.

A summary of research on machine learning-driven predictive analytics is provided in [2]. The findings indicate that machine learning is applied in predictive analysis within the healthcare domain in 56.7% of cases. Additionally, supervised learning accounts for 70% of these implementations, with Random Forest (RF) emerging as the most frequently used algorithm.

To enhance the efficiency of managing Electronic Health Records (EHR), researchers in [3] proposed a system incorporating a 3D human body model for data visualization while also digitizing paper-based medical records. The system leveraged Optical Character Recognition (OCR) technology to extract medical information. When tested on over 200 medical reports, it achieved a 100% success rate in digitization. However, the model's evaluation was limited to medical records written in Chinese script.

In [4], Kadadi and Agrawal (2014) investigated challenges related to big data integration and interoperability. Their research aimed to explore effective data integration techniques, address complexities within large datasets, and design an integration framework to resolve these issues. The study suggested several tools and technologies, including Hadoop, KARMA, and JNBridgePro, to improve data management. However, the real-world effectiveness of these solutions remained uncertain due to a lack of practical case studies.

Sinha and Jenckel (2019) explored OCR validation using Generative Adversarial Networks (GANs) in an unsupervised manner, as documented in [6]. Their experiment involved generating synthetic text images that closely resembled OCR-extracted text to improve validation. This approach enabled model evaluation even in cases where a softmax layer was unavailable. Despite this advancement, comparing generated images with original inputs posed challenges, especially since crucial stylistic attributes such as font styles (bold, italic) were not preserved.

Saluja and Punjabi (2019), in [7], proposed a method for improving OCR error correction in Indic languages using sub-word embeddings. Their model was trained to correct OCR errors by leveraging n-gram-based sub-word representations to accommodate complex linguistic structures. The fastText-based approach outperformed baseline models in terms of F-scores and Word Error Rates (WER), effectively handling out-of-vocabulary words. However, the accuracy of the results was highly dependent on the availability of high-quality training data. Discrepancies between training and test datasets led to increased WER for certain Malayalam words absent from the vocabulary.

These studies highlight the application of various technologies such as OCR, HCR, GANs, Hadoop, and FHIR for data digitization and standardization, as well as the integration of machine learning algorithms for predictive analytics.

Evaluation of Text Digitization Models

During the project's development, multiple models were assessed before selecting the most effective approach for text digitization.

1. Tesseract OCR with Regular Expressions

Process:

The input prescription file is first converted from PDF to an image using the **pdf2image** library.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, April 2025
Impact Factor: 7.67

The image undergoes preprocessing with the **OpenCV2** library.
The **Tesseract OCR Engine** extracts text from the processed image.
Using the **RegEx** library, extracted text is categorized into specific fields such as 'medicine_name' and 'patient_name.'
A **FastAPI** backend handles data extraction requests and returns results in **JSON format** for structured storage.

Limitations:

This approach worked well only for prescriptions following a predefined format, restricting scalability and adaptability.

2.2 Problem Statement

The continued use of handwritten medical notes and unorganized patient records in healthcare often results in delays, confusion, and mistakes in treatment. Because different hospitals and clinics use inconsistent formats, retrieving and understanding vital patient data becomes difficult—leading to lost information and unnecessary repeat tests.
MedSync tackles these issues by transforming physical medical documents into clean, standardized digital data. This system enhances accessibility, improves coordination among medical staff, and supports more accurate healthcare decisions, ultimately leading to safer and more efficient patient

III. METHODOLOGY

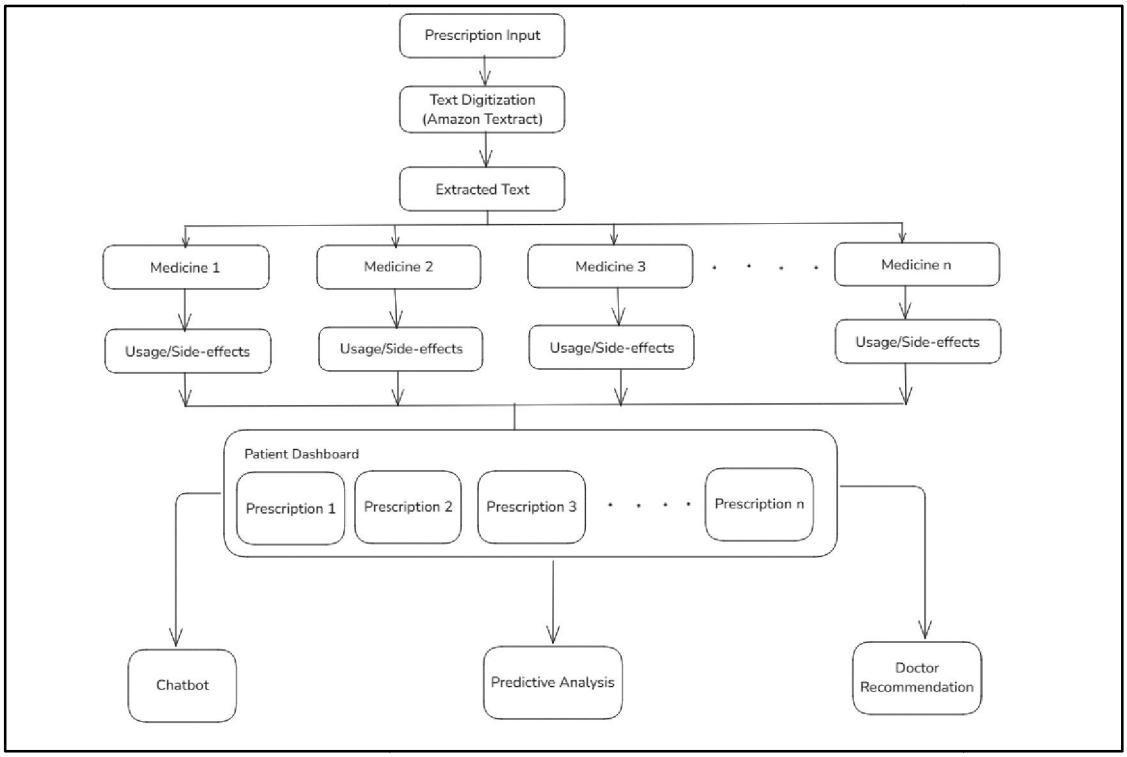

Figure 3.1 Flow Diagram

Figure 3.1illustrates the workflow of the MedSync system, which is segmented into the following key modules:

Text Extraction:

The user (or patient) begins by uploading an image of their prescription, typically in formats like JPG, JPEG, or HEIF. Once submitted, the image is processed using Amazon Textract—an OCR-based tool—which efficiently extracts the textual data from the image. The retrieved content is then shown on the output screen and securely saved within the system.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, April 2025
Impact Factor: 7.67

Medicine Details Retrieval:
When a user enters the name of a medication—either from their prescription or any other—they receive detailed information about it. The system uses *TF-IDF and Cosine Similarity* to match the input with entries in the medicine database. The output includes comprehensive data such as the medicine's ingredients, purpose, side effects, manufacturer, and user reviews.

AI Chatbot Support:
The platform includes a chatbot that responds to various health and medical questions. It generates detailed and relevant answers depending on the nature of the user's query. This chatbot functionality is powered by the *Groq API*.

Doctor Suggestion Module:
By entering symptoms or a diagnosed condition along with a preferred geographic location, users receive personalized doctor recommendations. The system returns a curated list of medical professionals specializing in the mentioned issue and practicing in the specified area.

IV. SYSTEM DESIGN

4.1 System Components

4.1.1 User Interface (Frontend):

- **Streamlit-Based Interface:** Provides an interactive and user-friendly web interface where users can upload prescription images, enter medicine-related queries, and interact with the chatbot.
- **Input Modules:** Users can either input text queries manually or upload prescription images in formats like JPG, JPEG, or HEIF.
- **Control Elements:** Buttons and fields are available for triggering actions such as uploading files, submitting queries, and viewing responses.

4.1.2 Processing Logic (Backend):

- **Main Controller Script:** Acts as the central logic hub, managing input requests and coordinating responses across various modules.
- **OCR Module:** Handles image-based inputs using Amazon Textract to extract text from prescription images accurately.
- **Medicine Information Module:** Processes text input from users to search and retrieve relevant drug details using text-matching algorithms like TF-IDF and cosine similarity.
- **Chatbot Handler:** Manages real-time health-related user queries using the Groq API to deliver informative responses.

4.1.3 AI Model Integration:

- **Amazon Textract & Groq API:** Amazon Textract is employed for high-precision OCR, while the Groq API powers the chatbot to understand and answer user queries effectively.

4.1.4 Environment and Security:

- **Secure API Management:** All API keys and sensitive credentials are stored as environment variables to ensure secure and authorized access to third-party services.

4.1.5 Session and Data Management:

- **Session Handling:** Maintains user session history to enable smooth multi-step interactions, such as continuing chat conversations and preserving previously entered inputs.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, April 2025
Impact Factor: 7.67

4.2 Backend Processing Workflow

4.2.1 User Interaction Flow:

- Users begin by accessing the MedSync web application through their browser.
- They can upload a prescription image or type in a medicine name or health-related question using the available input fields.

4.2.2 Request Handling and Logic Flow:

Based on the type of input—image or text—the backend chooses the appropriate module.

- **Image Upload (Prescription):** Sent to the OCR processing unit where Amazon Textract extracts medical data from the image.
- **Text Input (Medicine or Query):** Routed either to the medicine info module for data retrieval or the chatbot for query resolution.

4.2.3 AI Interaction and Processing:

The backend engages different AI services depending on the request:

- **For Image Inputs:** Amazon Textract is used to scan and interpret handwritten or printed prescriptions, converting them into structured text.
- **For Text Queries:** Groq API processes user inputs and provides clear, context-aware answers to medical questions.

4.2.4 Response Delivery:

- The results are presented through the Streamlit interface in a clean and user-friendly format.
- For chatbot responses, the system updates the conversation history in real-time to support fluid, ongoing interaction with the user.

V. RESULT AND ANALYSIS

Upon successful deployment of the selected methods and technologies, the outcomes achieved are as follows

Text Digitization

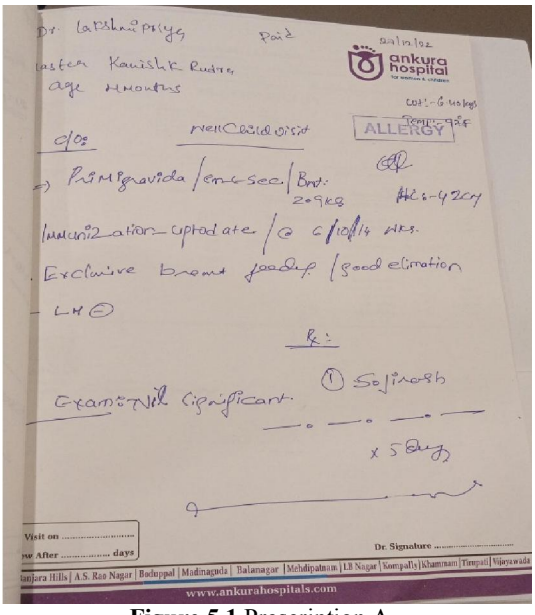

Figure 5.1. Prescription A

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, April 2025
Impact Factor: 7.67

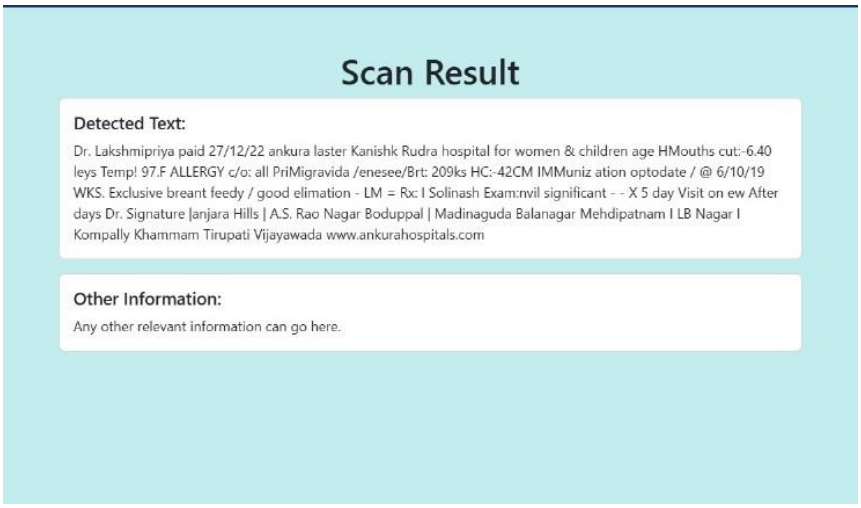

Figure 5.2. Digitized Text Using Amazon Textract

Chatbot:

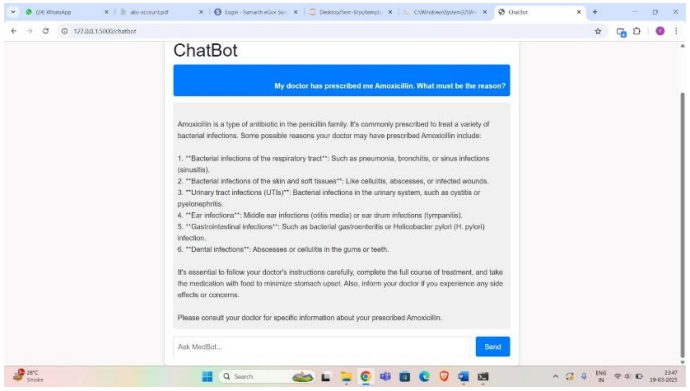

Figure 5.3. Chatbot

Medicine Information Finder:

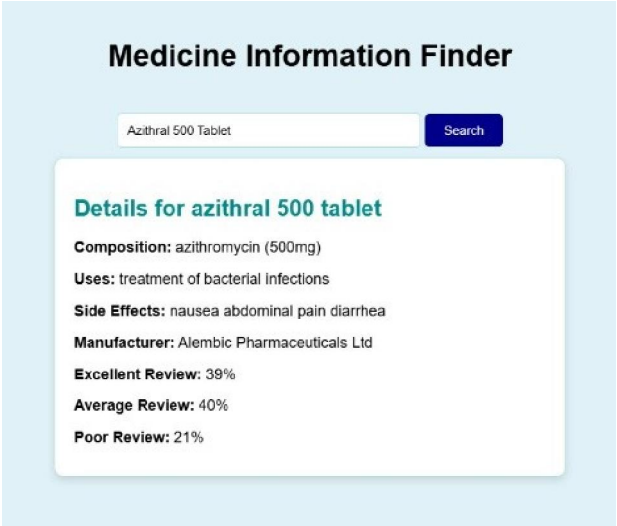

Figure 5.4. Medicine Information

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, April 2025
Impact Factor: 7.67

Doctor Recommendation:

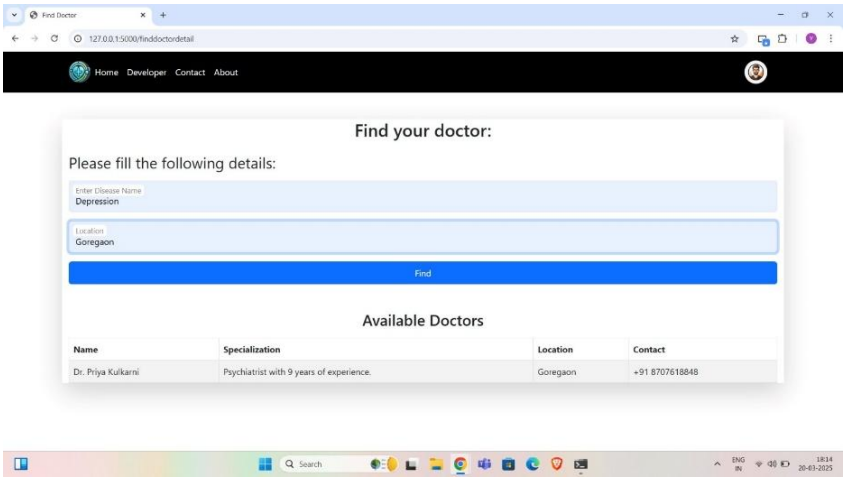

Figure 5.5. Doctor Recommendation

VI. CONCLUSION&FUTURE WORK

Conclusion

The MedSync project serves as a convenient and intuitive solution for managing patients' medical prescriptions, while also offering additional features such as a medicine information search tool, a healthcare chatbot, and doctor recommendations.

This functionality was made possible by addressing the limitations found in previous systems and methods reviewed during the literature survey, as well as by refining the early-stage text digitization models.

Future Work

Future enhancements to MedSync aim to boost its efficiency, accessibility, and intelligence. Expanding multilingual support through OCR and NLP would allow the system to handle prescriptions in various languages, making it more inclusive. Improving the medicine information finder with deep learning-based NER models and regularly updating the drug database would enhance accuracy and reliability. Upgrading the chatbot with transformer-based models like GPT or BERT could provide personalized, context-aware responses, and voice input integration would increase usability, especially for the elderly or disabled. Incorporating EHR and HMS integration with secure data sharing would offer a holistic solution for healthcare providers. Additionally, refining the doctor recommendation system using ML-based ranking and enabling telemedicine features would improve healthcare access. Performance could also be optimized through smart database indexing, caching, and distributed processing to handle growing user demands efficiently.

REFERENCES

[1] S. Rathod and S. Sarita, “Converting non-digitized health data to digital format,” *Asian Journal of Convergence in Technology*, vol. 6, no. 1, pp. 10–13, 2020.

[2] B. Loola, O. T. Khadija, and N. Souissi, “Predictive analysis using machine learning: Review of trends and methods,” in *Proc. Int. Symp. Advanced Electrical and Communication Technologies (ISAECT)*, 2020, pp. 1–6.

[3] N. Liu *et al.*, “A new data visualization and digitization method for building electronic health record,” in *Proc. IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM)*, 2020, pp. 2980–2982.

[4] A. Kadadi, R. Agrawal, C. Nyamful, and R. Atiq, “Challenges of data integration and interoperability in big data,” in *Proc. IEEE Int. Conf. Big Data*, 2014, pp. 38–40.

[5] V. Borisov, A. Minin, V. Basko, and A. Syskov, “FHIR data model for intelligent multimodal interface,” in *Proc. 26th Telecommunications Forum (TELFOR)*, 2018, pp. 420–425.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, April 2025
Impact Factor: 7.67

[6] A. Sinha, M. Jenckel, S. S. Bukhari, and A. Dengel, "Unsupervised OCR model evaluation using GAN," in *Proc. Int. Conf. Document Analysis and Recognition (ICDAR)*, 2019, pp. 1256–1261.
[7] R. Saluja, M. Punjabi, M. Carman, G. Ramakrishnan, and P. Chaudhuri, "Sub-word embeddings for OCR corrections in highly fusional Indic languages," in *Proc. Int. Conf. Document Analysis and Recognition (ICDAR)*, 2019, pp. 160–165.